Association for Women in Mathematics Series

Malena I. Español
Marta Lewicka
Lucia Scardia
Anja Schlömerkemper *Editors*

Research in Mathematics of Materials Science

AWM
ASSOCIATION FOR WOMEN IN MATHEMATICS

Springer

Association for Women in Mathematics Series

Volume 31

Series Editor
Kristin Lauter
Facebook
Seattle, WA, USA

Focusing on the groundbreaking work of women in mathematics past, present, and future, Springer's Association for Women in Mathematics Series presents the latest research and proceedings of conferences worldwide organized by the Association for Women in Mathematics (AWM). All works are peer-reviewed to meet the highest standards of scientific literature, while presenting topics at the cutting edge of pure and applied mathematics, as well as in the areas of mathematical education and history. Since its inception in 1971, The Association for Women in Mathematics has been a non-profit organization designed to help encourage women and girls to study and pursue active careers in mathematics and the mathematical sciences and to promote equal opportunity and equal treatment of women and girls in the mathematical sciences. Currently, the organization represents more than 3000 members and 200 institutions constituting a broad spectrum of the mathematical community in the United States and around the world.

Titles from this series are indexed by Scopus.

Malena I. Español • Marta Lewicka •
Lucia Scardia • Anja Schlömerkemper
Editors

Research in Mathematics of Materials Science

Springer

AWM
ASSOCIATION FOR
WOMEN IN MATHEMATICS

Editors

Malena I. Español
School of Mathematical and Statistical Sciences
Arizona State University
Tempe, AZ, USA

Marta Lewicka
Mathematics Department
University of Pittsburgh
Pittsburgh, PA, USA

Lucia Scardia
Department of Mathematics
Heriot-Watt University
Edinburgh, UK

Anja Schlömerkemper
Institute of Mathematics
Universität Würzburg
Würzburg, Germany

ISSN 2364-5733 ISSN 2364-5741 (electronic)
Association for Women in Mathematics Series
ISBN 978-3-031-04495-3 ISBN 978-3-031-04496-0 (eBook)
https://doi.org/10.1007/978-3-031-04496-0

Mathematics Subject Classification: 35Qxx, 49Jxx, 49Nxx, 49Sxx, 70Bxx, 70Cxx, 70Exx, 70Fxx, 70Hxx, 70Jxx, 70Kxx, 70Lxx, 70Qxx, 74Bxx, 74Cxx, 74Dxx, 74Hxx, 74Jxx, 74Kxx, 74Nxx, 74Pxx, 74Qxx, 74Rxx, 74Sxx

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Materials science is an interdisciplinary research field that incorporates the analytical and experimental techniques of engineering, physics, and chemistry. Its goal is to develop a comprehensive understanding of material characteristics and properties, and to gain a deeper insight into the material behavior across several length scales, from atomistic to macroscopic. Such insight is key to the design and discovery of new materials, as well as to the development of new technologies. In this context, *Research in Mathematics of Materials Science* seeks to provide a rigorous basis for future technological advances by developing tools for the accurate modeling of multi-scale phenomena. In the past decades, this field has seen rapid progress with an impressive level of success. Mathematics and materials science have thus developed strong ties and formed a partnership of mutual benefit.

This AWM Springer Volume highlights contributions of women mathematicians in the study of complex materials. The collected papers are both original research and reviews. The featured topics and methods draw on the fields of calculus of variations, partial differential equations, functional analysis, differential geometry and topology, as well as numerical analysis and mathematical modelling. Areas of applications include foams, fluid-solid interactions, liquid crystals, shape-memory alloys, magnetic suspensions, failure in solids, plasticity, viscoelasticity, homogenization, crystallization, grain growth, and phase-field models.

We hope that gathering such multifaceted scientific output in a single volume will facilitate knowledge exchange and help to identify new research problems. A parallel goal is to give visibility to the results of women researchers, and ultimately help address the gender gap in mathematics and science.

Tempe, AZ, USA — Malena I. Español
Pittsburgh, PA, USA — Marta Lewicka
Edinburgh, UK — Lucia Scardia
Würzburg, Germany — Anja Schlömerkemper
December 2021

Acknowledgements

We express gratitude to the authors who contributed their manuscripts and to the referees who provided valuable feedback. All articles in this volume have been peer reviewed. An acknowledgment extends to Kristin Lauter, editor of the AWM Series, for the opportunity of publishing this volume, and to Saveetha Balasundaram, our editor at Springer, for her guidance along the way.

Contents

Part II Review Papers

About the Editors

Malena I. Español is an Argentine-American applied mathematician specializing in numerical analysis and numerical linear algebra. Her research involves developing, analyzing, and applying mathematical models and numerical methods to problems arising in materials science and image processing. She is a faculty member in the School of Mathematical and Statistical Sciences at Arizona State University, USA.

Marta Lewicka is a mathematician specializing in analysis and PDEs. She has contributed results in the theory of nonlinear elasticity, hyperbolic systems of conservation laws, fluid dynamics, calculus of variations, potential theory, and differential games. She is a fellow of the American Mathematical Society and holds Professor's scientific title awarded by the President of the Republic of Poland. She works at the University of Pittsburgh, USA.

Lucia Scardia is a mathematician working in the calculus of variations and partial differential equations with applications in materials science. In particular, her interests include homogenization (deterministic and stochastic), harmonic analysis, non-local aggregation problems, dislocation theory, plasticity, and fracture and damage. She is an associate professor at Heriot-Watt University in Edinburgh, UK.

Anja Schlömerkemper is Professor of Mathematics specializing in mathematical analysis with applications to materials science and physics. She studies elastic and magnetic materials by involving methods from calculus of variations, PDEs, and homogenization theory. She is the president of the International Society for the Interaction of Mechanics and Mathematics, and she leads the chair of mathematics in the sciences at the University of Würzburg, Germany.

Part I
Research Papers

Interaction Between Oscillations and Singular Perturbations in a One-Dimensional Phase-Field Model

Annika Bach, Teresa Esposito, Roberta Marziani, and Caterina Ida Zeppieri

1 Introduction

In this note we study the asymptotic behaviour, via Γ-convergence, of one-dimensional integral functionals combining oscillations and singular perturbations occurring on two possibly different length scales, in the same spirit as [4, 10]. The functionals we consider are of Ambrosio–Tortorelli type and, for $\varepsilon > 0$ and $u, v \in W^{1,2}(a, b)$, they are defined as

$$F_\varepsilon(u, v) = \int_a^b \left(v^2 (u')^2 + \frac{(1-v)^2}{\varepsilon} + \varepsilon \varphi\Big(\frac{x}{\delta}\Big)(v')^2 \right) dx\,, \tag{1}$$

where $\varphi \in L^\infty(\mathbb{R})$ is a 1-periodic function. The scale-parameter $\delta = \delta(\varepsilon) > 0$ is infinitesimal as $\varepsilon \to 0$ and represents the characteristic length of some underlying heterogeneities. If

$$\alpha := \inf \varphi, \qquad \beta := \sup \varphi, \quad \text{with} \quad \alpha > 0,$$

then, up to a multiplicative constant, F_ε is bounded both from below and from above by the Ambrosio–Tortorelli functional [1, 2]; that is, we have

A. Bach
Dipartimento di Matematica “Guido Castelnuovo”, Sapienza Università di Roma, Rome, Italy
e-mail: annika.bach@uniroma1.it

T. Esposito · R. Marziani · C. I. Zeppieri (✉)
Angewandte Mathematik, WWU Münster, Münster, Germany
e-mail: teresa.esposito@uni-muenster.de; roberta.marziani@uni-muenster.de; caterina.zeppieri@uni-muenster.de

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_1

$$\begin{aligned}&\int_a^b v^2(u')^2\,dx + \int_a^b \left(\frac{(1-v)^2}{\varepsilon} + \varepsilon\alpha(v')^2\right)dx\\ &\qquad \le F_\varepsilon(u,v) \le \int_a^b v^2(u')^2\,dx + \int_a^b \left(\frac{(1-v)^2}{\varepsilon} + \varepsilon\beta(v')^2\right)dx.\end{aligned}$$

Therefore, as in the Ambrosio–Tortorelli approximation, the parameter ε determines the length scale of the diffuse approximation of the jump set of the limit variable. Indeed if $(u_\varepsilon, v_\varepsilon) \subset W^{1,2}(a,b) \times W^{1,2}(a,b)$ is a sequence along which F_ε is equibounded, then, necessarily, $v_\varepsilon \to 1$ in $L^2(a,b)$, while the first term in (1) favours those configurations where v_ε is asymptotically negligible, in the regions where u'_ε blows up. Then, as in the case of the Modica–Mortola functional [14, 15], v_ε makes a transition between 0 and 1 in a small layer of width proportional to ε. The cost of this transition is of order one and is bounded between the two constants $2\sqrt{\alpha}$ and $2\sqrt{\beta}$, the 2 appearing for symmetry reasons (cf. Remark 2). Moreover, the Γ-limit of F_ε (if it exists) shall satisfy

$$\int_a^b (u')^2\,dx + 2\sqrt{\alpha}\#S(u) \le \Gamma\text{-}\lim F_\varepsilon(u) \le \int_a^b (u')^2\,dx + 2\sqrt{\beta}\#S(u), \tag{2}$$

where $S(u)$ denotes the set of discontinuity points of u (and the limit variable v is omitted since it is equal to the constant function 1). The bounds in (2) then imply that the domain of the Γ-limit of F_ε is the space of piecewise-Sobolev functions $P\text{-}W^{1,2}(a,b)$. The latter coincides with the space of functions u which can be written as the sum of a Sobolev function $\tilde{u} \in W^{1,2}(a,b)$ and a piecewise-constant function $u^{\rm pc}$; thus $u' = \tilde{u}'$ and $S(u) = S(u^{\rm pc})$.

The main result of this note is Theorem 1 which establishes a Γ-convergence result for the functionals F_ε in every parameter regime; i.e., for every $\ell \in [0,+\infty]$, where

$$\ell := \lim_{\varepsilon\to 0} \frac{\varepsilon}{\delta(\varepsilon)}.$$

Specifically, we show that the sequence (F_ε) Γ-converges, with respect to the $L^1(a,b)$-convergence, to a functional which is always of Mumford–Shah type; i.e.,

$$F^\ell(u) = \int_a^b (u')^2\,dx + \mathbf{m}^\ell\#S(u)\,, \quad u \in P\text{-}W^{1,2}(a,b), \tag{3}$$

with a (constant) surface energy density $\mathbf{m}^\ell$ depending on the combined effect of the oscillations and the singular perturbation.

More precisely, we show that the lower bound in (2) is optimal when $\ell = 0$. That is, if $\varepsilon \ll \delta$, then $\mathbf{m}^0 = 2\sqrt{\alpha}$ and the Γ-limit of F_ε is given by the functional

$$F^0(u) = \int_a^b (u')^2\,dx + 2\sqrt{\alpha}\#S(u), \quad u \in P\text{-}W^{1,2}(a,b). \tag{4}$$

In this case a scale separation takes place. Indeed, formally, if in (1) we first take the Γ-limit in ε and keep δ fixed, we obtain the inhomogeneous free-discontinuity functionals (see [12])

$$\int_a^b (u')^2\,dx + 2\sum_{(a,b)\cap S(u)} \sqrt{\varphi\Big(\frac{x}{\delta}\Big)}, \quad u \in P\text{-}W^{1,2}(a,b),$$

whose Γ-limit as $\delta \to 0$ is exactly given by (4) (see, *e.g.*, [7, Section 9.3]).

For $\ell = +\infty$, which corresponds to the case $\delta \ll \varepsilon$, we also observe a scale separation. In fact being the oscillation parameter δ smaller than the approximation parameter ε, in this regime, the Γ-limit of F_ε is the same as that of the homogeneous functionals

$$\int_a^b v^2(u')^2\,dx + \int_a^b \bigg(\frac{(1-v)^2}{\varepsilon} + \varepsilon\,\varphi_{\text{hom}}(v')^2\bigg)\,dx,$$

where φ_{hom} is the harmonic mean of φ in $(0,1)$; i.e.,

$$\varphi_{\text{hom}} := \bigg(\int_0^1 \frac{1}{\varphi(t)}\,dt\bigg)^{-1}.$$

Therefore, passing to the limit as $\varepsilon \to 0$ gives

$$F^\infty(u) = \int_a^b (u')^2\,dx + 2\sqrt{\varphi_{\text{hom}}}\#S(u), \quad u \in P\text{-}W^{1,2}(a,b), \tag{5}$$

that is, $\mathbf{m}^\infty = 2\sqrt{\varphi_{\text{hom}}}$. We notice that, in general, $\varphi_{\text{hom}} \le \beta$.

Finally, in the case $\ell \in (0,+\infty)$ the parameters ε and δ, being of the same order, interact with one another producing a surface energy $\mathbf{m}^\ell$ which depends on their interplay according to the following formula

$$\mathbf{m}^\ell = \inf_{z\in[0,1)} \inf\bigg\{ \int_{\mathbb{R}} \Big((1-v)^2 + \varphi(\ell x)(v')^2\Big)\,dx \colon v \in W^{1,2}_{\text{loc}}(\mathbb{R}), \; v(z/\ell) = 0, \; \lim_{t\to\pm\infty} v(t) = 1 \bigg\}. \tag{6}$$

We notice that, in contrast to the typical optimal-profile problem for the Ambrosio–Tortorelli functional (cf. (26)) which determines both $\mathbf{m}^0$ and $\mathbf{m}^\infty$, the minimisation problem in (6) involves the (unscaled) Modica–Mortola term in F_ε on the whole real line, instead of $(0,+\infty)$. This is due to the presence of the inhomogeneity φ, which breaks the usual symmetry of the problem. Moreover, an additional optimisation on the parameter $z \in [0, 1)$ is needed to determine the "starting point" of an optimal transition. This feature makes the present problem different from the corresponding one for the Modica–Mortola functional considered in [3, 4] (see also [7, Chapter 9]).

Eventually, we conclude the limit analysis of the functionals F_ε by proving that the surface energy density $\mathbf{m}^\ell$ is continuous with respect to the parameter ℓ; i.e., we show that

$$\lim_{\ell \to 0^+} \mathbf{m}^\ell = \mathbf{m}^0 \quad \text{and} \quad \lim_{\ell \to +\infty} \mathbf{m}^\ell = \mathbf{m}^\infty.$$

We finally observe that the functionals in (1) have also a mechanical interpretation. Indeed they can be seen as a one-dimensional variational model for damage in heterogeneous materials, according to, e.g., [11, 13, 16, 17]. We also notice that due to the presence of the two interacting scales ε and δ, a Γ-convergence analysis for the corresponding n-dimensional model makes it necessary to resort to a more abstract method of proof, as shown in [6]. This method relies, among other, on the Γ-convergence analysis for general (scale-dependent, non-periodic) elliptic functionals recently developed in [5]. In particular, the result established in [5] shows that the Γ-limit of the n-dimensional counterpart of (1) is always of brittle type, this fact being a consequence of a volume-surface decoupling which takes place in the Γ-limit. On the other hand, the one-dimensional problem studied in this note can be solved directly, by hands, taking advantage of the simple form of the functionals F_ε and of the structure of the space of one-dimensional special functions of bounded variation, $SBV^2(a, b)$, which coincides with the space of piecewise-Sobolev functions $P\text{-}W^{1,2}(a, b)$. In particular, in the proof of the upper-bound inequality (in the three different scaling regimes), the structure of $P\text{-}W^{1,2}(a, b)$ allows us to treat the regular and singular part of the limit variable u separately, without resorting to the abstract decoupling result established in [5].

2 Setting of the Problem and Statement of the Main Result

In this section we define the phase-field functionals we are going to analyse and we state our main result.

Let $\varphi \in L^\infty(\mathbb{R})$ be a 1-periodic function and set

$$\alpha := \inf \varphi, \qquad \beta := \sup \varphi; \tag{7}$$

we additionally assume that $\alpha > 0$.

Let $\varepsilon > 0$ and let $\delta_\varepsilon >$ be such that $\lim_{\varepsilon \to 0} \delta_\varepsilon = 0$. For $a, b \in \mathbb{R}$ with $a < b$ we consider the one-dimensional integral functionals $F_\varepsilon \colon L^1(a, b) \times L^1(a, b) \longrightarrow [0, +\infty]$ defined by

$$F_\varepsilon(u,v) := \begin{cases} \displaystyle\int_a^b \Big(v^2(u')^2 + \frac{(1-v)^2}{\varepsilon} + \varepsilon\,\varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v')^2 \Big)\,dx & u,\ v \in W^{1,2}(a,b), \\ & 0 \le v \le 1, \\ +\infty & \text{otherwise.} \end{cases} \tag{8}$$

We notice that thanks to (7) the functionals F_ε satisfy

$$AT_\varepsilon^\alpha(u,v) \le F_\varepsilon(u,v) \le AT_\varepsilon^\beta(u,v)\,, \tag{9}$$

where, for $\lambda > 0$, AT_ε^λ is the one-dimensional Ambrosio–Tortorelli functional given by

$$AT_\varepsilon^\lambda(u,v) := \begin{cases} \displaystyle\int_a^b \Big(v^2(u')^2 + \frac{(1-v)^2}{\varepsilon} + \varepsilon\lambda(v')^2 \Big)\,dx & u,\ v \in W^{1,2}(a,b), \\ & 0 \le v \le 1, \\ +\infty & \text{otherwise.} \end{cases} \tag{10}$$

For later use it is convenient to define the localised functionals

$$F_\varepsilon(u,v,I) := \begin{cases} \displaystyle\int_I \Big(v^2(u')^2 + \frac{(1-v)^2}{\varepsilon} + \varepsilon\,\varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v')^2 \Big)\,dx & u,\ v \in W^{1,2}(a,b), \\ & 0 \le v \le 1, \\ +\infty & \text{otherwise,} \end{cases} \tag{11}$$

where $I \subset (a,b)$ is any open interval. Analogously, we define a localised version of the Modica–Mortola term in F_ε by setting

$$G_\varepsilon(v,I) := \begin{cases} \displaystyle\int_I \Big(\frac{(1-v)^2}{\varepsilon} + \varepsilon\,\varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v')^2 \Big)\,dx & v \in W^{1,2}(a,b),\ 0 \le v \le 1, \\ +\infty & \text{otherwise.} \end{cases} \tag{12}$$

As for the Ambrosio–Tortorelli functional, the Γ-limit of F_ε will be defined on a space of discontinuous functions. Then, to describe the domain of the limit functional, we need to introduce the space P-$W^{1,2}(a,b)$. The latter denotes the space of *piecewise* $W^{1,2}(a,b)$-functions defined on the interval (a,b). That is, $u \in P$-$W^{1,2}(a,b)$ if and only if there exists a *finite* partition of (a,b), $a = t_0 < t_1 < \ldots < t_M = b$, such that $u \in W^{1,2}(t_i, t_{i+1})$, for every $i = 1, \ldots, M-1$. The discontinuity set of a function $u \in P$-$W^{1,2}$ is denoted by $S(u)$ and it coincides with the minimal of such sets of points.

Let $PC(a,b)$ denote the space of *piecewise-constant* functions on (a,b); then it is easy to check that

$$P\text{-}W^{1,2}(a,b) = W^{1,2}(a,b) + PC(a,b), \tag{13}$$

that is, $u \in P\text{-}W^{1,2}(a,b)$ if and only if

$$u = \tilde{u} + u^{\mathrm{pc}}, \tag{14}$$

with $\tilde{u} \in W^{1,2}(a,b)$ and $u^{\mathrm{pc}} \in PC(a,b)$. We also notice that the sum in (13) is not a direct sum since the constant functions belong to $W^{1,2}(a,b) \cap PC(a,b)$; therefore, the decomposition in (14) is uniquely determined up to an additive constant.

Thanks to (14), for $u \in P\text{-}W^{1,2}(a,b)$ we have

$$u' = \tilde{u}' \quad \text{and} \quad S(u) = S(u^{\mathrm{pc}}).$$

Set

$$\ell := \lim_{\varepsilon \to 0} \frac{\varepsilon}{\delta_\varepsilon} \in [0, +\infty].$$

The following Γ-convergence theorem is the main result of this paper.

Theorem 1 *The sequence of functionals* (F_ε) *defined in* (8) $\Gamma(L^1 \times L^1)$*-converges to the functional* $F^\ell\colon L^1(a,b) \times L^1(a,b) \longrightarrow [0, +\infty]$ *defined as*

$$F^\ell(u,v) := \begin{cases} \displaystyle\int_a^b (u')^2\,dx + \mathbf{m}^\ell \# S(u) & u \in P\text{-}W^{1,2}(a,b),\ v = 1 \textit{ a.e. in } (a,b)\,, \\ +\infty & \textit{otherwise}\,. \end{cases} \tag{15}$$

Moreover, the constant $\mathbf{m}^\ell > 0$ *is defined as follows:*

1. *if* $\ell = 0$ *and* φ *is upper semicontinuous, then*

$$\mathbf{m}^0 := 2\sqrt{\alpha}; \tag{16}$$

2. *if* $\ell \in (0, +\infty)$*, then*

$$\mathbf{m}^\ell := \inf_{z \in [0,1)} \mathbf{m}^\ell_z\,, \tag{17}$$

with

$$\mathbf{m}^\ell_z := \inf\left\{ \int_{\mathbb{R}} \big((1-v)^2 + \varphi(\ell x + z)(v')^2\big)\,dx \colon v \in W^{1,2}_{\mathrm{loc}}(\mathbb{R}),\ 0 \le v \le 1, \right. \\ \left. v(0) = 0,\ v(\pm\infty) = 1 \right\}, \tag{18}$$

where $v(\pm\infty) := \lim_{x \to \pm\infty} v(x)$*;*

3. if $\ell = +\infty$, then

$$\mathbf{m}^{\infty} := 2\left(\int_0^1 \frac{1}{\varphi(t)}\,dt\right)^{-1/2}. \tag{19}$$

Eventually, the constant $\mathbf{m}^{\ell}$ satisfies

$$\lim_{\ell\to 0^+}\mathbf{m}^{\ell} = \mathbf{m}^0 \quad \textit{and} \quad \lim_{\ell\to+\infty}\mathbf{m}^{\ell} = \mathbf{m}^{\infty}\,, \tag{20}$$

provided φ is upper semicontinuous.

3 Preliminary Results

In this section we state and prove some preliminary results which will be used in what follows. We start recalling the convergence result for the 1-dimensional Ambrosio–Tortorelli functionals defined in (10) (see, e.g., [8, Theorem 3.15]).

Theorem 2 *For any $\lambda > 0$ the functionals AT_ε^λ defined in (10) $\Gamma(L^1\times L^1)$-converge as $\varepsilon \to 0$ to the functional*

$$MS^\lambda(u,v) := \begin{cases} \displaystyle\int_a^b (u')^2\,dx + 2\sqrt{\lambda}\#S(u) & u \in P\text{-}W^{1,2}(a,b),\ v = 1 \ \textit{a.e. in}\ (a,b)\,, \\ +\infty & \textit{otherwise}\,. \end{cases}$$

The next proposition establishes a compactness result for sequences with equibounded energy and a lower bound for the first term in F_ε, which is independent of the parameter regime.

Proposition 1 *Let F_ε be as in (8) and let $(u_\varepsilon, v_\varepsilon) \subset W^{1,2}(a,b)\times W^{1,2}(a,b)$ be such that*

$$u_\varepsilon \to u \ \ \textit{in} \ \ L^1(a,b) \quad \textit{and} \quad \sup_{\varepsilon>0} F_\varepsilon(u_\varepsilon, v_\varepsilon) < +\infty.$$

Then, there holds

1. *$v_\varepsilon \to 1$ in $L^2(a,b)$, $u \in P\text{-}W^{1,2}(a,b)$ and*

$$\liminf_{\varepsilon\to 0}\int_a^b v_\varepsilon^2 (u_\varepsilon')^2\,dx \ge \int_a^b (u')^2\,dx\,; \tag{21}$$

2. *If $S(u) = \{t_1,\dots,t_N\}$ and $I_1,\dots,I_N$ are pairwise disjoint open subintervals in (a,b) such that $t_i \in I_i$, for every $i = 1,\dots,N$, then there exist $s_\varepsilon^1,\dots,s_\varepsilon^N$ with $(s_\varepsilon^i) \subset I_i$ for every $\varepsilon > 0$, such that*

$$s_\varepsilon^i \to t_i \quad and \quad v_\varepsilon(s_\varepsilon^i) \to 0 \quad as\ \varepsilon \to 0, \tag{22}$$

for every $i = 1, \dots, N$.

Proof Thanks to (9), the proof readily follows from the corresponding one for the Ambrosio–Tortorelli functional (see, e.g., [8, Theorem 3.15]). □

Remark 1 Let $(u_\varepsilon, v_\varepsilon)$ be as in Proposition 1 and $I_1, \dots, I_N$, s_ε^i as in Proposition 1 (2). Since (1) implies that, up to subsequences, $v_\varepsilon \to 1$ a.e. in (a, b), we can find $r^i, \tilde{r}^i \in I_i$ with $r^i < s_\varepsilon^i < \tilde{r}^i$ such that

$$\lim_{\varepsilon\to 0} v_\varepsilon(r^i) = \lim_{\varepsilon\to 0} v_\varepsilon(\tilde{r}^i) = 1. \tag{23}$$

In particular, since v_ε is continuous, we can apply the Intermediate Value Theorem to deduce that, for any $\eta \in (0, 1/2)$ fixed, there exist $\tilde{s}_\varepsilon^i, r_\varepsilon^i, \tilde{r}_\varepsilon^i \in I_i$ (depending also on η) with $r_\varepsilon^i < \tilde{s}_\varepsilon^i < \tilde{r}_\varepsilon^i$ such that

$$v_\varepsilon(\tilde{s}_\varepsilon^i) = \eta\,,\ v_\varepsilon(r_\varepsilon^i) = v_\varepsilon(\tilde{r}_\varepsilon^i) = 1 - \eta \quad \text{and} \quad v_\varepsilon \le 1 - \eta \ \text{ in } [r_\varepsilon^i, \tilde{r}_\varepsilon^i]\,. \tag{24}$$

Set $M := \sup_\varepsilon F_\varepsilon(u_\varepsilon, v_\varepsilon)$; since by assumption $M < +\infty$, from (24) we infer

$$M \ge \int_{r_\varepsilon^i}^{\tilde{r}_\varepsilon^i} \frac{(1 - v_\varepsilon)^2}{\varepsilon}\, dx \ge \frac{\eta^2}{\varepsilon}(\tilde{r}_\varepsilon^i - r_\varepsilon^i) \text{ and } M \ge \alpha \int_{\tilde{s}_\varepsilon^i}^{\tilde{r}_\varepsilon^i} \varepsilon(v_\varepsilon')^2\, dx \ge \frac{\varepsilon\,\alpha(1 - 2\eta)^2}{\tilde{r}_\varepsilon^i - \tilde{s}_\varepsilon^i}\,,$$

where the last estimate follows from Jensen's Inequality. Therefore, for every $\varepsilon > 0$ we get

$$\frac{\alpha(1 - 2\eta)^2}{M} \le \frac{\tilde{r}_\varepsilon^i - \tilde{s}_\varepsilon^i}{\varepsilon} < \frac{\tilde{r}_\varepsilon^i - r_\varepsilon^i}{\varepsilon} \le \frac{M}{\eta^2} \tag{25}$$

and similarly for $\frac{\tilde{s}_\varepsilon^i - r_\varepsilon^i}{\varepsilon}$.

3.1 The Optimal-Profile Problem

In this subsection we study the minimisation problem defining the constant $\mathbf{m}^\ell$ in (17). The latter represents the minimal cost of a two-sided transition from the value 0 to the value 1, on the real line, in terms of the unscaled Modica–Mortola term in F_ε. We thus refer to the corresponding minimisation problem as the optimal-profile problem. The analysis of $\mathbf{m}^\ell$ will be useful both to prove the Γ-convergence result in the regime $\delta_\varepsilon \sim \varepsilon$ and to establish (20) in Theorem 1.

We start recalling some properties of the corresponding optimal-profile problem for the Ambrosio–Tortorelli functionals AT_ε^λ defined in (10).

Remark 2 Let $\lambda > 0$; arguing as in, e.g., [7, Chapter 6] it is immediate to check that

$$\begin{aligned}\sqrt{\lambda} &= \min\Big\{ \int_0^{+\infty} \big((1-v)^2 + \lambda\,(v')^2\big)\,dx \colon\ v \in W^{1,2}_{\mathrm{loc}}(0,+\infty),\\ &\qquad\qquad 0 \le v \le 1,\ v(0) = 0,\ v(+\infty) = 1 \Big\}\\ &= \inf_{T>0} \min\Big\{ \int_0^{T} \big((1-v)^2 + \lambda\,(v')^2\big)\,dx \colon\ v \in W^{1,2}(0,T),\\ &\qquad\qquad 0 \le v \le 1,\ v(0) = 0,\ v(T) = 1 \Big\}. \end{aligned} \tag{26}$$

Let $\mathbf{m}^\ell_z$ be as in (18); from (26) using a reflection argument and choosing either $\lambda = \alpha$ or $\lambda = \beta$, in view of (7) we get

$$2\sqrt{\alpha} \le \mathbf{m}^\ell_z \le 2\sqrt{\beta}\,, \tag{27}$$

for every $\ell \in (0,+\infty)$ and every $z \in [0,1)$.

The following lemma shows that the cost of an optimal profile depends continuously on the value attained by the competitors at zero.

Lemma 1 *For $\ell \in (0,+\infty)$, $z \in [0,1)$, and $t \in [0,1)$ let*

$$\begin{aligned}\mathbf{m}^\ell_z(t) := \inf\Big\{ \int_{\mathbb{R}} \big((1-v)^2 + \varphi(\ell x + z)(v')^2\big)\,dx \colon\ v \in W^{1,2}_{\mathrm{loc}}(\mathbb{R}),\ 0 \le v \le 1,\\ v(0) = t\ ,\, v(\pm\infty) = 1 \Big\}, \end{aligned} \tag{28}$$

and set

$$\mathbf{m}^\ell(t) := \inf_{z\in[0,1)} \mathbf{m}^\ell_z(t)\,, \tag{29}$$

so that, in particular, $\mathbf{m}^\ell(0) = \mathbf{m}^\ell$, *with* $\mathbf{m}^\ell$ *as in* (17). *Then* $\lim_{t\to 0} \mathbf{m}^\ell(t) = \mathbf{m}^\ell$.

Proof Let $\ell \in (0,+\infty)$ be fixed; let $z \in [0,1)$ be arbitrary and let $v \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$ be admissible for the infimum problem defining $\mathbf{m}^\ell_z$ in (18); i.e., in particular, $v(0) = 0$. For any $t \in [0,1)$ the function

$$v_t := \min\{v + t, 1\}$$

is admissible for the infimum problem defining $\mathbf{m}^\ell_z(t)$ and satisfies

$$\int_{\mathbb{R}} \big((1-v_t)^2 + \varphi(\ell x + z)(v_t')^2\big)\,dx \le \int_{\mathbb{R}} \big((1-v)^2 + \varphi(\ell x + z)(v')^2\big)\,dx\,.$$

Passing to the infimum in v and z we obtain both

$$\mathbf{m}^\ell(t) \le \mathbf{m}^\ell \quad \text{for every } t \in [0,1) \quad \text{and} \quad \limsup_{t\to 0} \mathbf{m}^\ell(t) \le \mathbf{m}^\ell. \tag{30}$$

Thus, to conclude it remains to show that

$$\liminf_{t\to 0} \mathbf{m}^\ell(t) \ge \mathbf{m}^\ell. \tag{31}$$

To this end, we fix $\eta \in (0,1/2)$ and for any $t \in (0,1/4)$ we choose $z_{\eta,t} \in [0,1)$ and $v_{\eta,t} \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$ with $0 \le v_{\eta,t} \le 1$, $v_{\eta,t}(0) = t$, $v_{\eta,t}(\pm\infty) = 1$ such that

$$\int_{\mathbb{R}} \big((1-v_{\eta,t})^2 + \varphi(\ell x + z_{\eta,t})(v_{\eta,t}')^2\big)\,dx \le \mathbf{m}^\ell(t) + \eta\,. \tag{32}$$

Since $v_{\eta,t}(\pm\infty) = 1$ and $v_{\varepsilon,t}$ is continuous, we can apply the Intermediate Value Theorem to find $T^1_{\eta,t}, T^2_{\eta,t}$ with $T^1_{\eta,t} < 0 < T^2_{\eta,t}$ such that

$$v_{\eta,t}(T^1_{\eta,t}) = v_{\eta_t}(T^2_{\eta,t}) = 1-\eta \quad \text{and} \quad v_{\eta,t} \le 1-\eta \;\text{ on }\; [T^1_{\eta,t}, T^2_{\eta,t}]\,. \tag{33}$$

Notice that the second condition in (33) together with (32) implies that

$$\mathbf{m}^\ell(t) + \eta \ge \int_{T^1_{\eta,t}}^{T^2_{\eta,t}} (1-v_{\eta,t})^2\,dx \ge \eta^2(T^2_{\eta,t} - T^1_{\eta,t})\,.$$

Thus, combining (27) and (30) yields

$$(T^2_{\eta,t} - T^1_{\eta,t}) \le \frac{2\sqrt{\beta}+\eta}{\eta^2} \quad \text{uniformly in } t\,. \tag{34}$$

Next we define

$$w_{\eta,t}(x) := \begin{cases} 1-(\eta+t)(x - T^1_{\eta,t} + 1) & \text{if } T^1_{\eta,t} - 1 \le x < T^1_{\eta,t}\,, \\ \max\{0, v_{\eta,t}(x) - t\} & \text{if } T^1_{\eta,t} \le x \le T^2_{\eta,t}\,, \\ 1+(\eta+t)(x - T^2_{\eta,t} - 1) & \text{if } T^2_{\eta,t} \le x < T^2_{\eta,t} + 1\,, \\ 1 & \text{otherwise in } \mathbb{R}\,, \end{cases}$$

(see Fig. 1).

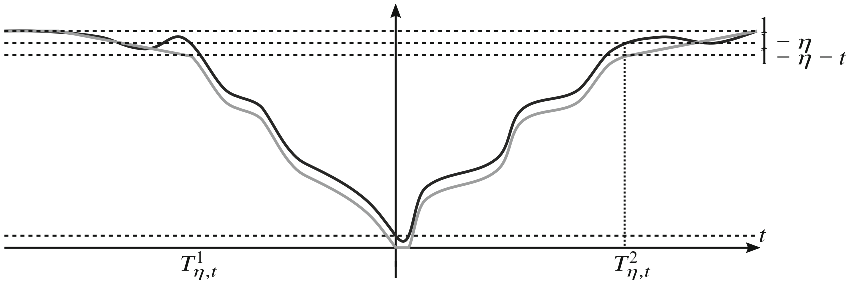


Fig. 1 The function $v_{\eta,t}$ (in dark grey) and the modification $w_{\eta,t}$ (in light grey)

The first condition in (33) ensures that $w_{\eta,t} \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$. Moreover, we have $w_{\eta,t}(0) = v_{\eta,t}(0) - t = 0$ and $v_{\eta,t}(\pm\infty) = 1$. In particular, $w_{\eta,t}$ is admissible for $\mathbf{m}^{\ell}_{z}$ for any $z \in [0, 1)$, so that

$$\mathbf{m}^{\ell} \leq \int_{\mathbb{R}} \left((1 - w_{\eta,t})^2 + \varphi(\ell x + z_{\eta,t})(w'_{\eta,t})^2\right) dx \,. \tag{35}$$

Further, since the map $s \mapsto (1-s)^2$ is decreasing on $(-\infty, 1)$ we get

$$\begin{aligned}
&\int_{T^1_{\eta,t}}^{T^2_{\eta,t}} \left((1 - w_{\eta,t})^2 + \varphi(\ell x + z_{\eta,t})(w'_{\eta,t})^2\right) dx \\
&\quad \leq \int_{T^1_{\eta,t}}^{T^2_{\eta,t}} \left((1 - v_{\eta,t} - t)^2 + \varphi(\ell x + z_{\eta,t})(v'_{\eta,t})^2\right) dx \\
&\quad \leq (1+\eta) \int_{T^1_{\eta,t}}^{T^2_{\eta,t}} \left((1 - v_{\eta,t})^2 + \varphi(\ell x + z_{\eta,t})(v'_{\eta,t})^2\right) dx + \Big(1 + \frac{1}{\eta}\Big) t^2 (T^2_{\eta,t} - T^1_{\eta,t})\,,
\end{aligned} \tag{36}$$

where the second inequality follows by expanding the square $(1 - v_{\eta,t} - t)^2$ and applying Young's Inequality to the term $2\sqrt{\eta}(1 - v_{\eta,t})\frac{t}{\sqrt{\eta}}$. Eventually, by definition of $w_{\eta,t}$, from (7) we infer

$$\begin{aligned}
&\int_{\mathbb{R}\setminus[T^1_{\eta,t}, T^2_{\eta,t}]} \left((1 - w_{\eta,t})^2 + \varphi(\ell x + z_{\eta,t})(w'_{\eta,t})^2\right) dx \\
&\quad \leq (\eta + t)^2 \bigg(\int_{T^1_{\eta,t}-1}^{T^1_{\eta,t}} (x - T^1_{\eta,t} + 1)^2 \, dx + \int_{T^2_{\eta,t}}^{T^2_{\eta,t}+1} (x - T^2_{\eta,t} - 1)^2 \, dx + 2\beta \bigg) \\
&\quad = 2(\eta + t)^2 \Big(\frac{1}{3} + \beta\Big)\,.
\end{aligned} \tag{37}$$

Thus, inserting (34) in (36) and combining (32) with (35)–(37) we deduce that

$$\mathbf{m}^{\ell} \leq (1+\eta)\big(\mathbf{m}^{\ell}(t)+\eta\big) + \Big(1+\frac{1}{\eta}\Big)t^2\frac{2\sqrt{\beta}+\eta}{\eta^2} + 2(\eta+t)^2\Big(\frac{1}{3}+\beta\Big)\,.$$

Passing in the above inequality first to the liminf in t and then to the limit as $\eta \to 0$ we finally obtain (31). □

Remark 3 We observe that for $\ell \in (0,+\infty)$, in general the strict inequality $\mathbf{m}^0 < \mathbf{m}^{\ell}$ holds. To prove it, assume φ is continuous with $0 < \alpha = \min\varphi < \max\varphi = \beta$. Then, the direct methods and a truncation argument provide us with a pair $(\bar{z},\bar{v}) \in [0,1) \times W^{1,2}_{\mathrm{loc}}(\mathbb{R})$ with $0 \leq \bar{v} \leq 1$, $\bar{v}(0)=0$, $\bar{v}(\pm\infty)=1$, such that

$$\mathbf{m}^{\ell} = \int_{\mathbb{R}} (1-\bar{v})^2 + \varphi(\ell x+\bar{z})(\bar{v}')^2\,dx\,. \tag{38}$$

Therefore, the Young Inequality yields

$$\begin{aligned}\mathbf{m}^{\ell} &= \int_{\mathbb{R}} (1-\bar{v})^2 + \varphi(\ell x+\bar{z})(\bar{v}')^2\,dx \\ &= \int_{\mathbb{R}} \Big((1-\bar{v})^2 + \alpha(\bar{v}')^2\Big)\,dx + \int_{\mathbb{R}} (\varphi(\ell x+\bar{z})-\alpha)(\bar{v}')^2\,dx \\ &\geq 2\sqrt{\alpha} + \int_{\mathbb{R}} (\varphi(\ell x+\bar{z})-\alpha)(\bar{v}')^2\,dx\,, \end{aligned} \tag{39}$$

with equality if and only if $\bar{v}$ satisfies $\alpha\,\bar{v}' = 1-\bar{v}$; i.e., $\bar{v} = 1-\exp(-|x|/\sqrt{\alpha})$. If this is the case, then $|\bar{v}'(x)| > 0$ for every $x \in \mathbb{R}\setminus\{0\}$, which will imply that the second term on the right-hand side of (39) is strictly positive by the assumptions on φ. Thus, the claim follows.

Remark 4 For later reference it is useful to observe that for every $\ell \in (0,+\infty)$ and $z \in [0,1)$ the constant $\mathbf{m}^{\ell}_z$ in (18) can be equivalently expressed in terms of a minimisation problem where the test functions are suitably shifted, instead of the integrand. Indeed, consider the shifted function $v_z := v(\,\cdot\,-\frac{z}{\ell})$; if $v \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$, then v_t belongs to $W^{1,2}_{\mathrm{loc}}(\mathbb{R})$, moreover $v(0)=0$, $v(\pm\infty)=1$ if and only if $v_z(\frac{z}{\ell})=0$, $v_z(\pm\infty)=1$. Therefore, since the change of variables $y = x+\frac{z}{\ell}$ gives

$$\int_{\mathbb{R}} \Big((1-v)^2 + \varphi(\ell x+z)(v')^2\Big)\,dx = \int_{\mathbb{R}} \Big((1-v_z)^2 + \varphi(\ell y)(v_z')^2\Big)\,dy\,,$$

passing to the infimum we get

$$\mathbf{m}_z^\ell = \inf\left\{ \int_{\mathbb{R}} \big((1-v)^2 + \varphi(\ell x)(v')^2\big)\,dx \colon v \in W^{1,2}_{\mathrm{loc}}(\mathbb{R}), \right.$$
$$\left. 0 \le v \le 1,\ v(\tfrac{z}{\ell}) = 0,\ v(\pm\infty) = 1 \right\}. \tag{40}$$

Finally, in the next proposition we prove an alternative formula for the surface density of the Γ-limit in the regime $\delta_\varepsilon \sim \varepsilon$ (cf. [5, Theorem 8.4]).

Proposition 2 *Let $\ell \in (0,+\infty)$ and set*

$$\widetilde{\mathbf{m}}^\ell := \inf\left\{ \int_{\mathbb{R}} \big((1-v)^2+\varphi(\ell x)(v')^2\big)\,dx \colon v \in W^{1,2}_{\mathrm{loc}}(\mathbb{R}),\ 0 \le v \le 1,\ v(\pm\infty) = 1, \right.$$
$$\left. \exists\, u \in W^{1,2}_{\mathrm{loc}}(\mathbb{R}) \text{ with } u(-\infty) = 0,\ u(+\infty) = 1 \text{ and } v\,u' = 0 \text{ a.e. in } \mathbb{R} \right\}.$$

Then $\widetilde{\mathbf{m}}^\ell = \mathbf{m}^\ell$, where $\mathbf{m}^\ell$ is as in (17).

Proof We first prove that $\mathbf{m}^\ell \ge \widetilde{\mathbf{m}}^\ell$.

To this end, let $\ell \in (0,+\infty)$ be fixed and $\eta \in (0,1/2)$ be arbitrary; using the expression of $\mathbf{m}_z^\ell$ in (40) we choose $z_\eta \in [0,1)$ and $v_\eta \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$ such that $v_\eta(\frac{z_\eta}{\ell}) = 0$, $v_\eta(\pm\infty) = 1$ and

$$\int_{\mathbb{R}} \big((1-v_\eta)^2 + \varphi(\ell x)(v_\eta')^2\big)\,dx \le \mathbf{m}^\ell + \eta\,. \tag{41}$$

Similarly as in Lemma 1 we can find $T_\eta^1, T_\eta^2, S_\eta^1, S_\eta^2$ with $T_\eta^1 < S_\eta^1 < \frac{z_\eta}{\ell} < S_\eta^2 < T_\eta^2$ satisfying the following conditions:

$$v_\eta(T_\eta^1) = v_\eta(T_\eta^2) = 1-\eta \quad \text{and} \quad v_\eta \le 1-\eta \ \text{ on } [T_\eta^1, T_\eta^2] \tag{42}$$
$$v_\eta(S_\eta^1) = v_\eta(S_\eta^2) = \eta^2 \quad \text{and} \quad v_\eta \le \eta^2 \ \text{ on } [S_\eta^1, S_\eta^2]\,. \tag{43}$$

We then define a pair $(u_\eta, v_\eta) \in W^{1,2}_{\mathrm{loc}}(\mathbb{R}) \times W^{1,2}_{\mathrm{loc}}(\mathbb{R})$ with $(u_\eta, v_\eta)(-\infty) = (0,1)$ and $(u_\eta, v_\eta)(+\infty) = (1,1)$ by setting

$$u_\eta(x) := \begin{cases} 0 & \text{if } x < S_\eta^1\,, \\[1ex] \dfrac{x - S_\eta^1}{S_\eta^2 - S_\eta^1} & \text{if } S_\eta^1 \le x \le S_\eta^2\,, \\[1ex] 1 & \text{if } x > S_\eta^2\,, \end{cases}$$

$$w_\eta(x) := \begin{cases} 1-(\eta+\eta^2)(x-T^1_\eta+1) & \text{if } T^1_\eta - 1 \le x < T^1_\eta\,, \\ \max\{0, v_\eta(x)-\eta^2\} & \text{if } T^1_\eta \le x \le T^2_\eta\,, \\ 1+(\eta+\eta^2)(x-T^2_\eta-1) & \text{if } T^2_\eta < x \le T^2_\eta + 1\,, \\ 1 & \text{otherwise in } \mathbb{R}\,. \end{cases}$$

Clearly, $u_\eta \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$, while (42) ensures that also $w_\eta \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$. Moreover, the second condition in (43) implies that $w_\eta \equiv 0$ on $[S^1_\eta, S^2_\eta]$, hence $w_\eta\, u'_\eta = 0$ a.e. in $\mathbb{R}$. In particular, w_η is admissible for $\widetilde{\mathbf{m}}^\ell$. Then it only remains to estimate its energy. This can be done arguing in a similar way as in Lemma 1. Namely, by repeating the computation in (36)–(37) now replacing t with η^2 leads to

$$\begin{aligned} \widetilde{\mathbf{m}}^\ell &\le \int_{\mathbb{R}} \big((1-w_\eta)^2 + \varphi(\ell x)(w'_\eta)^2\big)\,dx \\ &\le (1+\eta)\int_{\mathbb{R}} \big((1-v_\eta)^2 + \varphi(\ell x)(v'_\eta)^2\big)\,dx + \Big(1+\frac{1}{\eta}\Big)\eta^4(T^2_\eta - T^1_\eta) \\ &\qquad + 2(\eta+\eta^2)^2\Big(\frac{1}{3}+\beta\Big)\,. \end{aligned} \tag{44}$$

Moreover, as in (34), we deduce from (41) and (42) that $T^2_\eta - T^1_\eta \le \frac{2\sqrt{\beta}+\eta}{\eta^2}$. Inserting the latter in (44) and appealing to (41) yield

$$\widetilde{\mathbf{m}}^\ell \le (1+\eta)(\mathbf{m}^\ell + \eta) + 2(\eta+\eta^2)\Big(\sqrt{\beta}+\eta+\frac{1}{3}+\beta\Big)\,,$$

hence the desired inequality follows by the arbitrariness of $\eta > 0$.

We now show that $\mathbf{m}^\ell \le \widetilde{\mathbf{m}}^\ell$.

Let v be admissible for $\widetilde{\mathbf{m}}^\ell$; then there exist $u \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$ with $u(-\infty) = 0$, $u(+\infty) = 1$, and $v\,u' = 0$ a.e. in $\mathbb{R}$. Since $u \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$, the boundary conditions at $\pm\infty$ imply that u' cannot be equal to zero a.e. in $\mathbb{R}$. Since at the same time $v\,u' = 0$ a.e. in $\mathbb{R}$, we can find $\bar z \in \mathbb{R}$ with $v(\bar z) = 0$. Set $z := \ell\bar z - \lfloor \ell\bar z \rfloor \in [0,1)$ and $v_z := v(\cdot + (\bar z - \frac{z}{\ell}))$. Then $v_z(\frac{z}{\ell}) = 0$ and $v(\pm\infty) = 1$, while the 1-periodicity of φ together with the fact that $\ell\bar z - z = \lfloor \ell\bar z \rfloor \in \mathbb{Z}$ implies that

$$\begin{aligned} \int_{\mathbb{R}} \big((1-v)^2 + \varphi(\ell x)(v')^2\big)\,dx &= \int_{\mathbb{R}} \Big((1-v_z)^2 + \varphi\Big(\ell\Big(x+\bar z - \frac{z}{\ell}\Big)\Big)(v'_z)^2\Big)\,dx \\ &= \int_{\mathbb{R}} \big((1-v_z)^2 + \varphi(\ell x)(v'_z)^2\big)\,dx\,. \end{aligned}$$

Thus we conclude by passing to the infimum in v. □

Remark 5 The proof of Proposition 2 actually shows that

$$\mathbf{m}^\ell = \inf_{T>0} \inf \bigg\{ \int_{-T}^{T} \big((1-v)^2 + \varphi(\ell x)(v')^2\big)\, dx \colon v \in W^{1,2}(-T,T)\,,\ 0 \le v \le 1\,, \\ v(\pm T) = 1\,,\ \exists\, u \in W^{1,2}(-T,T) \text{ with } u(-T)=0,\ u(T)=1, \\ \text{and } v\,u' = 0 \text{ a.e. in } (-T,T) \bigg\}.$$

4 Oscillations on a Larger Scale than the Singular Perturbation

In this section we analyse the case when the oscillation parameter δ_ε is much larger than the singular-perturbation parameter ε; i.e., the case $\ell = 0$.

Throughout this section the function φ is additionally assumed to be upper semicontinuous.

Proposition 3 *Let $\ell = 0$ and assume that φ is upper semicontinuous; then the sequence (F_ε) defined in (8) Γ-converges to the functional $F^0 \colon L^1(a,b) \times L^1(a,b) \longrightarrow [0,+\infty]$ defined as*

$$F^0(u,v) := \begin{cases} \displaystyle\int_a^b (u')^2\, dx + \mathbf{m}^0 \# S(u) & u \in P\text{-}W^{1,2}(a,b)\,,\ v = 1 \text{ a.e. in } (a,b)\,, \\ +\infty & \text{otherwise}\,, \end{cases} \tag{45}$$

where $\mathbf{m}^0 := 2\sqrt{\alpha}$.

Proof Thanks to (9), from Theorem 2 we immediately deduce that

$$\Gamma\text{-}\liminf_{\varepsilon\to 0} F_\varepsilon(u,v) \ge \Gamma\text{-}\liminf_{\varepsilon\to 0} AT_\varepsilon^\alpha(u,v) = \int_a^b (u')^2\, dx + 2\sqrt{\alpha}\# S(u)\,,$$

which by definition of $\mathbf{m}^0$ gives the lower-bound inequality. It thus remains to establish the upper-bound inequality.

Let $u \in P\text{-}W^{1,2}(a,b)$; we construct a sequence $(u_\varepsilon, v_\varepsilon) \subset W^{1,2}(a,b) \times W^{1,2}(a,b)$ such that $(u_\varepsilon, v_\varepsilon) \to (u,1)$ in $L^1(a,b) \times L^1(a,b)$ and

$$\limsup_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon, v_\varepsilon) \le \mathbf{m}^0 \# S(u)\,.$$

Since the construction of the recovery sequence $(u_\varepsilon, v_\varepsilon)$ will be performed locally, close to a discontinuity point of u, we can assume without loss of generality that $S(u) = \{t_0\}$, with $t_0 \in (a,b)$.

Let now $\tilde{u} \in W^{1,2}(a,b)$ and $u^{\mathrm{pc}} \in PC(a,b)$ be as in (14); without loss of generality, we choose $u^{\mathrm{pc}} = s\chi_{(a,t_0)}$, with $s \in \mathbb{R}$.

For $\eta > 0$ let $y_\eta \in (0,1)$ satisfy

$$\varphi(y_\eta) \le \alpha + \eta\,. \tag{46}$$

Applying (26) with $\lambda = \alpha$ we find $T_\eta > 0$ and $v_\eta \in W^{1,2}(0,T_\eta)$ such that $0 \le v_\eta \le 1$, $v_\eta(0) = 0$, $v_\eta(T_\eta) = 1$, and

$$\int_0^{T_\eta} \big((1-v_\eta)^2 + \alpha\,(v_\eta')^2\big)\,dx \le \sqrt{\alpha} + \eta\,. \tag{47}$$

Finally, set

$$t_0^\varepsilon := \left\lfloor \frac{t_0}{\delta_\varepsilon} \right\rfloor \delta_\varepsilon\,, \quad y_\eta^\varepsilon := \delta_\varepsilon y_\eta\,, \tag{48}$$

and let $\xi_\varepsilon > 0$ be such that $\xi_\varepsilon \ll \varepsilon$. Then, a recovery sequence for $F^0(u,1)$ is defined as $(u_\varepsilon, v_\varepsilon) = (\tilde{u} + \bar{u}_\varepsilon, v_\varepsilon)$ with $(\bar{u}_\varepsilon, v_\varepsilon) \subset W^{1,2}(a,b) \times W^{1,2}(a,b)$ given by

$$\bar{u}_\varepsilon(x) := \begin{cases} 0 & \text{if } x \le t_0^\varepsilon + y_\eta^\varepsilon + \dfrac{\xi_\varepsilon}{2}\,, \\[2ex] \dfrac{2s}{\xi_\varepsilon}\Big(x - \Big(t_0^\varepsilon + y_\eta^\varepsilon + \dfrac{\xi_\varepsilon}{2}\Big)\Big) & \text{if } t_0^\varepsilon + y_\eta^\varepsilon + \dfrac{\xi_\varepsilon}{2} < x < t_0^\varepsilon + y_\eta^\varepsilon + \xi_\varepsilon\,, \\[2ex] s & \text{if } x \ge t_0^\varepsilon + y_\eta^\varepsilon + \xi_\varepsilon \end{cases}$$

and

$$v_\varepsilon(x) := \begin{cases} 0 & \text{if } |x - t_0^\varepsilon - y_\eta^\varepsilon| \le \xi_\varepsilon\,, \\[2ex] v_\eta\Big(\dfrac{|x - t_0^\varepsilon - y_\eta^\varepsilon| - \xi_\varepsilon}{\varepsilon}\Big) & \text{if } \xi_\varepsilon < |x - t_0^\varepsilon - y_\eta^\varepsilon| \le \xi_\varepsilon + \varepsilon T_\eta\,, \\[2ex] 1 & \text{if } |x - t_0^\varepsilon - y_\eta^\varepsilon| > \xi_\varepsilon + \varepsilon T_\eta\,, \end{cases}$$

(see Fig. 2).

We notice that since $t_0^\varepsilon + y_\eta^\varepsilon \to t_0$, then by construction $u_\varepsilon := \tilde{u} + \bar{u}_\varepsilon \to u$ in $L^1(a,b)$, further, $v_\varepsilon \to 1$ in $L^1(a,b)$ and a.e. in (a,b). Therefore, it remains to show that

$$\limsup_{\varepsilon\to 0} G_\varepsilon(v_\varepsilon,(a,b)) \le \int_a^b (u')^2\,dx + \mathbf{m}^0\,. \tag{49}$$

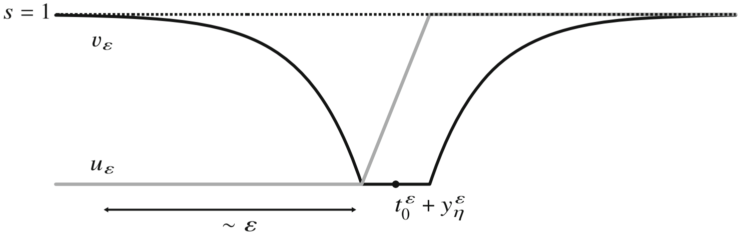


Fig. 2 Recovery sequence in the case $s = 1$

We start noticing that by construction

$$v_\varepsilon \bar{u}'_\varepsilon = 0 \ \text{ a.e.in } (a, b)$$

and, therefore,

$$\lim_{\varepsilon\to 0} \int_a^b v_\varepsilon^2 (u'_\varepsilon)^2 \, dx = \lim_{\varepsilon\to 0} \int_a^b v_\varepsilon^2 (\tilde{u}')^2 \, dx = \int_a^b (\tilde{u}')^2 \, dx \,, \tag{50}$$

the last equality following by the Dominated Convergence Theorem, since $\tilde{u}' \in L^2(a, b)$ and $0 \le v_\varepsilon \le 1$.

Moreover, by definition of v_ε we also have

$$G_\varepsilon(v_\varepsilon, (a, b)) \le 2 \int_{t_0^\varepsilon + y_\eta^\varepsilon + \xi_\varepsilon}^{t_0^\varepsilon + y_\eta^\varepsilon + \xi_\varepsilon + \varepsilon T_\eta} \left(\frac{(1 - v_\varepsilon)^2}{\varepsilon} + \varepsilon \varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v'_\varepsilon)^2 \right) dx + \frac{2\xi_\varepsilon}{\varepsilon} \,. \tag{51}$$

Since $\xi_\varepsilon \ll \varepsilon$, in (51) it only remains to estimate the integral on the right-hand side.

By a change of variables, recalling (48), and using the periodicity of φ we readily obtain

$$\begin{aligned} &\int_{t_0^\varepsilon + y_\eta^\varepsilon + \xi_\varepsilon}^{t_0^\varepsilon + y_\eta^\varepsilon + \xi_\varepsilon + \varepsilon T_\eta} \left(\frac{(1 - v_\varepsilon)^2}{\varepsilon} + \varepsilon \varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v'_\varepsilon)^2 \right) dx \\ &\quad = \int_0^{T_\eta} \left((1 - v_\eta(x))^2 + \varphi\Big(\frac{\varepsilon}{\delta_\varepsilon} x + y_\eta + \frac{\xi_\varepsilon}{\delta_\varepsilon}\Big)(v'_\eta(x))^2 \right) dx \,. \end{aligned} \tag{52}$$

Since φ is upper semicontinuous and $\xi_\varepsilon \ll \varepsilon \ll \delta_\varepsilon$, applying the reverse Fatou Lemma we infer

$$\limsup_{\varepsilon\to 0} \int_0^{T_\eta} \varphi\Big(\frac{\varepsilon}{\delta_\varepsilon} x + y_\eta + \frac{\xi_\varepsilon}{\delta_\varepsilon}\Big)(v'_\eta(x))^2 \, dx \le \int_0^{T_\eta} \varphi(y_\eta)(v'_\eta(x))^2 \, dx \,. \tag{53}$$

Therefore, gathering (51), (52), (53), and recalling the definition of y_η and v_η we get

$$
\begin{aligned}
\limsup_{\varepsilon\to 0} G_\varepsilon(v_\varepsilon,(a,b)) &\le 2\int_0^{T_\eta}\big((1-v_\eta)^2+\varphi(y_\eta)(v_\eta')^2\big)\,dx\\
&\le 2\int_0^{T_\eta}\big((1-v_\eta)^2+(\alpha+\eta)(v_\eta')^2\big)\,dx\\
&\le 2\Big(1+\frac{\eta}{\alpha}\Big)(\sqrt{\alpha}+\eta)=\Big(1+\frac{\eta}{\alpha}\Big)(\mathbf{m}^0+2\eta)\,.
\end{aligned}
\tag{54}
$$

Eventually, (49) follows by combining (50), (54) and letting $\eta \to 0$. □

Remark 6 We observe that in the proof of Proposition 3 the upper semicontinuity of φ is only needed to obtain the upper-bound inequality.

5 Oscillations on the Same Scale as the Singular Perturbation

In this section we analyse the case when the oscillation parameter δ_ε and the singular-perturbation parameter ε are of the same order; i.e., the case $\ell \in (0,+\infty)$.

On account of Lemma 1 and Proposition 2 we prove the following result.

Proposition 4 *Let $\ell \in (0,+\infty)$; then the sequence (F_ε) defined in (8) Γ-converges to the functional $F^\ell\colon L^1(a,b)\times L^1(a,b)\longrightarrow[0,+\infty]$ defined as*

$$
F^\ell(u,v) := \begin{cases} \displaystyle\int_a^b (u')^2\,dx + \mathbf{m}^\ell \# S(u) & u \in P\text{-}W^{1,2}(a,b)\,,\ v=1 \text{ a.e. in } (a,b)\,,\\ +\infty & \text{otherwise,}\end{cases}
$$

where $\mathbf{m}^\ell$ is as in (17).

Proof We prove separately the lower-bound and the upper-bound inequalities.

Step 1: Lower-bound inequality.

Let $(u,v) \in L^1(a,b)\times L^1(a,b)$ be arbitrary and let $(u_\varepsilon,v_\varepsilon)\subset W^{1,2}(a,b)\times W^{1,2}(a,b)$ be such that

$$
(u_\varepsilon,v_\varepsilon)\to(u,v) \ \text{ in } \ L^1(a,b)\times L^1(a,b) \quad \text{and} \quad \liminf_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon,v_\varepsilon) < +\infty\,.
$$

Then, up to subsequences (not relabelled) we can additionally assume that we have $\sup_{\varepsilon>0} F_\varepsilon(u_\varepsilon,v_\varepsilon) < +\infty$; therefore, Proposition 1 immediately yields that $u \in P\text{-}W^{1,2}(a,b)$, $v=1$ a.e. in (a,b) and

$$
\liminf_{\varepsilon\to 0}\int_a^b v_\varepsilon^2(u_\varepsilon')^2\,dx \ge \int_a^b (u')^2\,dx\,. \tag{55}
$$

Therefore, to prove the liminf inequality it suffices to show that

$$\liminf_{\varepsilon\to 0}\int_a^b \left(\frac{(1-v_\varepsilon)^2}{\varepsilon}+\varepsilon\,\varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v_\varepsilon')^2\right)dx \ge \mathbf{m}^\ell \# S(u)\,,$$

with $\mathbf{m}^\ell$ as in (17).

To this end we notice that if $S(u)=\emptyset$, then there is nothing to prove. Hence, we may assume that $S(u)=\{t_1,\dots,t_N\}$, with $N\ge 1$. Now, let $I_1,\dots,I_N$ be pairwise disjoint open intervals with $I_i\subset(a,b)$ and $t_i\in I_i$, for every $i=1,\dots,N$. We claim that

$$\liminf_{\varepsilon\to 0} G_\varepsilon(v_\varepsilon, I_i)\ge \mathbf{m}^\ell\,, \tag{56}$$

for every $i=1,\dots,N$, where G_ε is as in (12).

To prove the claim, we let $i\in\{1,\dots,N\}$ be arbitrary, and we invoke Proposition 1 (2) and Remark 1 to find $s^i_\varepsilon, r^i, \tilde r^i\in I_i$ with $r^i<s^i_\varepsilon<\tilde r^i$ satisfying

$$\lim_{\varepsilon\to 0} v_\varepsilon(s^i_\varepsilon)=0 \quad\text{and}\quad \lim_{\varepsilon\to 0} v_\varepsilon(r^i)=\lim_{\varepsilon\to 0} v_\varepsilon(\tilde r^i)=1\,.$$

Set $z^i_\varepsilon := \frac{s^i_\varepsilon}{\delta_\varepsilon}-\Big\lfloor\frac{s^i_\varepsilon}{\delta_\varepsilon}\Big\rfloor\in[0,1)$; thanks to the 1-periodicity of φ, the change of variables $y=\frac{x-s^i_\varepsilon}{\ell\delta_\varepsilon}$ yields

$$\begin{aligned} G_\varepsilon(v_\varepsilon, I_i) &\ge \int_{r^i}^{\tilde r^i}\left(\frac{(1-v_\varepsilon)^2}{\varepsilon}+\varepsilon\varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v_\varepsilon')^2\right)dx\\ &= \int_{\frac{r^i-s^i_\varepsilon}{\ell\delta_\varepsilon}}^{\frac{\tilde r^i-s^i_\varepsilon}{\ell\delta_\varepsilon}}\left(\frac{\ell\delta_\varepsilon}{\varepsilon}(1-w_\varepsilon)^2+\frac{\varepsilon}{\ell\delta_\varepsilon}\varphi(\ell y+z^i_\varepsilon)(w_\varepsilon')^2\right)dy\\ &\ge \gamma_\varepsilon\int_{\frac{r^i-s^i_\varepsilon}{\ell\delta_\varepsilon}}^{\frac{\tilde r^i-s^i_\varepsilon}{\ell\delta_\varepsilon}}\Big((1-w_\varepsilon)^2+\varphi(\ell y+z^i_\varepsilon)(w_\varepsilon')^2\Big)dy\,, \end{aligned} \tag{57}$$

where $w_\varepsilon(y)=v_\varepsilon(\ell\delta_\varepsilon y+s^i_\varepsilon)$ and

$$\gamma_\varepsilon := \min\left\{\frac{\ell\delta_\varepsilon}{\varepsilon},\frac{\varepsilon}{\ell\delta_\varepsilon}\right\}\to 1 \quad\text{as}\ \ \varepsilon\to 0\,. \tag{58}$$

Since $v_\varepsilon(r^i), v_\varepsilon(\tilde r^i)\to 1$, using a linear interpolation as in the proof of Lemma 1 we can extend w_ε to a function $w^i_\varepsilon\in W^{1,2}_{\rm loc}(\mathbb{R})$ with $0\le w^i_\varepsilon\le 1$ satisfying $w^i_\varepsilon(0)=v_\varepsilon(s^i_\varepsilon)$, $w_\varepsilon(\pm\infty)=1$ and such that

$$\begin{aligned}\int_{\frac{r^i - s^i_\varepsilon}{\ell\delta_\varepsilon}}^{\frac{\tilde r^i - s^i_\varepsilon}{\ell\delta_\varepsilon}} &\Big((1-w_\varepsilon)^2 + \varphi(\ell y + z^i_\varepsilon)(w'_\varepsilon)^2\Big)\,dy \\ &= \int_{\mathbb{R}} \Big((1-w^i_\varepsilon)^2 + \varphi(\ell y + z^i_\varepsilon)\big((w^i_\varepsilon)'\big)^2\Big)\,dy + o_\varepsilon(1)\,,\end{aligned} \tag{59}$$

as $\varepsilon \to 0$. Thus, since w^i_ε is admissible for $\mathbf{m}^\ell_{z^i_\varepsilon}(v_\varepsilon(s^i_\varepsilon)) \geq \mathbf{m}^\ell(v_\varepsilon(s^i_\varepsilon))$ and $v_\varepsilon(s^i_\varepsilon) \to 0$, gathering (57)–(59), passing to the liminf in ε and applying Lemma 1 yields

$$\liminf_{\varepsilon\to 0} G_\varepsilon(v_\varepsilon, I_i) \geq \lim_{\varepsilon\to 0}\gamma_\varepsilon\, \liminf_{\varepsilon\to 0}\mathbf{m}^\ell(v_\varepsilon(s^i_\varepsilon)) = \mathbf{m}^\ell\,,$$

hence (56). Eventually, summing over i we get

$$\liminf_{\varepsilon\to 0} G_\varepsilon(v_\varepsilon, (a,b)) \geq \sum_{i=1}^N \liminf_{\varepsilon\to 0} G_\varepsilon(v_\varepsilon, I_i) \geq \mathbf{m}^\ell N = \mathbf{m}^\ell \# S(u)\,,$$

which together with (55) gives the lower-bound inequality.

Step 2: Upper-bound inequality

As in the proof of Proposition 3 it suffices to construct a recovery sequence for $u = \tilde u + u^{\mathrm{pc}}$ with $\tilde u \in W^{1,2}(a,b)$ and $u^{\mathrm{pc}} = s\chi_{(a,t_0)}$, with $s \in \mathbb{R}$ and $t_0 \in (a,b)$. To this end, we fix $\eta > 0$ and according to Proposition 2 and Remark 5 we choose $T_\eta > 0$ and $(u_\eta, v_\eta) \in W^{1,2}(-T_\eta, T_\eta) \times W^{1,2}(-T_\eta, T_\eta)$ with $0 \leq v_\eta \leq 1$ satisfying $(u_\eta, v_\eta)(-T_\eta) = (0,1)$, $(u_\eta, v_\eta)(T_\eta) = (1,1)$, and $v_\eta\, u'_\eta = 0$ a.e. in $(-T_\eta, T_\eta)$ and

$$\int_{-T_\eta}^{T_\eta} \Big((1-v_\eta)^2 + \varphi(\ell x)(v'_\eta)^2\Big)\,dx \leq \mathbf{m}^\ell + \eta\,. \tag{60}$$

We extend (u_η, v_η) to $\mathbb{R}$ by setting $(u_\eta, v_\eta) := (\chi_{(0,+\infty)}, 1)$ in $\mathbb{R} \setminus (-T_\eta, T_\eta)$. Moreover, we set $t^\varepsilon_0 := \lfloor \frac{t_0}{\delta_\varepsilon} \rfloor \delta_\varepsilon$ and define the pairs $(u_\varepsilon, v_\varepsilon) := (s\bar u_\varepsilon + \tilde u, v_\varepsilon)$ with $(\bar u_\varepsilon, v_\varepsilon)$ given by

$$\bar u_\varepsilon(x) := u_\eta\Big(\frac{x - t^\varepsilon_0}{\ell\delta_\varepsilon}\Big) \quad\text{and}\quad v_\varepsilon(x) := v_\eta\Big(\frac{x - t^\varepsilon_0}{\ell\delta_\varepsilon}\Big).$$

By construction $u_\varepsilon \to u^{\mathrm{pc}} + \tilde u = u$ in $L^1(a,b)$, while $v_\varepsilon \to 1$ in $L^1(a,b)$ and a.e. in (a,b). It thus remains to estimate $F_\varepsilon(u_\varepsilon, v_\varepsilon)$. Since $v_\varepsilon u'_\varepsilon = 0$ a.e. in (a,b), as in (50) we deduce that

$$\lim_{\varepsilon\to 0}\int_a^b v^2_\varepsilon (u'_\varepsilon)^2\,dx = \lim_{\varepsilon\to 0}\int_a^b v^2_\varepsilon(\tilde u')^2\,dx = \int_a^b (\tilde u')^2\,dx\,. \tag{61}$$

Therefore, we are left to estimate $G_\varepsilon(v_\varepsilon, (a,b))$. By the choice of t^ε_0 and the 1-periodicity of φ, a change of variables yields

$$\begin{aligned} G_\varepsilon(v_\varepsilon,(a,b)) &= \int_{t_0^\varepsilon-\ell\delta_\varepsilon T_\eta}^{t_0^\varepsilon+\ell\delta_\varepsilon T_\eta} \left(\frac{(1-v_\varepsilon)^2}{\varepsilon}+\varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v_\varepsilon')^2\right)dx \\ &= \int_{-T_\eta}^{T_\eta}\left(\frac{\ell\delta_\varepsilon}{\varepsilon}(1-v_\eta)^2+\frac{\varepsilon}{\ell\delta_\varepsilon}\varphi(\ell x)(v_\eta')^2\right)dx \qquad (62) \\ &\le \widetilde{\gamma}_\varepsilon \int_{-T_\eta}^{T_\eta}\Big((1-v_\eta)^2+\varphi(\ell x)(v_\eta')^2\Big)\,dx\,, \end{aligned}$$

where

$$\widetilde{\gamma}_\varepsilon := \max\left\{\frac{\ell\delta_\varepsilon}{\varepsilon},\frac{\varepsilon}{\ell\delta_\varepsilon}\right\}\to 1 \quad \text{as } \varepsilon\to 0\,. \tag{63}$$

Using (60) and gathering (61)–(63) we readily obtain

$$\limsup_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon,v_\varepsilon) \le \int_a^b (u')^2\,dx + \mathbf{m}^\ell + \eta\,,$$

hence the upper-bound inequality follows by the arbitrariness of $\eta > 0$. □

6 Oscillations on a Smaller Scale than the Singular Perturbation

In this section we analyse the case when the oscillation δ_ε parameter is much smaller than the singular-perturbation parameter ε; i.e., the case $\ell = +\infty$.

Proposition 5 *Let $\ell = \infty$; then the sequence (F_ε) defined in (8) Γ-converges to the functional $F^\infty\colon L^1(a,b)\times L^1(a,b)\to[0,+\infty]$ defined as*

$$F^\infty(u,v) := \begin{cases} \displaystyle\int_a^b (u')^2\,dx + \mathbf{m}^\infty \# S(u) & u\in P\text{-}W^{1,2}(a,b),\ v=1 \text{ a.e. in } (a,b)\,, \\ +\infty & \text{otherwise}, \end{cases} \tag{64}$$

where

$$\mathbf{m}^\infty := 2\left(\int_0^1 \frac{1}{\varphi(t)}\,dt\right)^{-1/2}. \tag{65}$$

Proof It is convenient to introduce the constant

$$\varphi_{\mathrm{hom}} := \left(\int_0^1 \frac{1}{\varphi(t)}\,dt\right)^{-1},$$

so that $\mathbf{m}^\infty = 2\sqrt{\varphi_{\mathrm{hom}}}$. We now divide the proof into two steps.

Step 1: Lower-bound inequality.

For any $(u, v) \in L^1(a, b) \times L^1(a, b)$ let $(u_\varepsilon, v_\varepsilon) \subset W^{1,2}(a, b) \times W^{1,2}(a, b)$ be such that

$$(u_\varepsilon, v_\varepsilon) \to (u, v) \text{ in } L^1(a, b) \times L^1(a, b) \quad \text{and} \quad \liminf_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon, v_\varepsilon) < +\infty\,.$$

Arguing as in Proposition 4 we assume without loss of generality that we have $\sup_{\varepsilon>0} F_\varepsilon(u_\varepsilon, v_\varepsilon) < +\infty$ and we apply Proposition 1 to deduce that $u \in P\text{-}W^{1,2}(a, b)$, $v = 1$ a.e. in (a, b) and

$$\liminf_{\varepsilon\to 0} \int_a^b v_\varepsilon^2 (u_\varepsilon')^2\,dx \geq \int_a^b (u')^2\,dx\,.$$

We set $S(u) = \{t_1, \dots, t_N\}$ with $N \geq 1$ (if $S(u) = \emptyset$ there is nothing to prove) and we let $I_1, \dots, I_N$ be pairwise disjoint open intervals with $I_i \subset (a, b)$ and $t_i \in I_i$ for $i = 1, \dots, N$. Then if we show that

$$\liminf_{\varepsilon\to 0} G_\varepsilon(v_\varepsilon, I_i) \geq \mathbf{m}^\infty \quad \text{for every } i = 1, \dots, N, \tag{66}$$

with $\mathbf{m}^\infty$ as in (65) we are done.

We fix $\eta > 0$ and $i \in \{1, \dots, N\}$. By Proposition 1 and Remark 1 we can find $\tilde{s}_\varepsilon^i, r_\varepsilon^i, \tilde{r}_\varepsilon^i \in I_i$ with $r_\varepsilon^i < \tilde{s}_\varepsilon^i < \tilde{r}_\varepsilon^i$ such that

$$v_\varepsilon(\tilde{s}_\varepsilon^i) = \eta\,,\ v_\varepsilon(r_\varepsilon^i) = v_\varepsilon(\tilde{r}_\varepsilon^i) = 1 - \eta \quad \text{and} \quad v_\varepsilon \leq 1 - \eta \text{ in } [r_\varepsilon^i, \tilde{r}_\varepsilon^i]\,. \tag{67}$$

Then (25) implies that

$$\frac{\tilde{r}_\varepsilon^i - \tilde{s}_\varepsilon^i}{\varepsilon} \in \left[\frac{\alpha(1-2\eta)^2}{M}, \frac{M}{\eta^2}\right] \quad \text{for every } \varepsilon > 0\,, \tag{68}$$

where $M := \sup_{\varepsilon>0} F_\varepsilon(u_\varepsilon, v_\varepsilon) < +\infty$. Thanks to (67) the function $\tilde{v}_\varepsilon : \mathbb{R} \to [0, 1]$ given by

$$\tilde v_\varepsilon(x) := \begin{cases} \eta & \text{if } x < \tilde s^i_\varepsilon\,, \\ v_\varepsilon(x) & \text{if } \tilde s^i_\varepsilon \le x \le \tilde r^i_\varepsilon\,, \\ (1-\eta) + \eta \dfrac{x - r^i_\varepsilon}{\varepsilon} & \text{if } \tilde r^i_\varepsilon < x \le \tilde r^i_\varepsilon + \varepsilon\,, \\ 1 & \text{otherwise in } \mathbb{R} \end{cases} \tag{69}$$

belongs to $W^{1,2}_{\mathrm{loc}}(\mathbb{R})$. Moreover, set $t^i_\varepsilon := \left\lfloor \frac{\tilde s^i_\varepsilon}{\delta_\varepsilon} \right\rfloor \delta_\varepsilon \in (\tilde s^i_\varepsilon - \delta_\varepsilon, \tilde s^i_\varepsilon]$; using (7), by the definition of $\tilde v_\varepsilon$ we have

$$G_\varepsilon(v_\varepsilon, (\tilde s^i_\varepsilon, \tilde r^i_\varepsilon)) \ge G_\varepsilon(\tilde v_\varepsilon, (t^i_\varepsilon, \tilde r^i_\varepsilon + \varepsilon)) - \left(\frac{1}{3} + \beta\right)\eta^2 - (1-\eta)^2 \frac{\delta_\varepsilon}{\varepsilon}\,. \tag{70}$$

Eventually, by setting $w_\varepsilon(x) := \tilde v_\varepsilon(\varepsilon x + t^i_\varepsilon)$ and $T_\eta := \frac{M}{\eta^2} + 2$, the periodicity of φ, (68) and a change of variables yield

$$\begin{aligned} G_\varepsilon(\tilde v_\varepsilon, (t^i_\varepsilon, \tilde r^i_\varepsilon + \varepsilon)) &= \int_0^{T_\eta} \left((1 - w_\varepsilon)^2 + \varphi\Big(\frac{\varepsilon x}{\delta_\varepsilon}\Big)(w_\varepsilon')^2 \right) dx \\ &\ge \inf \left\{ \int_0^{T_\eta} \left((1-w)^2 + \varphi\Big(\frac{x}{\delta_\varepsilon/\varepsilon}\Big)(w')^2 \right) dx \colon w \in W^{1,2}(0, T_\eta)\,, \right. \\ &\qquad\qquad \left. w(0) = \eta\,,\ w(T_\eta) = 1 \right\}. \end{aligned} \tag{71}$$

Now, since $\delta_\varepsilon/\varepsilon \to 0$, by classical homogenisation (see, e.g., [7, Theorem 3.1]) we get

$$\begin{aligned} &\lim_{\varepsilon\to 0} \inf \left\{ \int_0^{T_\eta} \left((1-w)^2 + \varphi\Big(\frac{x}{\delta_\varepsilon/\varepsilon}\Big)(w')^2 \right) dx \colon w \in W^{1,2}(0, T_\eta)\,, \right. \\ &\qquad\qquad \left. w(0) = \eta\,,\ w(T_\eta) = 1 \right\} \\ &\quad = \min \left\{ \int_0^{T_\eta} \left((1-w)^2 + \varphi_{\mathrm{hom}}(w')^2 \right) dx \colon w \in W^{1,2}(0, T_\eta)\,, \right. \\ &\qquad\qquad \left. w(0) = \eta\,,\ w(T_\eta) = 1 \right\}. \end{aligned} \tag{72}$$

For any $w \in W^{1,2}(0, T_\eta)$ satisfying $w(0) = \eta$ and $w(T_\eta) = 1$, an application of the Modica–Mortola trick together with a change of variables yields

$$\begin{aligned}\int_0^{T_\eta}\Big((1-w)^2+\varphi_{\rm hom}(w')^2\Big)\,dx &\geq 2\sqrt{\varphi_{\rm hom}}\int_0^{T_\eta}(1-w)|w'|\,dx\\ &=2\sqrt{\varphi_{\rm hom}}\int_\eta^1(1-s)=(1-\eta)^2\sqrt{\varphi_{\rm hom}}\,,\end{aligned}$$

hence

$$\begin{aligned}&\min\bigg\{\int_0^{T_\eta}\Big((1-w)^2+\varphi_{\rm hom}(w')^2\Big)\,dx\colon w\in W^{1,2}(0,T_\eta)\,,\ w(0)=\eta\,,\ w(T_\eta)=1\bigg\}\\ &\geq(1-\eta)^2\sqrt{\varphi_{\rm hom}}\,.\end{aligned} \tag{73}$$

Finally, gathering together (70)–(73) we obtain

$$\liminf_{\varepsilon\to0}G_\varepsilon(v_\varepsilon,(\tilde{s}^i_\varepsilon,\tilde{r}^i_\varepsilon))\geq(1-\eta)^2\sqrt{\varphi_{\rm hom}}-\eta^2\bigg(\frac{1}{3}+\beta\bigg)\,. \tag{74}$$

Analogously it can be shown that

$$\liminf_{\varepsilon\to0}G_\varepsilon(v_\varepsilon,(r^i_\varepsilon,\tilde{s}^i_\varepsilon))\geq(1-\eta)^2\sqrt{\varphi_{\rm hom}}-\eta^2\bigg(\frac{1}{3}+\beta\bigg)\,. \tag{75}$$

Hence, from (74) and (75) we deduce that

$$\liminf_{\varepsilon\to0}G_\varepsilon(v_\varepsilon,I_i)\geq(1-\eta)^2\mathbf{m}^\infty-\eta^2\bigg(\frac{1}{3}+\beta\bigg)\,.$$

Eventually, by letting $\eta\to0$ we obtain (66) and, therefore, the lower bound.

Step 2: Upper-bound inequality.

Let $u\in P\text{-}W^{1,2}(a,b)$ be fixed; As in the proof of Proposition 3 we assume without loss of generality that $S(u)=\{t_0\}$ for some $t_0\in(a,b)$ and $u=\tilde{u}+u^{\rm pc}$ as in (14) with $\tilde{u}\in W^{1,2}(a,b)$ and $u^{\rm pc}=s\chi_{(a,t_0)}$ for some $s\in\mathbb{R}$.

We fix $\eta>0$; applying (26) with $\lambda=\varphi_{\rm hom}$ we find $T_\eta>0$ and $v_\eta\in W^{1,2}(0,T_\eta)$ satisfying $0\leq v_\eta\leq1$, $v_\eta(0)=0$, $v_\eta(T_\eta)=1$, and

$$\int_0^{T_\eta}(1-v_\eta)^2+\varphi_{\rm hom}(v'_\eta)^2\,dx\leq\sqrt{\varphi_{\rm hom}}+\eta\,. \tag{76}$$

By invoking the classical homogenisation theorem (see, *e.g.*, [9, Theorem 14.5]), for any $\sigma\searrow0$ we find a sequence $(w_\sigma)\subset W^{1,2}(0,T_\eta)$ such that $w_\sigma\to v_\eta$ in $L^2(0,T_\eta)$ as $\sigma\to0$, $w_\sigma(0)=0$, $w_\sigma(T_\eta)=1$ and

$$\lim_{\sigma\to 0}\int_0^{T_\eta}\Big((1-w_\sigma)^2+\varphi\Big(\frac{x}{\sigma}\Big)(w'_\sigma)^2\Big)\,dx=\int_0^{T_\eta}\Big((1-v_\eta)^2+\varphi_{\text{hom}}(v'_\eta)^2\Big)\,dx\,. \tag{77}$$

Now we let t_0^ε be as in (48), $\sigma_\varepsilon := \delta_\varepsilon/\varepsilon$ and $w_\varepsilon := w_{\sigma_\varepsilon}$ and define the pair $(u_\varepsilon, v_\varepsilon) \in W^{1,2}(a,b)\times W^{1,2}(a,b)$ by setting $u_\varepsilon := \tilde{u}+\bar{u}_\varepsilon$ with

$$\bar{u}_\varepsilon(x) := \begin{cases} 0 & \text{if } x \le t_0^\varepsilon + \dfrac{\delta_\varepsilon}{2}\,,\\[2mm] \dfrac{2s}{\delta_\varepsilon}\Big(x-\Big(t_0^\varepsilon+\dfrac{\delta_\varepsilon}{2}\Big)\Big) & \text{if } t_0^\varepsilon+\dfrac{\delta_\varepsilon}{2} < x < t_0^\varepsilon+\delta_\varepsilon\,,\\[2mm] s & \text{if } x \ge t_0^\varepsilon+\delta_\varepsilon\,,\end{cases}$$

and

$$v_\varepsilon(x) := \begin{cases} 0 & \text{if } |x-t_0^\varepsilon| \le \delta_\varepsilon\,,\\[2mm] w_\varepsilon\left(\dfrac{|x-t_0^\varepsilon|-\delta_\varepsilon}{\varepsilon}\right) & \text{if } \delta_\varepsilon < |x-t_0^\varepsilon| < \delta_\varepsilon+\varepsilon\,T_\eta\,,\\[2mm] 1 & \text{if } \delta_\varepsilon+\varepsilon\,T_\eta \le |x-t_0^\varepsilon|\,,\end{cases}$$

(see Fig. 3).

We claim that $(u_\varepsilon, v_\varepsilon)$ is a recovery sequence for $F^\infty(u,1)$. In fact, by construction $u_\varepsilon := \tilde{u}+\bar{u}_\varepsilon \to u$ in $L^1(a,b)$, $v_\varepsilon \to 1$ in $L^1(a,b)$ and a.e. in (a,b). Moreover, observing that

$$v_\varepsilon \bar{u}'_\varepsilon = 0 \text{ a.e. in } (a,b)$$

we get

$$\lim_{\varepsilon\to 0}\int_a^b v_\varepsilon^2(u'_\varepsilon)^2\,dx = \lim_{\varepsilon\to 0}\int_a^b v_\varepsilon^2(\tilde{u}')^2\,dx = \int_a^b (\tilde{u}')^2\,dx = \int_a^b (u')^2\,dx\,. \tag{78}$$

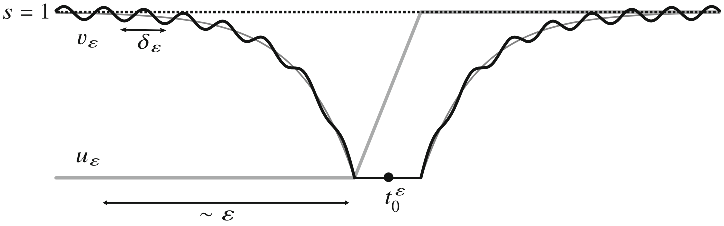


Fig. 3 Recovery sequence in the case $s = 1$; v_ε in dark grey is obtained by superposing oscillations on the rescaled optimal profile

On the other hand, by a change of variables and the periodicity of φ we deduce that

$$\begin{aligned} G_\varepsilon(v_\varepsilon,(a,b)) &\le 2\int_{t_0^\varepsilon+\delta_\varepsilon}^{t_0^\varepsilon+\delta_\varepsilon+\varepsilon T_\eta}\left(\frac{(1-v_\varepsilon)^2}{\varepsilon}+\varepsilon\varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v_\varepsilon')^2\right)dx+\frac{2\delta_\varepsilon}{\varepsilon}\\ &=2\int_0^{T_\eta}\left((1-w_\varepsilon)^2+\varphi\left(\frac{x}{\sigma_\varepsilon}\right)(w_\varepsilon')^2\right)dx+\frac{2\delta_\varepsilon}{\varepsilon}\,. \end{aligned}$$

The latter together with (77) and (76) yield

$$\limsup_{\varepsilon\to 0} G_\varepsilon(v_\varepsilon,(a,b))\le 2\sqrt{\varphi_{\rm hom}}+2\eta=\mathbf{m}^\infty+2\eta\,. \tag{79}$$

Finally, gathering together (78) and (79) we obtain

$$\limsup_{\varepsilon\to 0}\int_a^b\left(v_\varepsilon^2(u_\varepsilon')^2+\frac{(1-v_\varepsilon)^2}{\varepsilon}+\varphi\Big(\frac{x}{\delta_\varepsilon}\Big)(v_\varepsilon')^2\right)dx\le\int_a^b(u')^2\,dx+\mathbf{m}^\infty+2\eta\,.$$

Thus, upon replacing v_ε by $0\vee(v_\varepsilon\wedge 1)$ we conclude by the arbitrariness of $\eta>0$. □

7 Limit Analysis of $\mathbf{m}^\ell$

We conclude this note by analysing the convergence of the constant $\mathbf{m}^\ell$ as $\ell\to 0^+$ and $\ell\to+\infty$. Namely, we prove (20), thus concluding the proof of Theorem 1.

Proposition 6 *Let $\ell\in(0,+\infty)$ and $\mathbf{m}^\ell$ be as in* (17)*. Let moreover $\mathbf{m}^0$ and $\mathbf{m}^\infty$ be as in* (16) *and* (19)*, respectively. Then*

$$\lim_{\ell\to+\infty}\mathbf{m}^\ell=\mathbf{m}^\infty\,. \tag{80}$$

If φ is upper semicontinuous, it also holds

$$\lim_{\ell\to 0^+}\mathbf{m}^\ell=\mathbf{m}^0\,. \tag{81}$$

Proof The proof of (80) and (81) uses arguments which are similar to those employed in the proof of Proposition 5 and Proposition 3, respectively. For this reason, we only sketch this proof.

Step 1: Proof of (80).

We first show that

$$\liminf_{\ell\to+\infty}\mathbf{m}^\ell\ge\mathbf{m}^\infty\,. \tag{82}$$

To this end, we fix $\eta > 0$; using a similar argument as in the proof of Lemma 1 we can find $T_\eta > 0$ and for every $\ell \in (0, +\infty)$ a real number $z_{\eta,\ell} \in [0, 1)$ and $v_{\eta,\ell} \in W^{1,2}(-T_\eta, T_\eta)$ such that $0 \le v_{\eta,\ell} \le 1$, $v_{\eta,\ell}(\frac{z_{\eta,\ell}}{\ell}) = 0$, $v_\eta(\pm T_\eta) = 1$ and

$$\int_{-T_\eta}^{T_\eta} \Big((1 - v_{\eta,\ell})^2 + \varphi(\ell x)(v'_{\eta,\ell})^2\Big)\, dx \le \mathbf{m}^\ell + \eta\,. \tag{83}$$

Note that T_η can be chosen independently of ℓ. We now define $\tilde{v}_{\eta,\ell} \in W^{1,2}(0, T_\eta)$ by setting

$$\tilde{v}_{\eta,\ell} := \begin{cases} v_{\eta,\ell} & \text{if } \frac{z_{\eta,\ell}}{\ell} \le x \le T_\eta\,, \\ 0 & \text{if } 0 \le x < \frac{z_{\eta,\ell}}{\ell}\,. \end{cases}$$

Since $z_{\eta,\ell} \in [0, 1)$, we readily obtain

$$\begin{aligned} &\int_{\frac{z_{\eta,\ell}}{\ell}}^{T_\eta} \Big((1 - v_{\eta,\ell})^2 + \varphi(\ell x)(v'_{\eta,\ell})^2\Big)\, dx \\ &\quad \ge \int_0^{T_\eta} \Big((1 - \tilde{v}_{\eta,\ell})^2 + \varphi(\ell x)(\tilde{v}'_{\eta,\ell})^2\Big)\, dx - \frac{1}{\ell} \\ &\quad \ge \inf\bigg\{ \int_0^{T_\eta} \Big((1 - v)^2 + \varphi(\ell x)(v')^2\Big)\, dx \colon v \in W^{1,2}(0, T_\eta), \\ &\qquad\qquad\qquad\qquad v(0) = 0,\ v(T_\eta) = 1 \bigg\} - \frac{1}{\ell}\,. \end{aligned}$$

Thus, arguing as in the proof of Proposition 5, applying the classical homogenisation result together with the Modica–Mortola trick we deduce that

$$\liminf_{\ell \to +\infty} \int_{\frac{z_{\eta,\ell}}{\ell}}^{T_\eta} \Big((1 - v_{\eta,\ell})^2 + \varphi(\ell x)(v'_{\eta,\ell})^2\Big)\, dx \ge \frac{\mathbf{m}^\infty}{2}\,.$$

Since an analogous argument holds on $(-T_\eta, \frac{z_{\eta,\ell}}{\ell})$, in view of (83) we get

$$\liminf_{\ell \to +\infty} \mathbf{m}^\ell \ge \mathbf{m}^\infty - \eta\,,$$

from which we deduce (82) by letting $\eta \to 0$.

Then, it remains to prove that

$$\limsup_{\ell \to +\infty} \mathbf{m}^\ell \le \mathbf{m}^\infty\,.$$

We fix $\eta > 0$; arguing as in the proof of Proposition 5 Step 2 we use (26) together with the classical homogenisation result with boundary conditions to find $T_\eta > 0$ and a sequence $(v_{\eta,\ell})_\ell \subset W^{1,2}(0, T_\eta)$ satisfying $v_{\eta,\ell}(0) = 0$, $v_{\eta,\ell}(T_\eta) = 1$ and

$$\lim_{\ell \to +\infty} \int_0^{T_\eta} \Big((1 - v_{\eta,\ell})^2 + \varphi(\ell x)(v'_{\eta,\ell})^2\Big)\, dx \le \frac{\mathbf{m}^\infty}{2} + \eta\,. \tag{84}$$

Upon truncation we can additionally assume that $0 \le v_{\eta,\ell} \le 1$. Since $v_{\eta,\ell}(0) = 0$, the reflected function $\tilde{v}_{\eta,\ell}$ defined by setting $\tilde{v}_{\eta,\ell}(x) := v_{\eta,\ell}(|x|)$ belongs to $W^{1,2}(-T_\eta, T_\eta)$. Moreover, upon extending $\tilde{v}_{\eta,\ell}$ by 1 it is admissible for $\mathbf{m}_0^\ell$. Thus, (84) implies that

$$\limsup_{\ell \to +\infty} \mathbf{m}^\ell \le \limsup_{\ell \to +\infty} \mathbf{m}_0^\ell \le \lim_{\ell \to +\infty} 2 \int_0^{T_\eta} \Big((1 - v_{\eta,\ell})^2 + \varphi(\ell x)(v'_{\eta,\ell})^2\Big)\, dx \le \mathbf{m}^\infty + 2\eta\,,$$

which together with (82) gives (80) by the arbitrariness of $\eta > 0$.

Step 2: Proof of (81).

By definition of $\mathbf{m}^0$, from (27) we immediately deduce that $\liminf_{\ell \to 0} \mathbf{m}^\ell \ge \mathbf{m}^0$. To prove the opposite inequality, we fix $\eta > 0$ and choose $y_\eta \in (0, 1)$ such that $\sqrt{\varphi(y_\eta)} \le \sqrt{\alpha} + \eta$. Moreover, we set

$$v_\eta(x) := 1 - \exp\bigg(-\frac{|x|}{\sqrt{\varphi(y_\eta)}}\bigg)\,.$$

Then $v_\eta \in W^{1,2}_{\mathrm{loc}}(\mathbb{R})$, $0 \le v_\eta \le 1$ and v_η satisfies $v_\eta(0) = 0$, $v_\eta(\pm\infty) = 1$ and

$$\int_{\mathbb{R}} \Big((1 - v_\eta)^2 + \varphi(y_\eta)(v'_\eta)^2\Big)\, dx = 2\sqrt{\varphi(y_\eta)} \le \mathbf{m}^0 + 2\eta\,. \tag{85}$$

Since v_η is admissible for $\mathbf{m}^\ell_{y_\eta} \ge \mathbf{m}^\ell$, by the reverse Fatou Lemma and the upper semicontinuity of φ from (85) we deduce that

$$\begin{aligned}\limsup_{\ell \to 0^+} \mathbf{m}^\ell &\le \limsup_{\ell \to 0^+} \mathbf{m}^\ell_{y_\eta} \le \limsup_{\ell \to 0^+} \int_{\mathbb{R}} \Big((1 - v_\eta)^2 + \varphi(\ell x + y_\eta)(v'_\eta)^2\Big)\, dx \\ &\le \int_{\mathbb{R}} \bigg((1 - v_\eta)^2 + \limsup_{\ell \to 0^+} \varphi(\ell x + y_\eta)(v'_\eta)^2\bigg)\, dx \le \mathbf{m}^0 + 2\eta\,.\end{aligned}$$

By the arbitrariness of $\eta > 0$ this concludes the proof. □

Acknowledgments The work of R. Marziani and C. I. Zeppieri was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project 3160408400 and under the Germany Excellence Strategy EXC 2044-390685587, Mathematics Münster: Dynamics–Geometry–Structure.

References

1. L. Ambrosio, V. M. Tortorelli, Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence. Commun. Pure Appl. Math. **43**(8), 999–1036 (1990)
2. L. Ambrosio, V. M. Tortorelli, On the approximation of free-discontinuity problems. Boll. Un. Mat. Ital. (7) **6-B**(1), 105–123 (1992)
3. N. Ansini, A. Braides, V. Chiadò Piat, Interaction between homogenization and phase-transition processes. Proc. Steklov. Math. Inst. **236**, 373–385 (2002)
4. N. Ansini, A. Braides, V. Chiadò Piat, Gradient theory of phase transitions in composite media. Proc. Roy. Soc. Edinburgh Sect. A **133**, 265–296 (2003)
5. A. Bach, R. Marziani, C.I. Zeppieri, Γ-convergence and stochastic homogenisation of singularly-perturbed elliptic functionals (2021). Preprint. arXiv: 2102.09872
6. A. Bach, T. Esposito, R. Marziani, C.I. Zeppieri, Gradient damage models for heterogeneous materials (2022). Preprint. arXiv: 2205.13966
7. A. Braides, Γ-*Convergence for Beginners* (Oxford University Press, Oxford, 2002)
8. A. Braides, *Approximation of Free discontinuity Problems*. Springer Lect. Notes in Math. (vol. 1694) (Springer, New York, 1998)
9. A. Braides, A. Defranceschi, *Homogenization of Multiple Integrals* (Oxford University Press, New York, 1998)
10. A. Braides, C.I. Zeppieri, Multiscale analysis of a prototypical model for the interaction between microstructure and surface energy. Interfaces Free Bound. **11**, 61–118 (2009)
11. G. Dal Maso, F. Iurlano, Fracture models as Γ-limits of damage models. Commun. Pure Appl. Anal. **12**, 1657–1686 (2013)
12. M. Focardi, On the variational approximation of free-discontinuity problems in the vectorial case. Math. Models Methods Appl. Sci. **11**(4), 663–684 (2001)
13. F. Iurlano, Fracture and plastic models as Γ-limits of damage models under different regimes. Adv. Calc. Var. **6**, 165–189 (2013)
14. L. Modica, The gradient theory of phase transitions and the minimal interface criterion. Arch. Ration Mech. Anal. **98**, 123–142 (1987)
15. L. Modica, S. Mortola, Un esempio di Γ-convergenza. Boll. Un. Mat. Ital. **14-B**, 285–299 (1977)
16. K. Pham, H. Amor, J.-J. Marigo, C. Maurini, Gradient damage models and their use to approximate Brittle Fracture. Int. J. Damage Mech. **20**(4), 618–652 (2011)
17. K. Pham, J.-J. Marigo, Approche variationnelle de l'endommagement: II. Les modèles à gradient. Comptes Rendus Mécanique **338**, 199–206 (2010)

Grain Growth and the Effect of Different Time Scales

Katayun Barmak, Anastasia Dunca, Yekaterina Epshteyn, Chun Liu, and Masashi Mizuno

AMS 74N15; 35R37; 53C44; 49Q20

1 Introduction

Many technologically useful materials are polycrystals composed of a myriad of small monocrystalline grains separated by grain boundaries, see Figs. 1 and 2. Dynamics of grain boundaries play a crucial role in determining the grain structure and defining materials properties across multiple scales. Experimental and computational studies give useful insight into the geometric features and the crystallography of the grain boundary network in polycrystalline microstructures.

In this work, we consider two models for the motion of grain boundaries in a planar network with dynamic lattice misorientations and with drag of triple junctions. A classical model for the motion of grain boundaries in polycrystalline materials is growth by curvature, as a local evolution law for the grain boundaries

K. Barmak
Columbia University, New York, NY, USA
e-mail: kb2612@columbia.edu

A. Dunca
West High School, Salt Lake City, UT, USA
e-mail: anastasia.d960@slcstudents.org

Y. Epshteyn (✉)
University of Utah, Salt Lake City, UT, USA
e-mail: epshteyn@math.utah.edu

C. Liu
Illinois Institute of Technology, Chicago, IL, USA
e-mail: cliu124@iit.edu

M. Mizuno
Nihon University, Tokyo, Japan
e-mail: mizuno.masashi@nihon-u.ac.jp

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_2

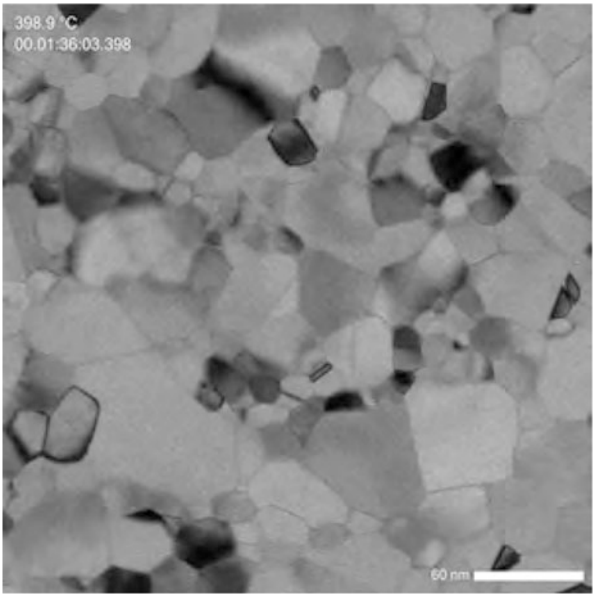


Fig. 1 Experimental microstructure: drift-corrected bright-field image of a 50 nm-thick Pt film from an instance of the in-situ grain growth experiment in the transmission electron microscope

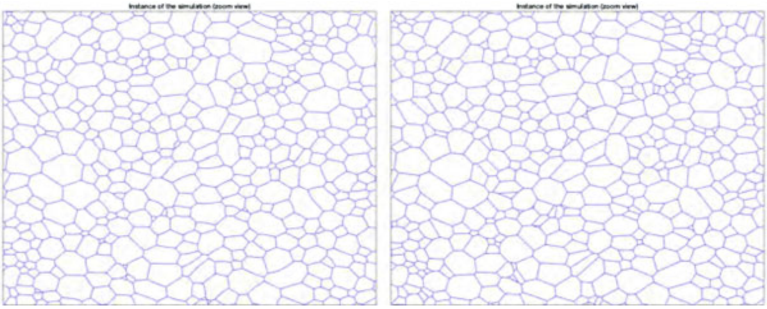

Instance of the simulation (zoom view)
Instance of the simulation (zoom view)

Fig. 2 Left figure - microstructure from simulation, model with curvature and finite mobility of the triple junctions (2): example of a time instance during the simulated evolution of a cellular network (zoom view). Right figure - microstructure from simulation, model without curvature (3): example of a time instance during the simulated evolution of a cellular network (zoom view)

due to Mullins and Herring [17, 28, 29], and see work on mean curvature flow, e.g., [11, 12, 15, 23, 25]. In addition, to have a well-posed model for the evolution of the grain boundary network, one has to impose a separate condition at the triple junctions where three grain boundaries meet [20]. A conventional choice is the Herring condition which is the natural boundary condition at the triple points for the grain boundary network at equilibrium [9, 10, 18, 20], and the references therein. There are several studies about grain boundary motion by mean curvature with the Herring condition at the triple junctions, see, for instance, [1, 4–8, 16, 20–22, 26, 37].

A standard assumption in the theory and simulations of grain growth is to address only the evolution of the grain boundaries/interfaces themselves and not the dynamics of the triple junctions. However, recent experimental work indicates that the motion of the triple junctions together with the anisotropy of the grain interfaces can have a significant effect on the resulting grain growth [7], see work on molecular dynamics simulation [36, 37], a recent work on dynamics of line defects [34, 38, 39], and a relevant work on numerical analysis of a vertex model [35]. The current work is a continuation of our previous work [13, 14], where we proposed a new model for the evolution of planar grain boundaries, which takes into account dynamic lattice misorientations (evolving anisotropy of grain boundaries or "grains rotations") and the mobility of the triple junctions. In [13, 14], using the energetic variational approach, we derived a system of geometric differential equations to describe the motion of such grain boundaries, and we established a local well-posedness result, as well as large time asymptotic behavior for the model. In addition, in [13], similar to our previous work on Grain Boundary Character Distribution, e.g., [4, 5] we conducted some numerical experiments for the 2D grain boundary network in order to illustrate the effect of time scales, e.g., of the mobility of triple junctions and of the dynamics of misorientations on how the grain boundary system decays energy and coarsens with time (note, in [13], we studied numerically only the model with curved grain boundaries). Our current goal is to conduct extensive numerical studies of two models, a model with curved grain boundaries and a model without curvature/"vertex model" of planar grain boundaries network with the dynamic lattice misorientations and with the drag of triple junctions [13, 14] and to further understand the effect of relaxation time scales, e.g., of the curvature of grain boundaries, mobility of triple junctions, and dynamics of misorientations on how the grain boundary system decays energy and coarsens with time. We also present and discuss relevant experimental results of grain growth in thin films.

The paper is organized as follows. In Sects. 2 and 3, we discuss and review important details and properties of the two models for grain boundary motion. In Sect. 4.1, we present and discuss relevant experimental findings of grain growth in thin films, and in Sect. 4.2 we conduct extensive numerical studies of the grain growth models.

2 Review of the Models with Single Triple Junction

In this paper we use recently developed models for the evolution of the planar grain boundary network with dynamic lattice misorientations and triple junction drag [13, 14] to study the effect of time scales of curvature of grain boundaries, dynamics of the triple junctions, and dynamics of the misorientations on grain growth. Thus, in this section for the reader's convenience, we first review the models which were originally developed in [13, 14].

Let us first recall the system for a single triple junction which was derived in [14]. The total grain boundary energy for such model is

$$\sum_{j=1}^{3} \sigma(\Delta^{(j)}\alpha)|\Gamma_t^{(j)}|. \tag{1}$$

Here, $\sigma : \mathbb{R} \to \mathbb{R}$ is a given surface tension, $\alpha^{(j)} = \alpha^{(j)}(t) : [0, \infty) \to \mathbb{R}$ is a time-dependent orientation of the grain $\theta = \Delta^{(j)}\alpha := \alpha^{(j-1)} - \alpha^{(j)}$ is a lattice misorientation of the grain boundary $\Gamma_t^{(j)}$ (difference in the orientation between two neighboring grains that share the grain boundary), and $|\Gamma_t^{(j)}|$ is the length of $\Gamma_t^{(j)}$. As a result of applying the maximal dissipation principle, in [14], the following model was derived,

$$\begin{cases} v_n^{(j)} = \mu\sigma(\Delta^{(j)}\alpha)\kappa^{(j)}, \quad \text{on } \Gamma_t^{(j)}, \ t > 0, \quad j = 1, 2, 3, \\ \dfrac{d\alpha^{(j)}}{dt} = -\gamma\Big(\sigma_\theta(\Delta^{(j+1)}\alpha)|\Gamma_t^{(j+1)}| - \sigma_\theta(\Delta^{(j)}\alpha)|\Gamma_t^{(j)}|\Big), \quad j = 1, 2, 3, \\ \dfrac{d\boldsymbol{a}}{dt}(t) = \eta \displaystyle\sum_{k=1}^{3} \sigma(\Delta^{(k)}\alpha)\frac{\boldsymbol{b}^{(k)}(0,t)}{|\boldsymbol{b}^{(k)}(0,t)|}, \quad t > 0, \\ \Gamma_t^{(j)} : \boldsymbol{\xi}^{(j)}(s,t), \quad 0 \le s \le 1, \quad t > 0, \quad j = 1, 2, 3, \\ \boldsymbol{a}(t) = \boldsymbol{\xi}^{(1)}(0,t) = \boldsymbol{\xi}^{(2)}(0,t) = \boldsymbol{\xi}^{(3)}(0,t), \quad \text{and} \quad \boldsymbol{\xi}^{(j)}(1,t) = \boldsymbol{x}^{(j)}, \quad j = 1, 2, 3. \end{cases} \tag{2}$$

In (2), $v_n^{(j)}$, $\kappa^{(j)}$, and $\boldsymbol{b}^{(j)} = \boldsymbol{\xi}_s^{(j)}$ are a normal velocity, a curvature, and a tangent vector of the grain boundary $\Gamma_t^{(j)}$, respectively. Note that s is not an arc length parameter of $\Gamma_t^{(j)}$, namely $\boldsymbol{b}^{(j)}$ is *not* necessarily a unit tangent vector. The vector $\boldsymbol{a} = \boldsymbol{a}(t) : [0, \infty) \to \mathbb{R}^2$ defines a position of the triple junction (triple junctions are where three grain boundaries meet), $\boldsymbol{x}^{(j)}$ is a position of the end point of the grain boundary. The three independent relaxation time scales $\mu, \gamma, \eta > 0$ (curvature, misorientation, and triple junction dynamics) are regarded as positive constants. Further, we assume in (2), $\alpha^{(0)} = \alpha^{(3)}$, $\alpha^{(4)} = \alpha^{(1)}$, and $\boldsymbol{b}^{(4)} = \boldsymbol{b}^{(1)}$, for simplicity. We also use notation $|\cdot|$ for a standard Euclidean vector norm. The complete details about model (2) can be found in the earlier work [14, Section 2]. Next, in [14], the curvature effect was relaxed, by taking the limit $\mu \to \infty$, and the reduced model without curvature was derived,

$$\begin{cases} \dfrac{d\alpha^{(j)}}{dt} = -\gamma\Big(\sigma_\theta(\Delta^{(j+1)}\alpha)|\boldsymbol{b}^{(j+1)}| - \sigma_\theta(\Delta^{(j)}\alpha)|\boldsymbol{b}^{(j)}|\Big), \quad j = 1, 2, 3, \\ \dfrac{d\boldsymbol{a}}{dt}(t) = \eta \displaystyle\sum_{j=1}^{3} \sigma(\Delta^{(j)}\alpha)\frac{\boldsymbol{b}^{(j)}}{|\boldsymbol{b}^{(j)}|}, \quad t > 0, \\ \boldsymbol{a}(t) + \boldsymbol{b}^{(j)}(t) = \boldsymbol{x}^{(j)}, \quad j = 1, 2, 3. \end{cases} \tag{3}$$

In (3), we consider $\boldsymbol{b}^{(j)}(t)$ as a grain boundary. Note that, similar to (2), the system of equations (3) can also be derived from the energetic variational principle for the total grain boundary energy (1) (with $|\Gamma_t^{(j)}|$ replaced by $|\boldsymbol{b}^{(j)}|$).

Remark 1

a) As was discussed in [14], the reduced model without curvature effect (3) is not a standard ODE system. This is the ODE system where each variable is locally constrained. Moreover, local well-posedness result (e.g., local existence result) for the original model (2) will not imply local well-posedness result for the reduced system (3). It is not known if the reduced model (3) is a small perturbation of (2).

b) The reduced model (3) captures the dynamics of the orientations/misorientations and the triple junctions. At the same time, it was more accessible for the mathematical analysis than the model (2). In addition, the system (3) is a generalization to higher dimension and dynamic misorientations of the model from [5, 8]. In this paper, we will compare and contrast through extensive numerical studies the model with the curvature effect (2) and the reduced model (3).

To establish local well-posedness result for model (3) in [14], the surface tension σ was assumed to be C^3, positive, and minimized at 0, namely

$$\sigma(\theta) \geq \sigma(0) > 0, \tag{4}$$

for $\theta \in \mathbb{R}$. In addition, it was assumed convexity of $\sigma(\theta)$, for all $\theta \in \mathbb{R}$,

$$\sigma_\theta(\theta)\theta \geq 0, \qquad \text{and} \quad \sigma_{\theta\theta}(0) > 0, \tag{5}$$

and

$$\sigma_\theta(\theta) = 0 \text{ if and only if } \theta = 0. \tag{6}$$

Let us review some of the important theoretical results established for (3) in previous work [13, 14]. First, consider the equilibrium state of the system (3), namely

$$\begin{cases} 0 = -\left(\sigma_\theta(\Delta^{(j+1)}\alpha_\infty)|\boldsymbol{b}_\infty^{(j+1)}| - \sigma_\theta(\Delta^{(j)}\alpha_\infty)|\boldsymbol{b}_\infty^{(j)}|\right), \quad j = 1, 2, 3, \\ \mathbf{0} = \displaystyle\sum_{j=1}^{3} \sigma(\Delta^{(j)}\alpha_\infty)\frac{\boldsymbol{b}_\infty^{(j)}}{|\boldsymbol{b}_\infty^{(j)}|}, \\ \boldsymbol{a}_\infty = \boldsymbol{x}^{(1)} - \boldsymbol{b}_\infty^{(1)} = \boldsymbol{x}^{(2)} - \boldsymbol{b}_\infty^{(2)} = \boldsymbol{x}^{(3)} - \boldsymbol{b}_\infty^{(3)}. \end{cases} \tag{7}$$

As in [13, 14], assume, for each $i = 1, 2, 3$,

$$\left| \sum_{j=1, j\neq i}^{3} \frac{\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)}}{|\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(i)}|} \right| > 1. \tag{8}$$

The assumption (8) implies that fixed points $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$, and $\boldsymbol{x}^{(3)}$ cannot belong to the single line. Furthermore, (8) is equivalent to the condition that in the triangle with vertices $\boldsymbol{x}^{(1)}\boldsymbol{x}^{(2)}\boldsymbol{x}^{(3)}$, all three angles are less than $\frac{2\pi}{3}$. Next, from the assumptions (8), (5)–(6), associated equilibrium system (7) becomes,

$$\begin{cases} \displaystyle\sum_{j=1}^{3} \frac{\boldsymbol{b}_{\infty}^{(j)}}{|\boldsymbol{b}_{\infty}^{(j)}|} = \boldsymbol{0}, \\ \boldsymbol{a}_{\infty} + \boldsymbol{b}_{\infty}^{(j)} = \boldsymbol{x}^{(j)}, \quad j = 1, 2, 3. \end{cases} \tag{9}$$

In [14], it was shown that the assumptions (5)–(6) imply $\alpha_{\infty}^{(1)} = \alpha_{\infty}^{(3)} = \alpha_{\infty}^{(3)}$, hence $\Delta^{(j)}\alpha_{\infty} = 0$ for $j = 1, 2, 3$ for the equilibrium system (7) (note that in this case, for the purpose of mathematical modeling, one can still assume a "fictitious" grain boundary with the same orientation on each side of the grain boundary. In addition, in this work we study the grain boundary system before it reaches a state of constant orientations, see Sect. 4.)

We also have energy dissipation principle for the system (3),

Proposition 1 (Energy Dissipation [14, Proposition 5.1]) *Let* $(\boldsymbol{\alpha}, \boldsymbol{a})$ *be a solution of* (3) *on* $0 \le t \le T$*, and let* $E(t)$*, given by* (1)*, be the total grain boundary energy of the system. Then, for all* $0 < t \le T$*,*

$$E(t) + \frac{1}{\gamma}\int_0^t \left|\frac{d\boldsymbol{\alpha}}{dt}(\tau)\right|^2 d\tau + \frac{1}{\eta}\int_0^t \left|\frac{d\boldsymbol{a}}{dt}(\tau)\right|^2 d\tau = E(0). \tag{10}$$

Next, define, constant as in [13],

$$C_1 := \inf\left\{\sum_{j=1}^{3} |\boldsymbol{x}^{(j)} - \boldsymbol{a}| : \text{There exists } j = 1, 2, 3 \text{ such that } |\boldsymbol{a} - \boldsymbol{a}_{\infty}| \ge \frac{1}{2}|\boldsymbol{b}_{\infty}^{(j)}|\right\}. \tag{11}$$

Assume also that an initial data $(\boldsymbol{\alpha}_0, \boldsymbol{a}_0)$ satisfies,

$$E(0) = \sum_{j=1}^{3} \sigma(\Delta^{(j)}\alpha_0)|\boldsymbol{a}_0 - \boldsymbol{x}^{(j)}| < \sigma(0)C_1. \tag{12}$$

Then, one can establish the global existence result for the model (3),

Theorem 1 (Global Existence [13, Theorem 4.1]) *Let* $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)} \in \mathbb{R}^2$*,* $\boldsymbol{a}_0 \in \mathbb{R}^2$*, and* $\boldsymbol{\alpha}_0 \in \mathbb{R}^3$ *be the initial data for the system* (3)*. Assume* (8)*, and let* $\boldsymbol{a}_{\infty}$ *be a unique solution of the equilibrium system* (9)*. Further, assume condition* (12)*. Then there exists a unique global in time solution* $(\boldsymbol{\alpha}, \boldsymbol{a})$ *of* (3)*.*

We also have the following large time asymptotic behavior results for the solution of system (3),

Proposition 2 (Large Time Asymptotic [13, Proposition 5.1]) *Let $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$, $\boldsymbol{x}^{(3)} \in \mathbb{R}^2$, $\boldsymbol{a}_0 \in \mathbb{R}^2$, and $\boldsymbol{\alpha}_0 \in \mathbb{R}^3$ be the initial data for the system (3). We assume that the initial data satisfy (12), and we also impose the same assumptions as in Theorem 1. Define α_∞ as,*

$$\alpha_\infty := \frac{\alpha_0^{(1)} + \alpha_0^{(2)} + \alpha_0^{(3)}}{3}. \tag{13}$$

Let $\boldsymbol{a}_\infty$ be a solution of the equilibrium system (9) and $(\boldsymbol{\alpha}, \boldsymbol{a})$ be a time global solution of (3). Then,

$$\boldsymbol{\alpha}(t) \to \alpha_\infty(1, 1, 1), \quad \boldsymbol{a}(t) \to \boldsymbol{a}_\infty, \tag{14}$$

as $t \to \infty$.

Theorem 2 (Large Time Asymptotic [13, Theorem 5.1]) *There is a small constant $\epsilon_1 > 0$ such that, if $|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_\infty| + |\boldsymbol{a}_0 - \boldsymbol{a}_\infty| < \epsilon_1$, then the associated global solution $(\boldsymbol{\alpha}, \boldsymbol{a})$ of the system (3) satisfies,*

$$|\boldsymbol{\alpha}(t) - \boldsymbol{\alpha}_\infty| + |\boldsymbol{a}(t) - \boldsymbol{a}_\infty| \le C_2 e^{-\lambda^\star t}, \tag{15}$$

for some positive constants $C_2, \lambda^\star > 0$.

Remark 2 The decay order $\lambda^\star$ in (15) is explicitly estimated as,

$$\lambda^\star \ge \lambda, \tag{16}$$

where λ depends on γ, η, $\sigma_{\theta\theta}(0)$, $\sigma(0)$ and on the smallest positive eigenvalues of the linearized operators for the equations of the orientation $\boldsymbol{\alpha}$ and of the triple junction $\mathbf{a}$.

Corollary 1 (Large Time Asymptotic [13, Corollary 5.1]) *Under the same assumption as in Theorem 2, the associated grain boundary energy $E(t)$ satisfies,*

$$E(t) - E_\infty \le C_3 e^{-\lambda^\star t}, \tag{17}$$

for some positive constant $C_3 > 0$, where

$$E_\infty := \sigma(0) \sum_{j=1}^{3} |\boldsymbol{b}_\infty^{(j)}|.$$

Proof For the reader's convenience, we will review the proof from [13]. Since $\alpha_\infty^{(1)} = \alpha_\infty^{(2)} = \alpha_\infty^{(3)}$, we obtain

$$
\begin{aligned}
E(t) - E_\infty &= \sum_{j=1}^{3} \left(\sigma(\Delta^{(j)}\alpha(t)) |\boldsymbol{b}^{(j)}(t)| - \sigma(0) |\boldsymbol{b}_\infty^{(j)}| \right) \\
&\le \sum_{j=1}^{3} \left(\sigma(0) |\boldsymbol{b}^{(j)}(t) - \boldsymbol{b}_\infty^{(j)}| + \left(\sigma(\Delta^{(j)}\alpha(t)) - \sigma(0) \right) |\boldsymbol{b}^{(j)}(t)| \right) \\
&\le \sum_{j=1}^{3} \left(\sigma(0) |\boldsymbol{a}^{(j)}(t) - \boldsymbol{a}_\infty| + \left(C_4 |\Delta^{(j)}\alpha(t)| \right) |\boldsymbol{b}^{(j)}(t)| \right) \\
&\le \sum_{j=1}^{3} \left(\sigma(0) |\boldsymbol{a}^{(j)}(t) - \boldsymbol{a}_\infty| + 2C_4 |\boldsymbol{b}^{(j)}(t)| |\boldsymbol{\alpha}(t) - \boldsymbol{\alpha}_\infty| \right),
\end{aligned}
\tag{18}
$$

where $C_4 = \sup_{|\theta|<2\epsilon_1} |\sigma_\theta(\theta)|$. Using the dissipation estimate (10) and the exponential decay estimate (15), we obtain (17). □

Remark 3 Note that the obtained exponential decay to equilibrium, see estimates (15) and (17) was obtained by considering linearized problem, Lemma 5.1 in [13]. Consideration of the model with curvature - with finite μ, (2) and of the nonlinear problem instead of linearized problem could lead to potential power laws estimates for the decay rates. See also discussion and numerical studies in Sect. 4.

3 Extension to Grain Boundary Network

In this section, we review the extension of the results to a grain boundary network $\{\Gamma_t^{(j)}\}$. As in [13, 14], we define the total grain boundary energy of the network, like,

$$
E(t) = \sum_j \sigma(\Delta^{(j)}\alpha) |\Gamma_t^{(j)}|, \tag{19}
$$

where $\Delta^{(j)}\alpha$ is a misorientation, a difference between the lattice orientation of the two neighboring grains which form the grain boundary $\Gamma_t^{(j)}$. Then, the energetic variational principle leads to a full model (network model analog of a single triple junction system (2)),

$$
\begin{cases}
v_n^{(j)} = \mu\sigma(\Delta^{(j)}\alpha)\kappa^{(j)}, \quad \text{on } \Gamma_t^{(j)},\ t > 0, \\
\dfrac{d\alpha^{(k)}}{dt} = -\gamma \dfrac{\delta E}{\delta \alpha^{(k)}}, \\
\dfrac{d\boldsymbol{a}^{(l)}}{dt} = \eta \displaystyle\sum_{\boldsymbol{a}^{(l)} \in \Gamma_t^{(j)}} \left(\sigma(\Delta^{(j)}\alpha) \frac{\boldsymbol{b}^{(j)}}{|\boldsymbol{b}^{(j)}|} \right), \quad t > 0.
\end{cases}
\tag{20}
$$

As in [14], we consider the relaxation parameters, $\mu \to \infty$, and we further assume that the energy density $\sigma(\theta)$ is an even function with respect to the misorientation $\theta = \Delta^{(j)}\alpha$, that is, the misorientation effects are symmetric with respect to the difference between the lattice orientations. Then, the problem (20) reduces to (network model analog of a single triple junction system (3)),

$$\begin{cases} \Gamma_t^{(j)} \text{ is a line segment between some } \boldsymbol{a}^{(l_{j,1})} \text{ and } \boldsymbol{a}^{(l_{j,2})}, \\ \dfrac{d\alpha^{(k)}}{dt} = -\gamma \displaystyle\sum_{\substack{\text{grain with } \alpha^{(k')} \text{ is the neighbor of the grain with } \alpha^{(k)} \\ \Gamma_t^{(j)} \text{ is formed by the two grains with } \alpha^{(k)} \text{ and } \alpha^{(k')}}} |\Gamma_t^{(j)}|\sigma_\theta(\alpha^{(k)} - \alpha^{(k')}), \\ \dfrac{d\boldsymbol{a}^{(l)}}{dt} = \eta \displaystyle\sum_{\boldsymbol{a}^{(l)} \in \Gamma_t^{(j)}} \left(\sigma(\Delta^{(j)}\alpha) \frac{\boldsymbol{b}^{(j)}}{|\boldsymbol{b}^{(j)}|} \right). \end{cases} \tag{21}$$

To obtain the global solution of the system (21) in [13], we studied the system before the critical events, and we first considered an associated energy minimizing state, $(\alpha_\infty^{(k)}, \boldsymbol{a}_\infty^{(l)})$ of (21). The critical events are the disappearance events, e.g., disappearance of the grains and/or grain boundaries during coarsening of the system, facet interchange and splitting of unstable junctions. Then, $(\alpha_\infty^{(k)}, \boldsymbol{a}_\infty^{(l)})$ satisfies,

$$\begin{cases} \Gamma_\infty^{(j)} \text{ is a line segment between some } \boldsymbol{a}_\infty^{(l_{j,1})} \text{ and } \boldsymbol{a}_\infty^{(l_{j,2})}, \\ 0 = -\gamma \displaystyle\sum_{\substack{\text{grain with } \alpha^{(k')} \text{ is the neighbor of the grain with } \alpha^{(k)} \\ \Gamma_t^{(j)} \text{ is formed by the two grains with } \alpha^{(k)} \text{ and } \alpha^{(k')}}} |\Gamma_\infty^{(j)}|\sigma_\theta(\alpha_\infty^{(k)} - \alpha_\infty^{(k')}), \\ \mathbf{0} = \eta \displaystyle\sum_{\boldsymbol{a}_\infty^{(l)} \in \Gamma_\infty^{(j)}} \left(\sigma(\Delta^{(j)}\alpha_\infty) \frac{\boldsymbol{b}_\infty^{(j)}}{|\boldsymbol{b}_\infty^{(j)}|} \right). \end{cases} \tag{22}$$

Hence, the total energy E_∞ of the grain boundary network (22) is

$$E_\infty = \sum_j \sigma(\Delta^{(j)}\alpha_\infty)|\boldsymbol{b}_\infty^{(j)}| = \inf\left\{ \sum_j \sigma(\Delta^{(j)}\alpha)|\boldsymbol{b}^{(j)}| \right\}. \tag{23}$$

Remark 4 Note, we assumed in (21)–(22) that the total number of grains, grain boundaries, and triple junctions are the same as in the initial configuration (assumption of no critical events in the network).

Further, if there is a neighborhood $U^{(l)} \subset \mathbb{R}^2$ of $\boldsymbol{a}_\infty^{(l)}$ such that

$$E_\infty < \sum_j |\boldsymbol{b}^{(j)}| \tag{24}$$

for all $\boldsymbol{a}^{(l)} \in U^{(l)}$, one can obtain a priori estimate for the triple junctions, and, hence, obtain the time global solution of (21). Note that, the assumption (24) is related to the boundary condition of the line segments $\Gamma_t^{(j)}$. Further, if the energy minimizing state is unique, then we can proceed with the same argument as in Lemma 4.1 in [13] and obtain the global solution (21) near the energy minimizing state.

Remark 5 Note that, the solution of (22) may not be unique even though the grain orientations are constant (misorientation is zero) [13].

The asymptotics of the grain boundary networks are rather nontrivial. Our arguments in [13] were based on the uniqueness of the equilibrium state (9). However, we do not know the uniqueness of solutions of the equilibrium state for the grain boundary network (22). Thus, in general we cannot take a full limit for the large time asymptotic behavior of the solution of the network model (21). But, one can show, the following result instead,

Corollary 2 ([13, Corollary 6.1]) *In a grain boundary network* (21)*, assume that the initial configuration is sufficiently close to an associated energy minimizing state* (22)*. Then, there is a global solution* $(\alpha^{(k)}, \boldsymbol{a}^{(l)})$ *of* (21)*. Furthermore, there exists a time sequence* $t_n \to \infty$ *such that* $(\alpha^{(k)}(t_n), \boldsymbol{a}^{(l)}(t_n))$ *converges to an associated equilibrium configuration* (22)*.*

4 Experiments and Numerical Simulations

In this section we present results of some experiments in thin films and numerical study of the grain growth using models of planar grain boundary network from Sect. 3. The energetics and connectivity of the grain boundary network play a crucial role in determining the properties of a material across multiple scales, see also Sects. 2 and 3. Therefore, our main focus here is to develop a better understanding of the energetic properties of the experimental and computational microstructures.

4.1 Experimental Results: Grain Boundary Character Distribution

To more fully characterize a microstructure, it is necessary to consider the types and energies of the constituent grain boundaries, in addition to geometric features such as grain size. Indeed, experiments and simulations over the past 30+ years have led to the discovery and notion of the Grain Boundary Character Distribution (GBCD) [2, 3, 19, 30, 32, 33]. *The GBCD, denoted by ρ, is an empirical distribution of the relative area (in 3D) or relative length (in 2D) of interface/grain boundaries with a given misorientation and boundary normal.* The GBCD can be viewed as a leading

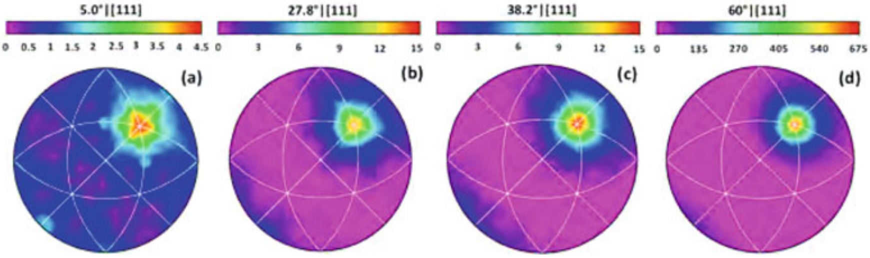


Fig. 3 Experiments: (a-d) Grain boundary character distribution of 100 nm-thick, as-deposited Al film with a mean grain size of approximately 100 nm for four given misorientations. Misorientations are specified as angle-axis pairs. Pseudosymmetry cleanup of the crystal orientation maps was used in generating the figures. The scale is multiples of random distribution

statistical descriptor to characterize the texture of the grain boundary network (see, e.g., [2, 3, 5, 19, 30, 33]).

Figure 3 presents the GBCD for four different misorientations for an as-deposited aluminum film with near random orientation distribution. The details of film deposition, sample preparation, and precession electron diffraction crystal orientation mapping in the transmission electron microscope are given in [31]. However, in contrast to [31], the orientation data were subjected to the same cleanup procedure as for the grain size distribution, namely the pseudosymmetry cleanup procedure detailed in [24] with the exception of the 60°|[111] boundaries, which are clearly abundant and should not be removed. The minimum grain size of the dilation cleanup step was 20 pixels.

Given that grain boundaries have five crystallographic degrees of freedom - three to specify the misorientation across the grain boundary, and two to define the normal to the boundary, the two-dimensional graphical presentation of the GBCD as in Fig. 3 is achieved in the following manner. To begin, a given misorientation is selected, for example, 5°|[111]. The rotation axis, here [111], is given by the Miller indices of the crystallographic direction that is common to both grains on either side of the boundary. The misorientation angle is usually, but not always, chosen to be within the fundamental zone of misorientations, which for cubic crystals has a minimum of zero and a maximum of 62.8°. Common choices of angles are either those of low angle boundaries, with rotation angles of less than 15 degrees, or those of coincident site lattice (CSL) type. In Fig. 3, the selected rotation angles about the [111] axis of 27.8°|[111], 38.2°|[111], and 60°|[111] correspond to CSL designations $\Sigma 13b$, $\Sigma 7$, and $\Sigma 3$, respectively. The numerical value in the Σ designation is the reciprocal of the number of atomic sites that are coincident in the crystallographic plane perpendicular to rotation axis. For face centered cubic crystals, the Miller indices of this plane are the same as the Miller indices of the misorientation axis, e.g., the (111) plane for the [111] rotation axis. The letters a or b in the Σ designation then indicate different angle-axis pairs with the same number of coincident sites. Note that the CSL designation does not specify the grain boundary plane that is present in the sample; rather it specifies only a given misorientation.

Next, the grain boundary planes present in the experimental sample for the given misorientation are represented by the crystallographic directions normal to the planes in standard stereographic projections, such as those in Fig. 3. The use of stereographic projection rather than other types of projections in single crystal or bicrystal crystallography of materials has been common practice. Its choice is based on the fact that it is an angle-preserving projection that does not depend on the size of the crystal (from nano to macro). For cubic crystals, the standard projection has the [001] cubic crystal axis pointing out of the page thereby projecting onto the page as the origin of the plot at the center of the (projected equatorial) circle. In Fig. 3, the [100] crystallographic axis points to the right, and the [010] crystallographic axis points up, thereby defining a right-handed axis set.

The stereographic projections of the boundary plane normals such as those of Fig. 3 then show the abundance of grain boundary plane normals in multiples of random distribution (MRD) on the thermal scale. The MRD is similar to a probability density plot, but its integrated value is 2, rather than 1, since every grain boundary segment is counted twice, once for the grain on the one side of the boundary and once for the grain on the other side of the boundary. When the direction normal to the boundary plane and the misorientation axis are the same, the grain boundary is termed a twist boundary, since the axis of rotation is normal to the observed boundary plane. In Fig. 3, a high relative intensity is seen at the position of the [111] twist boundaries for all four selected misorientations. If, on the other hand, the high intensities were seen as bands along a great circle ninety degrees away from the chosen misorientation axis, then the boundaries would have been designated as tilt boundaries, with the misorientation axis in the plane of the grain boundary. In effect, GBCD plots such as those of Fig. 3 make manifest texture formation in the grain boundary network, see also numerical experiments Sect. 4.2.

The most striking feature of Fig. 3 is the very high abundance of 60°|[111] boundaries, which show a population of several hundred times MRD. Given that the majority of the boundary planes were also found to be (111), this sample is said to have a large population of coherent $\Sigma 3$, or the so-called coherent twin boundaries. $\Sigma 3$ boundaries constitute approximately one quarter of all the boundaries in this sample. In contrast, for a "bulk" aluminum sample, i.e., in an aluminum sample with mean grain size of 23 μm, the population of $\Sigma 3$ boundaries is more than ten times lower [31]. The very high population of $\Sigma 3$ boundaries in the thin film sample of Fig. 3 is likely a result of the structure forming processes that take place during film deposition, rather than a result of normal grain growth. The evolution of the grain boundary network and the GBCD of this sample towards equilibrium or steady state will be determined by the dynamics of the grain boundaries and the relaxation time scales for the boundary curvature, misorientation, and triple junctions, for which models and simulations are presented in the current work. We note that in experimental samples where GBCD has reached steady state, the GBCD averaged over its five crystallographic parameters is inversely related to the grain boundary energy density similar to the GBCD extracted from grain growth models, Sect. 4.2. Laboratory-based experimental quantification of grain boundary dynamics via in-situ annealing experiments similar to the experiment in Fig. 1, together with intermittent mapping of crystal orientations for determination of the evolving GBCD, will be the key to connecting more closely experimental findings to mathematical and computational models of grain growth. These experiments are the subject of the ongoing research.

4.2 Numerical Experiments

Here, we present several numerical experiments to illustrate the effects of different time scales, such as the dynamic orientations/misorientations (grains "rotations") and mobility of the triple junctions, as well as we compare the grain growth model with curvature (20) and model without curvature (21), as described in Sects. 2 and 3.

In particular, the main goal of our numerical experiments is to illustrate the time scales effect of curvature—through grain boundary mobility μ, mobility of the triple junctions η, and misorientation parameter γ on how the grain boundary system decays energy and coarsens with time. For that we will numerically study evolution of the total grain boundary energy,

$$E(t) = \sum_j \sigma(\Delta^{(j)}\alpha)|\Gamma_t^{(j)}|, \tag{25}$$

whereas before, $\Delta^{(j)}\alpha$ is a misorientation of the grain boundary $\Gamma_t^{(j)}$, and $|\Gamma_t^{(j)}|$ is the length of the grain boundary. We will also consider the growth of the average area, defined as,

$$A(t) = \frac{4}{N(t)}, \tag{26}$$

here 4 is the total area of the sample, and $N(t)$ is the total number of grains at time t. The growth of the average area is closely related to the coarsening rate of the grain system that undergoes critical/disappearance events. However, it is important to note that critical events include not only grain disappearance but also facet/grain boundary disappearance, facet interchange, and splitting of unstable junctions, for more details about numerical modeling of critical events in 2D, see, e.g., [8, 21]. Further, we will investigate the distribution of the grain boundary character distribution (GBCD) $\rho(\Delta^{(j)}\alpha)$ at a final time of the simulations T_∞ (defined below) under a simplified assumption on a grain boundary energy density, namely that $\sigma(\Delta^{(j)}\alpha)$ is only a function of the misorientation, see also Sects. 2–3. The GBCD (in this context) is an empirical statistical measure of the relative length (in 2D) of the grain boundary interface with a given lattice misorientation,

$$\rho(\Delta^{(j)}\alpha, t) = \text{ relative length of interface of lattice misorientation } \Delta^{(j)}\alpha \text{ at time } t, \quad \text{normalized so that } \int_{\Omega_{\Delta^{(j)}\alpha}} \rho d\Delta^{(j)}\alpha = 1, \tag{27}$$

where we consider $\Omega_{\Delta^{(j)}\alpha} = [-\frac{\pi}{4}, \frac{\pi}{4}]$ in the numerical experiments below (for planar grain boundary network, it is reasonable to consider such range for the misorientations). For more details, see, for example, [5, 13]. In all our tests below, we compare the GBCD at T_∞ to the stationary solution of the Fokker–Planck equation, the Boltzmann distribution for the grain boundary energy density $\sigma(\Delta^{(j)}\alpha)$,

$$\rho_D(\Delta^{(j)}\alpha) = \frac{1}{Z_D} e^{-\frac{\sigma(\Delta^{(j)}\alpha)}{D}}, \quad \text{with partition function, i.e., normalization factor} \quad Z_D = \int_{\Omega_{\Delta^{(j)}\alpha}} e^{-\frac{\sigma(\Delta^{(j)}\alpha)}{D}} d\Delta^{(j)}\alpha, \tag{28}$$

[4–6, 8]. We employ the Kullback–Leibler relative entropy test to obtain a unique "temperature-like" parameter D and to construct the corresponding Boltzmann distribution for the GBCD at T_∞ as it was originally done in [4–6, 8]. Note, as we also discussed in Sect. 4.1, GBCD is a primary candidate to characterize texture of the grain boundary network, and is inversely related to the grain boundary energy density as discovered in experiments and simulations. The reader can consult, for example, [4–6, 8] for more details about GBCD and the theory of the GBCD. In the numerical experiments in this paper, we consider two choices for the grain boundary energy density as plotted in Fig. 4 and given below,

$$\sigma(\Delta^{(j)}\alpha) = 1 + 0.25 \sin^2(2\Delta^{(j)}\alpha) \text{ and } \sigma(\Delta^{(j)}\alpha) = 1 + 0.25 \sin^4(2\Delta^{(j)}\alpha).$$

We consider simulation of 2D grain boundary network using the algorithm based on the sharp interface approach [13] with dynamic misorientation and finite mobility

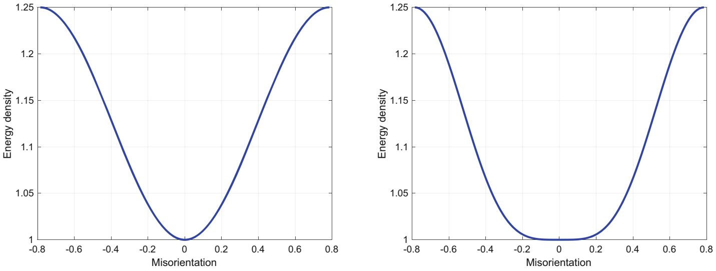


Fig. 4 Grain boundary energy density function $\sigma(\Delta\alpha)$: *(a) Left plot,* $\sigma = 1+0.25\sin^2(2\Delta\alpha)$ and *(b) Right plot,* $\sigma = 1 + 0.25\sin^4(2\Delta\alpha)$

of the triple junctions which we also extended to a model without curvature (21). Note that the algorithm [13] is a further extension of the algorithm from [4, 8]. We recall that in the numerical scheme we work with a variational principle. The cornerstone of the algorithm, which assures its stability, is the discrete dissipation inequality for the total grain boundary energy that holds when either the discrete Herring boundary condition ($\eta \to \infty$) or discrete "dynamic boundary condition" (finite mobility η of the triple junctions, third equation of (20) or of (21)) is satisfied at the triple junctions. We also recall that in the numerical algorithm for model (20) we impose the Mullins' theory (first equation of (20)) as the local evolution law for the grain boundaries (and the time scale μ is kept finite). For model (21), $\mu \to \infty$, hence the dynamics of the grain boundaries are defined by the evolution of the triple junctions (the third equation of (21)) and by the grains rotation (the second equation of (21)). The reader can consult [4, 8, 13] for more details about numerical algorithm based on the sharp interface approach.

In all the numerical tests below we initialized our system with 10^4 cells/grains with normally distributed misorientation angles at initial time $t = 0$. We also assume that the final time of the simulations T_∞ is the time when approximately 80% of grains disappeared from the system, namely the time when only about 2000 cells/grains remain. The final time is selected based on the system (20) with no dynamic misorientations ($\gamma = 0$) and with the Herring condition at the triple junctions ($\eta \to \infty$) and, it is selected to ensure that statistically significant number of grains still remain in the system and that the system reached its statistical steady state. Therefore, all the numerical results which are presented below are for the grain boundary system that undergoes critical/disappearance events.

First, we study the effect of dynamics of triple junctions on the dissipation and coarsening of the system, see Figs. 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19

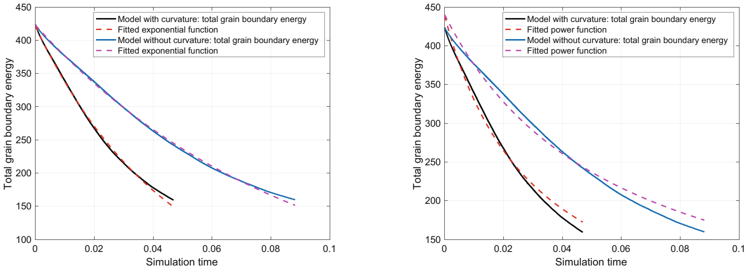


Fig. 5 One run of 2D trial with 10000 initial grains: *(a) Left plot,* total grain boundary energy plot, model with curvature (solid black) versus fitted exponential decaying function $y(t) = 420.9\exp(-22.13t)$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted exponential decaying function $y(t) = 422.8\exp(-11.64t)$ (dashed magenta); *(b) Right plot*, total grain boundary energy plot, model with curvature (solid black) versus fitted power function $y_1(t) = 438.8797(1.0 + 32.9489t)^{-1}$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted power function $y_1(t) = 439.9588(1.0 + 17.1792t)^{-1}$ (dashed magenta). Mobility of the triple junctions is $\eta = 10$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^2(2\Delta\alpha)$

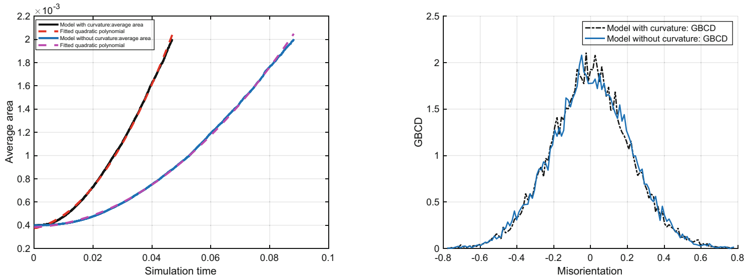


Fig. 6 *(a) Left plot,* one run of 2D trial with 10000 initial grains: Growth of the average area of the grains, model with curvature (solid black) versus fitted quadratic polynomial function $y(t) = 0.6575t^2 + 0.004668t + 0.0003745$ (dashed red). Growth of the average area of the grains, model without curvature (solid blue) versus fitted quadratic polynomial function $y(t) = 0.2025t^2 + 0.001016t + 0.0003844$ (dashed magenta); *(b) Right plot*, GBCD (black curve, model with curvature) and GBCD (blue curve, model without curvature) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains. Mobility of the triple junctions is $\eta = 10$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^2(2\Delta\alpha)$

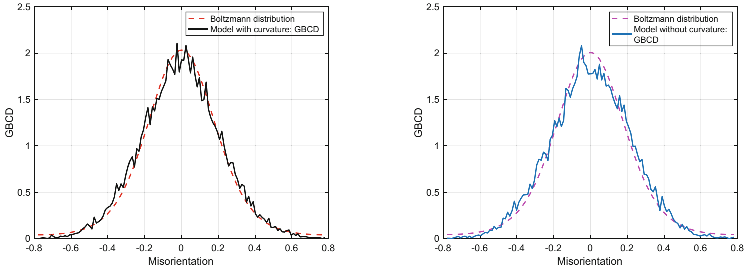


Fig. 7 *(a) Left plot,* model with curvature, GBCD (black curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.0641$ (dashed red curve). *(b) Right plot*, model without curvature, GBCD (blue curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.0655$ (dashed magenta curve). Mobility of the triple junctions is $\eta = 10$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^2(2\Delta\alpha)$

20, 21, and 22 (we consider different values of misorientation parameter γ for these tests). We observe that for smaller values of the mobility of the triple junctions η, the energy decay $E(t)$ is well-approximated by an exponential function for both models, for the model with curvature (20) and for the model without curvature (21), see Figs. 5 and 8 (left plots). This is consistent with the results of our theory, see Sects. 2 and 3 and [13, 14], even though, the theoretical results are obtained under assumption of no critical events and $\mu \to \infty$ (for grain growth model without curvature). This result indicates that for lower mobility of the triple junctions η, the dynamics of triple junctions have a dominant effect on the grain growth, see model (21). This explains the similarity in the energy decay for grain growth model with curvature (20) and without curvature (21) when $\eta = 10$, Figs. 5 and 8. In comparison, we

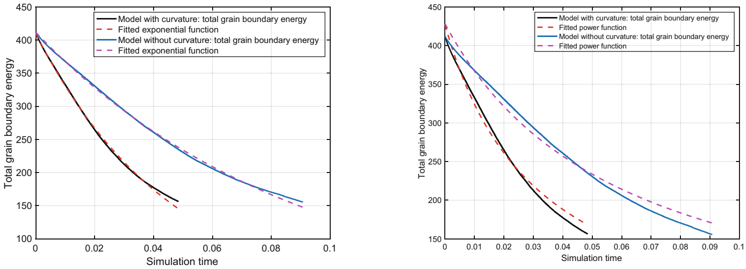


Fig. 8 One run of 2D trial with 10000 initial grains: *(a) Left plot,* total grain boundary energy plot, model with curvature (solid black) versus fitted exponential decaying function $y(t) = 409.8\exp(-21.38t)$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted exponential decaying function $y(t) = 411.6\exp(-11.3t)$ (dashed magenta); *(b) Right plot*, total grain boundary energy plot, model with curvature (solid black) versus fitted power function $y_1(t) = 426.9841(1.0{+}31.746t)^{-1}$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted power function $y_1(t) = 428.2145(1.0 + 16.6556t)^{-1}$ (dashed magenta). Mobility of the triple junctions is $\eta = 10$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^4(2\Delta\alpha)$

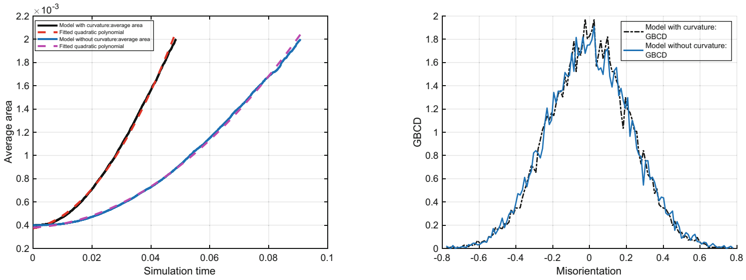


Fig. 9 *(a) Left plot,* one run of 2D trial with 10000 initial grains: Growth of the average area of the grains, model with curvature (solid black) versus fitted quadratic polynomial function $y(t) = 0.6258t^2 + 0.004538t + 0.0003732$ (dashed red). Growth of the average area of the grains, model without curvature (solid blue) versus fitted quadratic polynomial function $y(t) = 0.1866t^2 + 0.001377t + 0.0003799$ (dashed magenta); *(b) Right plot*, GBCD (black curve, model with curvature) and GBCD (blue curve, model without curvature) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains. Mobility of the triple junctions is $\eta = 10$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^4(2\Delta\alpha)$

also present fit to a power law decaying function, see Figs. 5 and 8 (right plots). The power law function does not seem to give as good approximation in this case.

However, for a larger value of $\eta = 100$, Figs. 11, 14, 17, and 20, we obtain that the total grain boundary energy does not follow exponential decay anymore for the model with curvature (20), but rather the energy decay is closer to a power law. Thus, the curvature time scale-the grain boundary evolution has a dominant effect for large η. However, for the model without curvature (21), the energy decay is still well approximated by the exponential function which is consistent with the theory, Sects. 2 and 3. Note also that the numerically observed energy decay rates

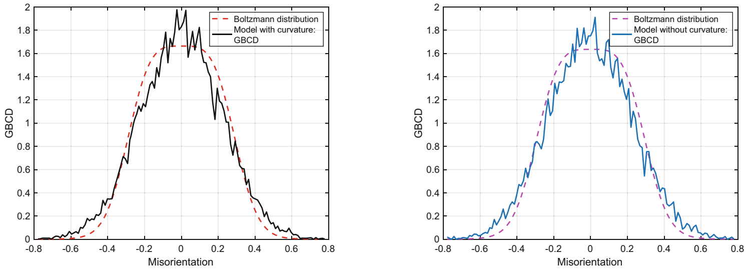


Fig. 10 *(a) Left plot,* model with curvature, GBCD (black curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.035$ (dashed red curve). *(b) Right plot*, model without curvature, GBCD (blue curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.035$ (dashed magenta curve). Mobility of the triple junctions is $\eta = 10$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25 \sin^4(2\Delta\alpha)$

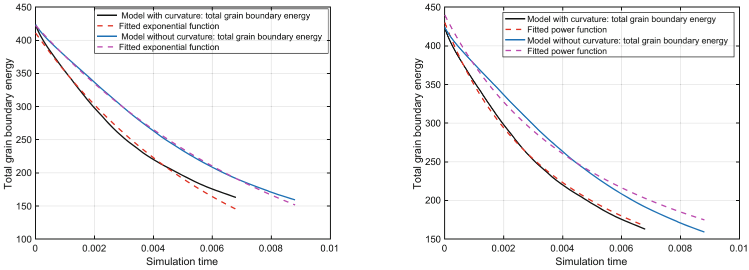


Fig. 11 One run of 2D trial with 10000 initial grains: *(a) Left plot,* total grain boundary energy plot, model with curvature (solid black) versus fitted exponential decaying function $y(t) = 411\exp(-153t)$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted exponential decaying function $y(t) = 422.4\exp(-116.3t)$ (dashed magenta); *(b) Right plot*, total grain boundary energy plot, model with curvature (solid black) versus fitted power function $y_1(t) = 430.0278(1.0 + 231.6960t)^{-1}$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted power function $y_1(t) = 439.8212(1.0 + 171.9395t)^{-1}$ (dashed magenta). Mobility of the triple junctions is $\eta = 100$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^2(2\Delta\alpha)$

increase with the mobility η of the triple junctions which is also consistent with the developed theory [13]. In addition, we observe that the average area grows as a quadratic function in time for the finite mobility η of the triple junctions, Figs. 6, 9, 12, 15, 18, and 21 (left plots) and see also our earlier work [13]. We also observe that the coarsening rate of grain growth slows down with the smaller η. In addition, we note that the energy decay in our numerical tests is consistent with the growth of the average area. Moreover, we observe that neither dynamics of the triple junctions nor curvature show as much of an effect on the GBCD, see Figs. 6 (right plot)-7, 9 (right plot), 10, 12 (right plot), 13, 15 (right plot), 16, 18 (right plot), 19 and 21 (right plot), 22. (Note, the "temperature" like parameter D also accounts for various critical events–grains disappearance, facet/grain boundary

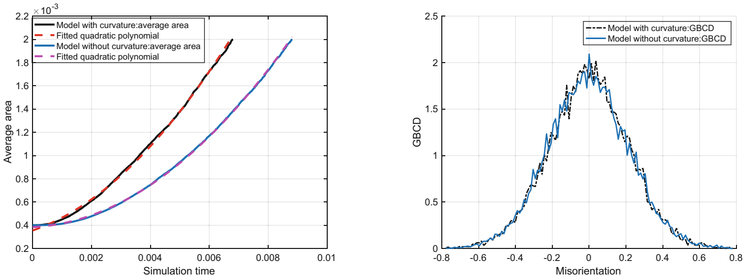


Fig. 12 *(a) Left plot,* one run of 2D trial with 10000 initial grains: Growth of the average area of the grains, model with curvature (solid black) versus fitted quadratic polynomial function $y(t) = 23.47t^2 + 0.08748t + 0.0003549$ (dashed red). Growth of the average area of the grains, model without curvature (solid blue) versus fitted quadratic polynomial function $y(t) = 19.74t^2 + 0.01157t + 0.0003843$ (dashed magenta); *(b) Right plot*, GBCD (black curve, model with curvature) and GBCD (blue curve, model without curvature) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains. Mobility of the triple junctions is $\eta = 100$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^2(2\Delta\alpha)$

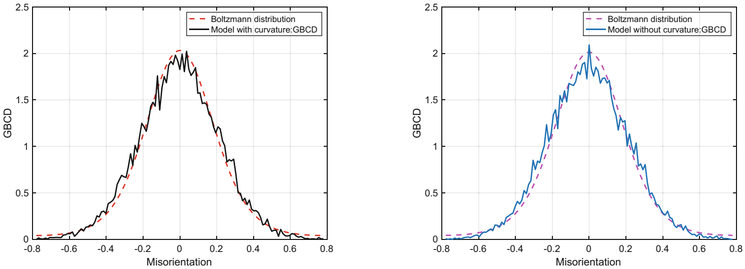


Fig. 13 *(a) Left plot,* model with curvature, GBCD (black curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with “temperature”- $D \approx 0.0641$ (dashed red curve). *(b) Right plot*, model without curvature, GBCD (blue curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with “temperature”- $D \approx 0.0651$ (dashed magenta curve). Mobility of the triple junctions is $\eta = 100$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^2(2\Delta\alpha)$

disappearance, facet interchange, splitting of unstable junctions. It will be part of our future study to understand how D depends on the critical events).

For the other series of tests, we vary the misorientation parameter γ, second equation of (20) or of (21) (and we set the mobility of the triple junctions $\eta = 100$, third equation of (20) or of (21)). We do not observe as much of an effect on the energy decay or average area growth in this case, but we observe the significant effect on the GBCD and the diffusion coefficient/“temperature”-like parameter D, see Figs. 5, 6, 7, 8, 9, 10, 11, 12 13, 14, 15, and 16 (with the misorientation parameter $\gamma = 1$) and Figs. 17, 18, 19 20, 21, and 22 (with larger values of the misorientation parameter γ). As concluded from our numerical results, larger values of γ give smaller diffusion coefficient/”temperature”-like parameter D,

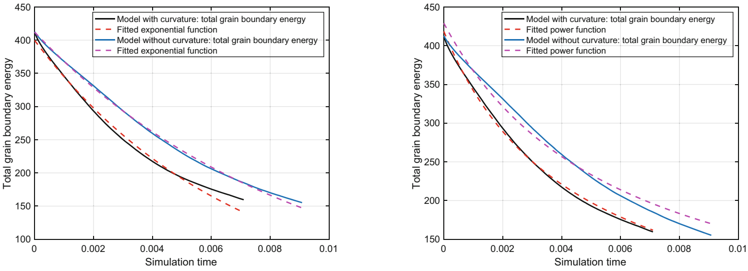


Fig. 14 One run of 2D trial with 10000 initial grains: *(a) Left plot,* total grain boundary energy plot, model with curvature (solid black) versus fitted exponential decaying function $y(t) = 399.7\exp(-147.2t)$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted exponential decaying function $y(t) = 412.1\exp(-113.3t)$ (dashed magenta); *(b) Right plot*, total grain boundary energy plot, model with curvature (solid black) versus fitted power function $y_1(t) = 418.3970(1.0 + 223.2641t)^{-1}$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted power function $y_1(t) = 428.9782(1.0 + 167.5042t)^{-1}$ (dashed magenta). Mobility of the triple junctions is $\eta = 100$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^4(2\Delta\alpha)$

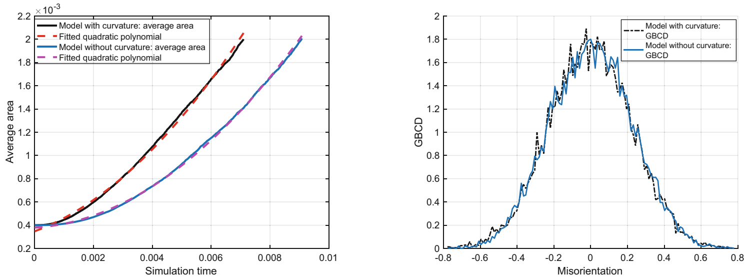


Fig. 15 *(a) Left plot,* one run of 2D trial with 10000 initial grains: Growth of the average area of the grains, model with curvature (solid black) versus fitted quadratic polynomial function $y(t) = 20.63t^2 + 0.09393t + 0.0003472$ (dashed red). Growth of the average area of the grains, model without curvature (solid blue) versus fitted quadratic polynomial function $y(t) = 18.31t^2 + 0.01553t + 0.0003786$ (dashed magenta); *(b) Right plot*, GBCD (black curve, model with curvature) and GBCD (blue curve, model without curvature) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains. Mobility of the triple junctions is $\eta = 100$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25\sin^4(2\Delta\alpha)$

and hence higher GBCD peak near misorientation 0. This is consistent with our theory that basically, larger misorientation parameter γ produces direct motion of misorientations towards equilibrium state of zero misorientations, see Sect. 2 and also [13]. Furthermore, from all of our numerical experiments with dynamic misorientations and with different triple junction mobilities, we observe that the GBCD at time T_∞ is well-approximated by the Boltzmann distribution for the grain boundary energy density see Figs. 7, 10, 13, 16, 19, and 22, as well as consistent with experimental findings as discussed in Sect. 4.1, which is similar to the work in [4–6, 8], but more detailed analysis needs to be done for a system that

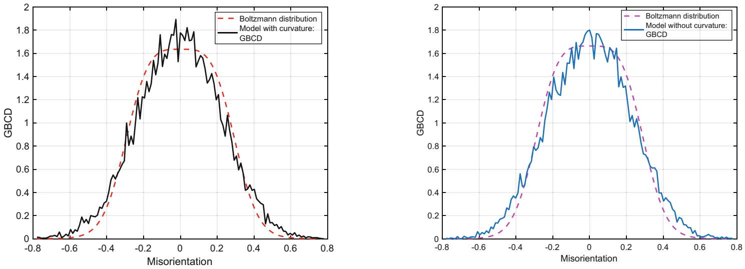


Fig. 16 *(a) Left plot,* model with curvature, GBCD (black curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.037$ (dashed red curve). *(b) Right plot*, model without curvature, GBCD (blue curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.035$ (dashed magenta curve). Mobility of the triple junctions is $\eta = 100$ and the misorientation parameter $\gamma = 1$. Grain boundary energy density $\sigma = 1 + 0.25 \sin^4(2\Delta\alpha)$

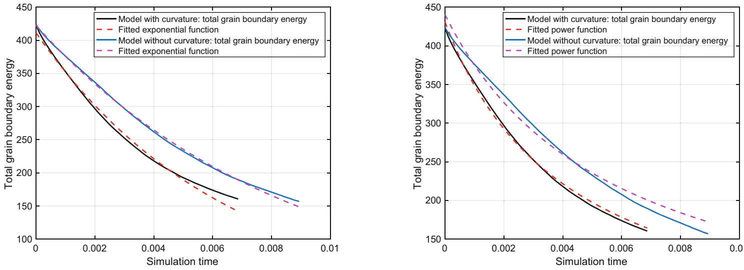


Fig. 17 One run of 2D trial with 10000 initial grains: *(a) Left plot,* total grain boundary energy plot, model with curvature (solid black) versus fitted exponential decaying function $y(t) = 411.2 \exp(-154.9t)$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted exponential decaying function $y(t) = 422.2 \exp(-116.8t)$ (dashed magenta); *(b) Right plot*, total grain boundary energy plot, model with curvature (solid black) versus fitted power function $y_1(t) = 430.8310(1.0 + 236.0718t)^{-1}$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted power function $y_1(t) = 440.1947(1.0 + 173.8526t)^{-1}$ (dashed magenta). Mobility of the triple junctions is $\eta = 100$, the misorientation parameter $\gamma = 250$ (curvature model) and $\gamma = 300$ (vertex model). Grain boundary energy density $\sigma = 1 + 0.25 \sin^2(2\Delta\alpha)$

undergoes critical events to understand the relation between GBCD, "temperature"-like/diffusion parameter D, and different relaxation time scales, as well as the effect of the time scales on the dissipation mechanism and certain coarsening rates.

Remark 6 Note that, we performed 3 runs for each numerical test presented in this work. We report results of a single run for the energy decay and growth of the average area (the results from the other two runs for each test were very similar to the presented ones), and we illustrate averaged over the 3 runs the GBCD statistics. The curve-fitting for the energy and the average area plots was done using Matlab [27] toolbox cftool.

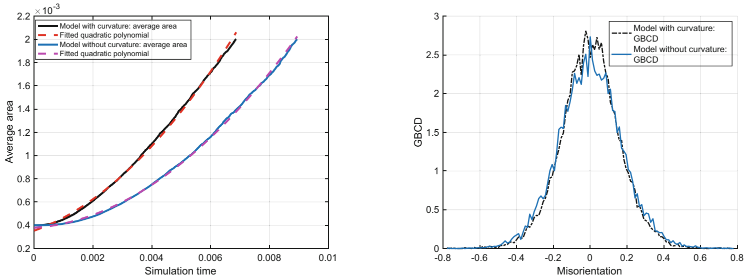


Fig. 18 *(a) Left plot,* one run of 2D trial with 10000 initial grains: Growth of the average area of the grains, model with curvature (solid black) versus fitted quadratic polynomial function $y(t) = 23.18t^2 + 0.08941t + 0.0003532$ (dashed red). Growth of the average area of the grains, model without curvature (solid blue) versus fitted quadratic polynomial function $y(t) = 18.56t^2 + 0.01824t + 0.000378$ (dashed magenta); *(b) Right plot*, GBCD (black curve, model with curvature) and GBCD (blue curve, model without curvature) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains. Mobility of the triple junctions is $\eta = 100$, the misorientation parameter $\gamma = 250$ (curvature model) and $\gamma = 300$ (vertex model). Grain boundary energy density $\sigma = 1 + 0.25\sin^2(2\Delta\alpha)$

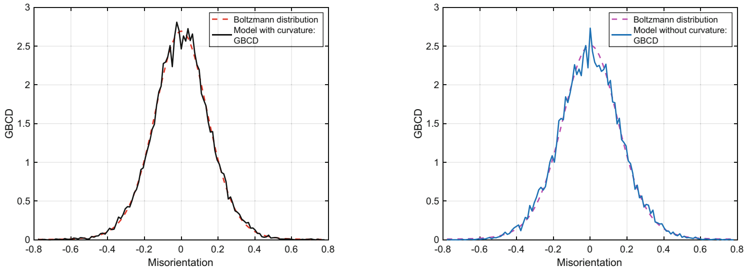


Fig. 19 *(a) Left plot,* model with curvature, GBCD (black curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.0397$ (dashed red curve). *(b) Right plot*, model without curvature, GBCD (blue curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.0448$ (dashed magenta curve). Mobility of the triple junctions is $\eta = 100$, the misorientation parameter $\gamma = 250$ (model with curvature) and $\gamma = 300$ (model without curvature). Grain boundary energy density $\sigma = 1 + 0.25\sin^2(2\Delta\alpha)$

5 Conclusion

In this work, we conducted extensive numerical studies of the two models developed in [13, 14]: a model with curved grain boundaries and a model without curvature/"vertex model" of planar grain boundaries network with the dynamic lattice misorientations and with the drag of triple junctions. The goal of our study was to further understand the effect of relaxation time scales, e.g., of the curvature of grain

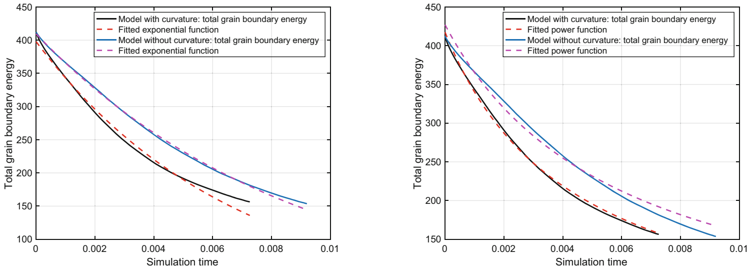


Fig. 20 One run of 2D trial with 10000 initial grains: *(a) Left plot,* total grain boundary energy plot, model with curvature (solid black) versus fitted exponential decaying function $y(t) = 397.8\exp(-147.8t)$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted exponential decaying function $y(t) = 409.5\exp(-113.2t)$ (dashed magenta); *(b) Right plot*, total grain boundary energy plot, model with curvature (solid black) versus fitted power function $y_1(t) = 417.4031(1.0 + 226.6032t)^{-1}$ (dashed red). Total grain boundary energy plot, model without curvature (solid blue) versus fitted power function $y_1(t) = 427.0061(1.0 + 168.5772t)^{-1}$ (dashed magenta) Mobility of the triple junctions is $\eta = 100$, the misorientation parameter $\gamma = 1000$ (curvature model) and $\gamma = 1500$ (vertex model). Grain boundary energy density $\sigma = 1 + 0.25\sin^4(2\Delta\alpha)$

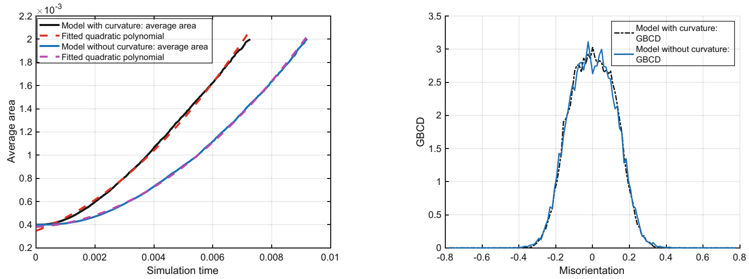


Fig. 21 *(a) Left plot,* one run of 2D trial with 10000 initial grains: Growth of the average area of the grains, model with curvature (solid black) versus fitted quadratic polynomial function $y(t) = 19.81t^2 + 0.09408t + 0.0003476$ (dashed red). Growth of the average area of the grains, model without curvature (solid blue) versus fitted quadratic polynomial function $y(t) = 17.81t^2 + 0.01484t + 0.0003807$ (dashed magenta); *(b) Right plot*, GBCD (black curve, model with curvature) and GBCD (blue curve, model without curvature) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains. Mobility of triple junctions is $\eta = 100$, the misorientation parameter $\gamma = 1000$ (curvature model) and $\gamma = 1500$ (vertex model). Grain boundary energy density $\sigma = 1 + 0.25\sin^4(2\Delta\alpha)$

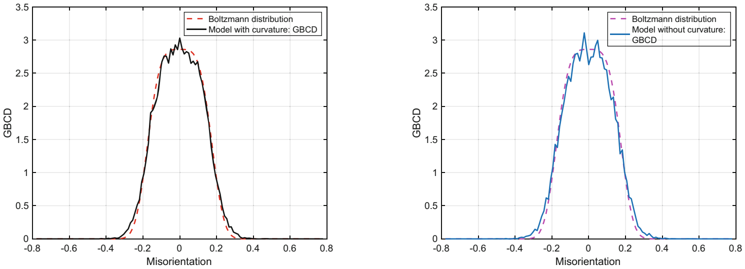


Fig. 22 *(a) Left plot,* model with curvature, GBCD (black curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.005$ (dashed red curve). *(b) Right plot*, model without curvature, GBCD (blue curve) at T_∞ averaged over 3 runs of 2D trials with 10000 initial grains versus Boltzmann distribution with "temperature"- $D \approx 0.005$ (dashed magenta curve). Mobility of the triple junctions is $\eta = 100$, the misorientation parameter $\gamma = 1000$ (model with curvature) and $\gamma = 1500$ (model without curvature). Grain boundary energy density $\sigma = 1 + 0.25 \sin^4(2\Delta\alpha)$

boundaries, mobility of triple junctions, and dynamics of misorientations on how the grain boundary system decays energy and coarsens with time. We also presented and discussed relevant experimental results of grain growth in thin films.

Acknowledgments The authors are grateful to David Kinderlehrer for the fruitful discussions, inspiration, and motivation of the work. Matthew Patrick and Amirali Zangiabadi are thanked for assistance with the experimental work. Katayun Barmak acknowledges partial support of NSF DMS-1905492, Yekaterina Epshteyn acknowledges partial support of NSF DMS-1905463, Chun Liu acknowledges partial support of NSF DMS-1950868, and Masashi Mizuno acknowledges partial support of JSPS KAKENHI Grant No. 18K13446, 22K03376.

References

1. H. Abels, H. Garcke, L. Müller, Stability of spherical caps under the volume-preserving mean curvature flow with line tension. Nonlinear Anal. **117**, 8–37 (2015)
2. B.L. Adams, D. Kinderlehrer, W.W. Mullins, A.D. Rollett, S. Ta'asan, Extracting the relative grain boundary free energy and mobility functions from the geometry of microstructures. Scripta Mater. **38**(4), 531–536 (1998)
3. B.L. Adams, S. Ta'Asan, D. Kinderlehrer, I. Livshits, D.E. Mason, C.-T. Wu, W.W. Mullins, G.S. Rohrer, A.D. Rollett, D.M. Saylor, Extracting grain boundary and surface energy from measurement of triple junction geometry. Interface Sci. **7**(3), 321–337 (1999)
4. P. Bardsley, K. Barmak, E. Eggeling, Y. Epshteyn, D. Kinderlehrer, S. Ta'asan, Towards a gradient flow for microstructure. Atti Accad. Naz. Lincei Rend. Lincei Mat. Appl. **28**(4), 777–805 (2017)
5. K. Barmak, E. Eggeling, M. Emelianenko, Y. Epshteyn, D. Kinderlehrer, R. Sharp, S. Ta'asan, Critical events, entropy, and the grain boundary character distribution. Phys. Rev. B **83**, 134117 (2011)

6. K. Barmak, E. Eggeling, M. Emelianenko, Y. Epshteyn, D. Kinderlehrer, S. Ta'asan, Geometric growth and character development in large metastable networks. Rend. Mat. Appl. (7) **29**(1), 65–81 (2009)
7. K. Barmak, E. Eggeling, D. Kinderlehrer, R. Sharp, S. Ta'asan, A.D. Rollett, K.R. Coffey, Grain growth and the puzzle of its stagnation in thin films: The curious tale of a tail and an ear. Progr. Mater. Sci. **58**(7), 987–1055 (2013)
8. K. Barmak, E. Eggeling, M. Emelianenko, Y. Epshteyn, D. Kinderlehrer, R. Sharp, S. Ta'asan, An entropy based theory of the grain boundary character distribution. Discr. Contin. Dyn. Syst. **30**(2), 427–454 (2011)
9. K.A. Brakke, *The Motion of a Surface by Its Mean Curvature*. Mathematical Notes, vol. 20 (Princeton University Press, Princeton, NJ, 1978)
10. L. Bronsard, F. Reitich, On three-phase boundary motion and the singular limit of a vector-valued Ginzburg-Landau equation. Arch. Rational Mech. Anal. **124**(4), 355–379 (1993)
11. Y.G. Chen, Y. Giga, S. Goto, Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations. J. Differ. Geom. **33**(3), 749–786 (1991)
12. K. Ecker, *Regularity Theory for Mean Curvature Flow*. Progress in Nonlinear Differential Equations and their Applications, vol. 57 (Birkhäuser Boston, Inc., Boston, MA, 2004)
13. Y. Epshteyn, C. Liu, M. Mizuno, Large time asymptotic behavior of grain boundaries motion with dynamic lattice misorientations and with triple junctions drag. Commun. Math. Sci. **19**(5), 1403–1428 (2021)
14. Y. Epshteyn, C. Liu, M. Mizuno, Motion of grain boundaries with dynamic lattice misorientations and with triple junctions drag. SIAM J. Math. Anal. **53**(3), 3072–3097 (2021)
15. L.C. Evans, J. Spruck, Motion of level sets by mean curvature. I. J. Differ. Geom. **33**(3), 635–681 (1991)
16. H. Garcke, Y. Kohsaka, D. Ševčovič, Nonlinear stability of stationary solutions for curvature flow with triple function. Hokkaido Math. J. **38**(4), 721–769 (2009)
17. C. Herring, *Surface Tension as a Motivation for Sintering* (Springer Berlin Heidelberg, Berlin, Heidelberg, 1999), pp. 33–69
18. L. Kim, Y. Tonegawa, On the mean curvature flow of grain boundaries. Ann. Inst. Fourier (Grenoble) **67**(1), 43–142 (2017)
19. D. Kinderlehrer, I. Livshits, G.S. Rohrer, S. Ta'asan, P. Yu, Mesoscale simulation of the evolution of the grain boundary character distribution, in *Recrystallization and Grain Growth, pts 1 and 2*, vols. 467–470(Part 1-2) (2004), pp. 1063–1068
20. D. Kinderlehrer, C. Liu, Evolution of grain boundaries. Math. Models Methods Appl. Sci. **11**(4), 713–729 (2001)
21. D. Kinderlehrer, I. Livshits, S. Ta'asan, A variational approach to modeling and simulation of grain growth. SIAM J. Sci. Comput. **28**(5), 1694–1715 (2006)
22. R.V. Kohn, Irreversibility and the statistics of grain boundaries. Physics, 4, 33 (2011)
23. T. Laux, F. Otto, Convergence of the thresholding scheme for multi-phase mean-curvature flow. Calc. Var. Part. Differ. Equa. **55**(5), 74 (2016). Art. 129
24. X. Liu, A.P. Warren, N.T. Nuhfer, A.D. Rollett, K.R. Coffey, K. Barmak, Comparison of crystal orientation mapping-based and image-based measurement of grain size and grain size distribution in a thin aluminum film. Acta Mater. **79**, 138–145 (2014)
25. C. Mantegazza, *Lecture Notes on Mean Curvature Flow*. Progress in Mathematics, vol. 290 (Birkhäuser/Springer Basel AG, Basel, 2011)
26. C. Mantegazza, M. Novaga, V.M. Tortorelli, Motion by curvature of planar networks. Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) **3**(2), 235–324 (2004)
27. Matlab MathWorks Inc., Matlab. version 9.4.0 (r2018a). The MathWorks Inc., Natick, MA, 2018
28. W.W. Mullins, Two-dimensional motion of idealized grain boundaries. J. Appl. Phys. **27**(8), 900–904 (1956)
29. W.W. Mullins, Theory of thermal grooving. J. Appl. Phys. **28**(3), 333–339 (1957)
30. G.S. Rohrer, Influence of interface anisotropy on grain growth and coarsening. Annu. Rev. Mater. Res. **35**, 99–126 (2005)

31. G.S. Rohrer, X. Liu, J. Liu, A. Darbal, X. Chen, M.A. Berkson, N.T. Nuhfer, K.R. Coffey, K. Barmak, The grain boundary character distribution of a highly twinned nanocrystalline aluminum thin film compared to bulk microcrystalline aluminum. J. Mater. Sci. **52**, 9819–9833 (2017)
32. G.S. Rohrer, D.M. Saylor, B. El Dasher, B.L. Adams, A.D. Rollett, P. Wynblatt, The distribution of internal interfaces in polycrystals. Z. Metallkd. **95**, 1–18 (2004)
33. A.D. Rollett, S.-B. Lee, R. Campman, G.S. Rohrer, Three-dimensional characterization of microstructure by electron back-scatter diffraction. Annu. Rev. Mater. Res. **37**, 627–658 (2007)
34. S.L. Thomas, C. Wei, J. Han, Y. Xiang, D.J. Srolovitz, Disconnection description of triple-junction motion. Proc. Natl. Acad. Sci. **116**(18), 8756–8765 (2019)
35. C.E. Torres, M. Emelianenko, D. Golovaty, D. Kinderlehrer, S. Ta'asan, Numerical analysis of the vertex models for simulating grain boundary networks. SIAM J. Appl. Math. **75**(2), 762–786 (2015)
36. M. Upmanyu, D.J. Srolovitz, L.S. Shvindlerman, G. Gottstein, Molecular dynamics simulation of triple junction migration. Acta Mater. **50**(6), 1405–1420 (2002)
37. M. Upmanyu, D.J. Srolovitz, L.S. Shvindlerman, G. Gottstein, Triple junction mobility: A molecular dynamics study. Interface Sci. **7**(3), 307–319 (1999)
38. L. Zhang, J. Han, Y. Xiang, D.J. Srolovitz, Equation of motion for a grain boundary. Phys. Rev. Lett. **119**, 246101 (2017)
39. L. Zhang, Y. Xiang, Motion of grain boundaries incorporating dislocation structure. J. Mech. Phys. Solids **117**, 157–178 (2018)

Regularity of Minimizers for a General Class of Constrained Energies in Two-Dimensional Domains with Applications to Liquid Crystals

Patricia Bauman and Daniel Phillips

1 Introduction.

In this paper we consider minimizers to a singular constrained energy functional of the form

$$J(v) = \int_{\Omega} (F(v, Dv) + f(v))dx, \tag{1}$$

where Ω is a bounded C^2 domain in $\mathbb{R}^n$ and $n \geq 2$. We assume that f is defined and real valued on an open, bounded, convex set $\mathcal{K}$ in $\mathbb{R}^q$ with $q \geq 1$ and that $f(v) \to \infty$ as $v \to \partial\mathcal{K}$ for v in $\mathcal{K}$. We extend the definition of f to all of $\mathbb{R}^q$ by setting $f(v) = \infty$ for v in $\mathbb{R}^q \setminus \mathcal{K}$. Thus we assume throughout this paper that

$$\begin{cases} f : \mathcal{K} \to \mathbb{R}, \quad f \in C^2(\mathcal{K}), \quad D^2 f \geq -M I_q \text{ on } \mathcal{K}, \\ \lim\limits_{\substack{v \to \partial\mathcal{K} \\ v \in \mathcal{K}}} f(v) = \infty, \text{ and } f(v) = \infty \text{ on } \mathbb{R}^q \setminus \mathcal{K}, \end{cases} \tag{2}$$

where $M \geq 0$. We also assume the following structure conditions on F:

P. Bauman (✉) · D. Phillips
Purdue University, Department of Mathematics, West Lafayette, IN, USA
e-mail: baumanp@purdue.edu; phillips@purdue.edu

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_3

$$\begin{cases} F(v,P) = A_{ij}^{\alpha\beta}(v)p_\alpha^i p_\beta^j + B_i^\alpha(v)p_\alpha^i \quad \text{for } v \in \overline{\mathcal{K}},\, P \in \mathbb{M}^{q\times n}, \\ A_{ij}^{\alpha\beta}(v)p_\alpha^i p_\beta^j \geq \lambda|P|^2 \quad \text{for } v \in \overline{\mathcal{K}},\, P \in \mathbb{M}^{q\times n}, \\ \text{where } A_{ij}^{\alpha\beta},\, B_i^\alpha \in C^2(\overline{\mathcal{K}}) \ \text{ and } \lambda > 0. \end{cases} \tag{3}$$

We use the convention in this paper that repeated indices are summed. In this case i and j go from 1 to n and α and β go from 1 to q. Here $\mathbb{M}^{q\times n}$ denotes the set of $q \times n$ real-valued matrices. We define

$$M_1 = max[\{sup_{v\in\overline{\mathcal{K}}}|A_{ij}^{\alpha\beta}(v)|\}, \{sup_{v\in\overline{\mathcal{K}}}|B_i^\alpha(v)|\}]$$

and $M_2 \equiv max\{\|A_{ij}^{\alpha\beta}\|_{C^2(\overline{\mathcal{K}})}, \|B_i^\alpha\|_{C^2(\overline{\mathcal{K}})}\}$.

The energy functional J is defined for all v in

$$H^1(\Omega;\overline{\mathcal{K}}) = \{v \in H^1(\Omega;\mathbb{R}^q) : v(x) \in \overline{\mathcal{K}} \text{ almost everywhere in } \Omega\}.$$

By assumption (2) on f, if $u \in H^1(\Omega;\overline{\mathcal{K}})$ and $J(u) < \infty$, then $u(x) \in \mathcal{K}$ for almost every x in Ω.

Given $\underline{\mathrm{u}}$ in $H^1(\Omega;\overline{\mathcal{K}})$ such that $J(\underline{\mathrm{u}}) < \infty$, it follows from direct methods in the calculus of variations (see [7]) that minimizers exist in the space $A_{\underline{u}} = \{v \in H^1(\Omega;\overline{\mathcal{K}}) : v - \underline{u} \in H_0^1(\Omega;\mathbb{R}^q)\}$. In this paper we will refer to such minimizers as *finite energy minimizers* of J in Ω.

A question of interest for applications is whether minimizers u of $J(\cdot)$ are smooth and whether they satisfy $u(x) \in \mathcal{K}$ for all x in Ω. One of the difficulties in analyzing their regularity is to find finite energy variations in $H^1(\Omega;\overline{\mathcal{K}})$ from which one can extract useful information about their properties.

Our results are for $n = 2$. We prove the following main theorem:

Theorem 1 *Assume Ω is a bounded C^2 domain in $\mathbb{R}^n$, $n = 2$, $\mathcal{K}$ is an open bounded, convex set in $\mathbb{R}^q$, $q \geq 1$, and (2) and (3) hold. If $u \in H^1(\Omega;\overline{\mathcal{K}})$ is a finite energy minimizer for J in Ω, then u is in $C^{2,\delta}(\Omega)$ for all $0 < \delta < 1$, $u(\Omega) \subset \mathcal{K}$, and u satisfies the equilibrium equation*

$$\mathrm{div}\, \mathrm{F_P}(\mathrm{u}, \mathrm{Du}) - \mathrm{F_u}(\mathrm{u}, \mathrm{Du}) = \mathrm{f_u}(\mathrm{u}) \quad \textit{on } \Omega. \tag{4}$$

Moreover, if Ω' is an open set in Ω such that $\Omega' \subset\subset \Omega$, then dist $(u(\Omega'), \partial\mathcal{K}) \geq c > 0$ where c depends only on $J(u)$, Ω, dist $(\Omega', \partial\Omega)$, M_2, M, and λ.

Minimizers u of (1) were investigated by Evans, Kneuss, and Tran in [5]. Assuming that $n \geq 2$, $F = F(v, Dv)$ satisfies certain growth conditions and is uniformly strictly quasi-convex, and f satisfies (2) with $M = 0$, they proved the following partial regularity result: there is an open subset Ω_0 of Ω such that

$|\Omega \setminus \Omega_0| = 0$, $u \in C^2(\Omega_0)$, and $u(\Omega_0) \subset \mathcal{K}$. In particular it follows that u satisfies the equilibrium equation (4) on Ω_0 in this case.

They also proved in [5] that if, in addition, $F = F(Dv)$, $F(\cdot)$ is convex, and $f(v)$ is smooth and convex on $\mathcal{K}$, then a finite energy minimizer u is in $H^2_{loc}(\Omega)$. We use a similar approach here for part of our analysis. They considered variations $u + t\phi_k$ where ϕ_k is given by

$$\begin{aligned}\phi_k(x;h) &= [\zeta^2(x)u(x+he_k) + \zeta^2(x-he_k)u(x-he_k) - (\zeta^2(x)+\zeta^2(x-he_k))u(x)]h^{-2}\\ &= \nabla_k^{-h}(\zeta^2\nabla_k^h u) \quad \text{for } 1 \le \text{k} \le \text{n},\end{aligned}$$

for $\zeta \in C^2_c(\Omega)$ and $\nabla^h_k w(x) = h^{-1}[w(x+he_k) - w(x)]$ for $h \neq 0$ and sufficiently small. They showed that for $0 \le t < \bar{t}(h)$, $u(x) + t\phi_k(x) \in \overline{\mathcal{K}}$. Using the definition of ϕ_k and the convexity of f they proved that

$$\int_\Omega f(u+t\phi_k)dx \le \int_\Omega f(u)dx \quad \text{for } 0 \le \text{t} < \bar{\text{t}}(\text{h}),$$

and hence

$$0 \le J(u+t\phi_k) - J(u) \le \int_\Omega (F(D(u+t\phi_k)) - F(Du))dx.$$

Dividing by t and letting $t \to 0$ gives

$$0 \le \int_\Omega (F_P(Du) : D\phi_k)dx \text{ for } 1 \le k \le n.$$

For $F = F(Dv)$ this inequality leads to $u \in H^2_{\text{loc}}(\Omega)$. (See [5], Thm 4.1.)

Here we have $F = F(v, Dv)$ and f is not convex. However, our assumption (2) implies that

$$f_0(v) \equiv f(v) + \frac{M}{2}|v|^2$$

is a convex function on $\mathcal{K}$ and by [5]

$$\int_\Omega f_0(u+t\phi_k)dx \le \int_\Omega f_0(u)dx \quad \text{for } 0 \le \text{t} < \bar{\text{t}}(\text{h}). \tag{5}$$

Using this we can argue just as above to show that a minimizer u of J satisfies

$$0 \le \int\limits_{\Omega} (F_P(u, Du) : D\phi_k + F_u(u, Du) \cdot \phi_k - Mu \cdot \phi_k) dx. \tag{6}$$

From this and (3) it follows that

$$\frac{\lambda}{2} \int\limits_{\Omega} |\nabla^h Du|^2 \zeta^2 \, dx \le C_0 \int\limits_{\Omega} |\nabla^h u|^2 |Du|^2 \zeta^2 \, dx + C_1, \tag{7}$$

where λ is the constant defined in our assumption (3).

To obtain an H^2_{loc} estimate, we additionally need that u is continuous and a second inequality. In Sect. 2, we show that when $n = 2$, our assumptions (2), (3), and a result in [4] imply that finite energy minimizers of J are continuous in Ω. We then construct additional variations $u + tw_l$ in $H^1(\Omega; \overline{\mathcal{K}})$ so that w_l satisfies (5) and (6), with ϕ_k replaced by w_l. We use this to prove a second inequality, from which we obtain the H^2_{loc} regularity of minimizers. In Sect. 3 we use this result to prove Theorem 1.

Our formulation of the constrained energy (1) and the assumptions (2) and (3) is motivated by the constrained Landau-de Gennes Q-tensor energy for nematic liquid crystals. This energy is given by

$$I_{LdG}[Q] = \int\limits_{\Omega} \left[G(Q, DQ) + \Psi_b(Q)\right] dx, \tag{8}$$

where

$$\begin{aligned} G(Q, DQ) =& L_1 |DQ|^2 + L_2 \cdot D_{x_j} Q_{ij} \cdot D_{x_k} Q_{ik} + L_3 \cdot D_{x_j} Q_{ik} \cdot D_{x_k} Q_{ij} \\ &+ L_4 \cdot Q_{\ell k} \cdot D_{x_\ell} Q_{ij} \cdot D_{x_k} Q_{ij} + L_5 \cdot \epsilon_{\ell jk} \cdot Q_{\ell i} \cdot D_{x_j} Q_{ki} \\ &\equiv L_1 I_1 + L_2 I_2 + L_3 I_3 + L_4 I_4 + L_5 I_5 \end{aligned}$$

and $\Psi_b(Q) = T f_{ms}(Q) - \kappa |Q|^2$. The constants L_1, L_2, L_3, L_4, and L_5 are material-dependent elastic constants, T and κ are positive constants, and $\epsilon_{\ell jk}$ is the Levi-Civita tensor. The function $f_{ms}(Q)$ is a specific function (called the Maier–Saupe potential) defined on

$$\mathcal{M} = \{Q \in \mathbb{M}^{3\times 3} : Q = Q^t, \text{ tr } Q = 0, \text{ and } -\frac{1}{3} < \lambda(Q) < \frac{2}{3} \\ \text{for all eigenvalues } \lambda(Q) \text{ of } Q\}.$$

It is defined abstractly using probability densities on a sphere that represent possible orientations of liquid crystal molecules. (See Sect. 4 for the definition of f_{ms}.) The bulk term Ψ_b and the Maier–Saupe potential f_{ms} were introduced and investigated in the papers [3] by Ball and Majumdar and [9] by Katriel, Kventsel, Luckhurst, and Sluckin. It is known that f_{ms} is convex. Moreover, f_{ms} is bounded below and $f_{ms}(Q) \to \infty$ for Q in $\mathcal{M}$ with $Q \to \partial\mathcal{M}$; hence the same is true for $\Psi_b(Q)$. As in (2), we set $\Psi_b(Q) = \infty$ for $Q \in S_0 \setminus \mathcal{M}$ where $S_0 = \{Q \in \mathbb{M}^{3\times 3} : Q = Q^t, tr\ Q = 0\}$. Thus Ψ_b blows up at $\partial\mathcal{M}$ as in (2), with f replaced by Ψ_b and $\mathcal{K}$ replaced by $\mathcal{M}$. It follows that finite energy minimizers Q of I_{LdG} satisfy $Q(x) \in \mathcal{M}$ almost everywhere in Ω, so that the eigenvalues of $Q(x)$ are in $(-\frac{1}{3}, \frac{2}{3})$ almost everywhere in Ω. Conditions on $L_1, \cdots, L_4$ have been identified so that minimizers of I_{LdG} exist in

$$A_{Q_0} \equiv \{Q \in H^1(\Omega; \overline{\mathcal{M}}) : Q - Q_0 \in H_0^1(\Omega; \mathbb{M}^{3\times 3})\}$$

provided that $Q_0 \in H^1(\Omega; \overline{\mathcal{M}})$ and that $I_{LdG}[Q_0] < \infty$. (See (38) and [10].) It was stated in [3] that for Ω in $\mathbb{R}^3$, if $Q_0(\overline{\Omega}) \subset\subset \mathcal{M}$, minimizers in A_{Q_0} of the energy

$$\int\limits_{\Omega} [L_1|DQ)|^2 + \Psi_b(Q)]dx$$

with $L_1 > 0$ are smooth in Ω and valued in $\mathcal{M}$; thus their eigenvalues are in $(-\frac{1}{3}, \frac{2}{3})$ at all points in Ω. A sketch of a proof of this statement is included in [1]. (See also [4].) Such minimizers are called "physically realistic." Additional features for minimizers, $\tilde{Q}$ of I_{LdG} with Ω in $\mathbb{R}^3$, were obtained by Geng and Tong in [8]. In particular, assuming specific conditions on $G(Q, DQ)$ they proved higher integrability properties for $|D\tilde{Q}|$.

The elastic term with coefficient L_4 in I_{LdG} is called the "cubic term." When $L_4 \neq 0$, the energy density is quasilinear. This makes it difficult to analyze the behavior of minimizers in this case.

Physicists have computed the elastic coefficients $L_1, \cdots, L_4$ in terms of the elastic coefficients $K_1, \cdots, K_4$ that account for the elastic energy of splay, twist, and bend that occur in the well-known Frank energy density, which models liquid crystals in terms of functions $n = n(x)$ valued in $\mathbb{S}^2$. They found that $L_4 = 0$ if and only if $K_1 = K_3$, which is nonphysical for many applications. Thus it is desirable to consider the energy I_{LdG} with $L_4 \neq 0$. It is interesting to note that when $L_4 \neq 0$, the unconstrained Landau-de Gennes energy given by (8) with $\Psi_b(Q)$ replaced by a polynomial is unbounded from below. Thus minimizers do not exist in general for boundary value problems with this energy. See [3].

For Ω in $\mathbb{R}^2$, we proved Hölder continuity of finite energy minimizers in [4] under general conditions for energy functionals of the form

$$\mathcal{F}(Q) = \int_{\Omega} [F_e(Q(x), DQ(x)) + f_b(Q(x))]dx$$

by using harmonic and elliptic replacements to construct finite energy comparison functions. In particular we established that finite energy minimizers to the quasilinear constrained energy I_{LdG} in (8) under the coercivity condition (38) are Hölder continuous in Ω. We also proved under the additional assumption $L_2 = L_3 = 0$ that finite energy minimizers for (8) satisfy the "physicality condition," $Q(x) \in \mathcal{M}$ for all $x \in \Omega$. Here we establish this property without requiring the additional assumption.

In Sect. 4 of this paper we describe a connection between the constrained energies $J[u]$ and $I_{LdG}[Q]$. Using Theorem 1, we prove in Theorem 2 that under appropriate coercivity conditions on $L_1, \cdots, L_4$, finite energy minimizers of I_{LdG} with all elasticity terms are in $C^2(\Omega)$; moreover, they satisfy a strong physicality condition: if Ω' is an open set such that $\Omega' \subset\subset \Omega$, then $Q(\Omega') \subset \mathcal{M}$ and $\operatorname{dist}(Q(\Omega'), \partial\Omega) \geq c > 0$. Thus in compact subsets of Ω, the eigenvalues of minimizers are contained in closed subintervals of the physical range $(-\frac{1}{3}, \frac{2}{3})$.

2 Continuity and H^2_{loc} Estimates for Minimizers in Two-Dimensional Domains

Assume that Ω is a bounded C^2 domain in $\mathbb{R}^2$. Let $\Lambda = \{x \in \Omega : u(x) \in \partial\mathcal{K}\}$. In this section we will show that finite energy minimizers u of J are locally Hölder continuous in Ω, C^2 in $\Omega \setminus \Lambda$, and globally Hölder continuous in $\overline{\Omega}$ if their boundary values are sufficiently smooth. In addition, they are in $H^2_{\text{loc}}(\Omega)$.

Our proof of the first statement is an application of (2), (3), and a result in [4] for two-dimensional domains. We will ultimately prove (in Sect. 3) that $\Lambda = \emptyset$.

Proposition 1 *Assume that Ω is a bounded C^2 domain in $\mathbb{R}^2$. Assume $u = u(x)$ is in $H^1(\Omega; \overline{\mathcal{K}})$ and u is a finite energy minimizer of J in Ω. Let $\Lambda = \{x \in \Omega : u(x) \in \partial\mathcal{K}\}$.*

(a) If Ω' is a connected open set with $\Omega' \subset\subset \Omega$, there exist constants $0 < \sigma < 1$ and $c_1 > 0$ such that $\omega(d) = c_1 d^\sigma$ is a modulus of continuity for u in Ω'. The constants σ and c_1 depend only on $J(u)$, Ω, dist $(\Omega', \partial\Omega)$, M_1, and the constants M and λ in (2) and (3).

(b) If $u_0 \in H^1(\Omega; \overline{\mathcal{K}})$ and $J(u_0) < \infty$ such that $u_0 \in C^{0,1}(\partial\Omega; \overline{\mathcal{K}})$ and $\int_{\partial\Omega} f(u_0)ds < \infty$ and if $u \in H^1(\Omega; \overline{\mathcal{K}})$ is a minimizer of J in $A_{u_0} = \{v \in H^1(\Omega; \overline{\mathcal{K}}) : v - u_0 \in H^1_0(\Omega; \mathbb{R}^q)\}$, then there exists a constant $0 < \beta < 1$ such that $u \in C^\beta(\overline{\Omega}; \overline{\mathcal{K}})$. The modulus of continuity, $\omega(d) = c_2 d^\beta$, has constants depending only on $J(u)$, Ω, M_1, M, λ and u_0.

(c) The minimizer u is continuous in Ω and $C^{2,\delta}$ in the open set $\Omega \setminus \Lambda$ for all $0 < \delta < 1$.

Proof In [4] we investigated finite energy minimizers $\tilde{Q}(x)$ of a constrained energy of the form

$$\tilde{J}(Q) = \int_\Omega (F_e(Q, DQ) + f_b(Q))dx$$

over all $Q \in H^1(\Omega; \overline{\mathcal{M}})$ such that $Q - Q_0 \in H_0^1(\Omega; S_0)$ and $\tilde{J}(Q_0) < \infty$, where $\mathcal{M}$ is an open bounded convex subset of $S_0 = \{Q \in \mathbb{M}^{3\times 3} : Q = Q^t \text{ and tr } Q = 0\}$ and $\mathbb{M}^{3\times 3}$ is the set of 3×3 real-valued matrices. Note that S_0 is isometrically isomorphic to $\mathbb{R}^5$. Here $f_b(Q) = g_b(Q) - \kappa|Q|^2 + b_0$ for $Q \in S_0$ (and ∞ otherwise) and it is assumed that g_b is a smooth convex function defined on $\mathcal{M}$ such that $g_b(Q) \to \infty$ as $Q \to \partial\mathcal{M}$ with $Q \in \mathcal{M}$. Since $f(v) = f_0(v) - \frac{M}{2}|v|^2$, our assumptions (2) and (3) on $f(v)$ and $F(v, Dv)$ correspond to the assumptions (1.2) and (1.4) on $f_b(Q)$ and $F_e(Q, DQ)$ in [4] that were used to prove the same Hölder continuity on $\tilde{Q} = \tilde{Q}(x)$ that we wish to prove here for $u = u(x)$. The change from energy densities that depend on the variable $Q \in S_0$ to those that depend on $u \in \mathbb{R}^q$ is a trivial one, and the arguments in the proofs of Theorem 1 and 2 go through to prove (a) and (b).

To prove (c), we first note that by (a), u is continuous in Ω and hence $u^{-1}(\mathcal{K}) = \Omega \setminus \Lambda$ is an open set. To verify that u is $C^{2,\delta}$ on this set, we argue as in [4], Corollary 2. Indeed, assume $B_{4r}(x_0) \subset\subset \Omega \setminus \Lambda$. Note that f is bounded and C^2 on a neighborhood of $u(B_{4r}(x_0))$. We can then take smooth first variations for J about u supported in $B_{4r}(x_0)$ to conclude that u is a weak solution of (4) on $B_{4r}(x_0)$. We can apply the result from [7], Ch. VI, Proposition 1 asserting that in two space dimensions a continuous weak solution of (3)–(4) with $f_u(u(x))$ bounded is in $W^{2,p}(B_{3r}(x_0))$ for some $p > 2$ and thus its first derivatives are Hölder continuous on $B_{2r}(x_0)$. Now we can apply techniques from linear elliptic theory in [7], Ch. III. Taking (3) into account these lead to $u \in C^{2,\delta}(B_r(x_0))$. □

Our next objective is to show that $u \in H^2_{\text{loc}}(\Omega)$. To define variations $u + tw_l$ that will provide a proof, we will need several properties of the convex potential

$$f_0(v) = f(v) + \frac{M}{2}|v|^2$$

in a family of cones C with vertices in the convex set $\mathcal{K}$. For ease of notation, assume from now on without loss of generality that 0 is in $\mathcal{K}$. Since $\mathcal{K}$ is a bounded convex set in $\mathbb{R}^q$, it is starlike with respect to 0. Let $\mathbb{S}^{q-1} = \partial B_1(0)$ where $B_1(0)$ is the open unit ball in $\mathbb{R}^q$ centered at 0. Let $g : \mathbb{S}^{q-1} \to \mathbb{R}^+$ be in $C^{0,1}(\mathbb{S}^{q-1})$ such that the map

$$\nu \in \mathbb{S}^{q-1} \to g(\nu)\nu \in \mathbb{R}^q$$

is a parametrization of $\partial\mathcal{K}$. Define $0 < m_1 < m_2$ by

$$m_1 = inf\{g(\nu) : \nu \in \mathbb{S}^{q-1}\} \text{ and } m_2 = sup\{g(\nu) : \nu \in \mathbb{S}^{q-1}\}. \tag{9}$$

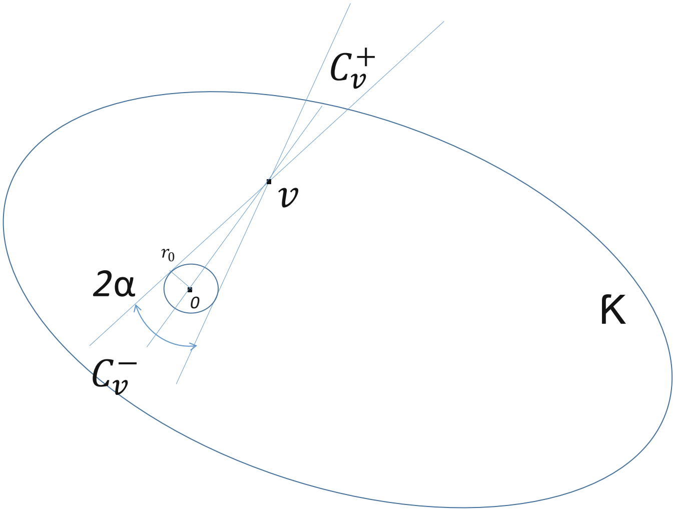


Fig. 1 The cones C_v^- and C_v^+

Define $G(x) : \overline{B_1(0)} \to \overline{\mathcal{K}}$ by

$$G(x) = \begin{cases} g(\frac{x}{|x|})x & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases} \tag{10}$$

Thus G is a bi-Lipschitz continuous map from $\overline{B_1(0)}$ onto $\overline{\mathcal{K}}$. Let $\mathcal{Y}_\mu = G(B_\mu(0))$ for $0 < \mu < 1$. Then $\mathcal{Y}_\mu$ is an open convex subset of $\mathcal{K}$ and $\mathcal{Y}_\mu \uparrow \mathcal{K}$ as $\mu \uparrow 1$. Fix $r_0 > 0$ and $0 < \mu_0 < 1$ so that $\overline{B_{r_0}(0)} \subset \mathcal{Y}_{\mu_0}$.

Definition We define a family of cones C as follows: For each v in $\mathcal{K} \setminus \overline{B_{r_0}(0)}$, we define the cone C_v^- to be the closed half-cone with vertex v, axis containing the ray from v to 0, and aperture $\alpha = \alpha(v)$ in $(0, \frac{\pi}{2})$ determined by $\sin\alpha = \frac{r_0}{|v|}$. (See Fig. 1). The cone C_v^+ is the reflection of C_v^- about the point v, i.e.,

$$C_v^+ = \{w = v + \xi : \xi \in \mathbb{R}^q \text{ and } v - \xi \in C_v^-\}.$$

We define C to be the family of all cones, C_v^- and C_v^+, with v in $\mathcal{K} \setminus \overline{B_{r_0}(0)}$.

For v as above, the ball $\overline{B_{r_0}(0)}$ is contained in C_v^- and is tangent to its boundary. Thus each ray in C_v^- with initial point v intersects $\partial B_{r_0}(0)$ at least once. Also $r_0 < |v| \leq m_2$ and thus there exists $\alpha_0 \in (0, \frac{\pi}{2})$ such that

$$1 > \sin\alpha \geq \frac{r_0}{m_2} \equiv \sin\alpha_0 > 0 \text{ for all } v \in \mathcal{K} \setminus \overline{B_{r_0}(0)}. \tag{11}$$

The result below follows from (11) and the symmetry of C_v^- and C_v^+.

Proposition 2 *Assume $v \in \mathcal{K} \setminus \overline{B_{r_0}(0)}$, z is a point on the axis of C_v^+ with $z \neq v$, and γ satisfies $|z - v| \geq \gamma > 0$. If $r > 0$ satisfies $r \leq \gamma \sin \alpha_0$, then*

$$\overline{B_r(z)} \subset C_v^+. \tag{12}$$

Proof The hypotheses ensure that

$$r \leq \gamma \sin \alpha \leq |z - v| \sin \alpha$$

for $\alpha = \alpha(v)$. It follows from this and the symmetry of C_v^- and C_v^+ that $\overline{B_r(z)} \subset C_v^+$. □

Proposition 3 *If $v \in \mathcal{K} \setminus \mathcal{Y}_\mu$ and $\mu > \mu_0$, then $C_v^+ \cap \mathcal{K} \subset \mathcal{K} \setminus \mathcal{Y}_\mu$.*

Proof If not, there exists $w \in C_v^+ \cap \mathcal{K}$ such that $w \notin \mathcal{K} \setminus \mathcal{Y}_\mu$ and hence $w \in \mathcal{Y}_\mu$. Thus $w \neq v$ and $w = v + \xi$ for some $\xi \neq 0$. The ray with initial point v that passes through $v - \xi$ is in C_v^- and hence contains a point y in $\partial B_{r_0}(0) \cap C_v^-$. Since $\mu > \mu_0$, $\partial B_{r_0}(0) \subset \mathcal{Y}_\mu$. By the convexity of $\mathcal{Y}_\mu$, the segment $\overline{yw}$ is contained in $\mathcal{Y}_\mu$. But $v \in \overline{yw} \subset \mathcal{Y}_\mu$, which contradicts the fact that $v \in \mathcal{K} \setminus \mathcal{Y}_\mu$. □

Define $s_0 = \max\{f_0(v) : v \in \partial B_{r_0}(0)\}$. Since f_0 is continuous on $\mathcal{K}$ and $f_0(v) \to \infty$ as $v \to \partial \mathcal{K}$ with v in $\mathcal{K}$, there exists a constant μ_1 in $(\mu_0, 1)$ such that

$$f_0(v) \geq 1 + s_0 \text{ for all } v \text{ in } \mathcal{K} \setminus \mathcal{Y}_{\mu_1}. \tag{13}$$

From now on, we fix $0, r_0, \alpha_0, m_1, m_2, 0 < \mu_0 < \mu_1 < 1$, and s_0 as above. We then have the following:

Lemma 1 *For any v in $\mathcal{K} \setminus \mathcal{Y}_{\mu_1}$ and w in $C_v^+ \cap \mathcal{K}$ such that $w \neq v$, we have $\nabla f_0(w) \cdot (w - v) > 0$.*

Proof If this is false, there exists v in $\mathcal{K} \setminus \mathcal{Y}_{\mu_1}$ and $w \in C_v^+ \cap \mathcal{K}$ such that $w \neq v$ and $\nabla f_0(w) \cdot (w - v) \leq 0$. By Proposition 3, $w \in \mathcal{K} \setminus \mathcal{Y}_{\mu_1}$. Consider a linear path given by $p(t) = v + t \frac{w-v}{|w-v|}$ for $\underline{t} \leq t \leq \bar{t}$ where $p(\underline{t}) \in \partial B_{r_0}(0)$ and $\bar{t} = |v - w|$. Setting $h(t) = f_0(p(t))$ we have $h'(\bar{t}) \leq 0$. By definition of m_2 (see (9)), we have

$$0 < \bar{t} - \underline{t} = |p(\bar{t}) - p(\underline{t})| = |w - p(\underline{t})| \leq |w| + |p(\underline{t})| \leq 2m_2.$$

Since $h(t)$ is convex, we then have

$$0 \geq h'(\bar{t}) \geq \frac{h(\bar{t}) - h(\underline{t})}{\bar{t} - \underline{t}} \geq \frac{h(\bar{t}) - h(\underline{t})}{2m_2}.$$

This is impossible since $m_2 > 0$ and by definition of s_0,

$$h(\bar{t}) - h(\underline{t}) = f_0(w) - f_0(p(\underline{t})) \geq (1 + s_0) - s_0 = 1.$$

□

Corollary 1 *If $v \in \mathcal{K} \setminus \mathcal{Y}_{\mu_1}$ and $\mathbf{l} = \mathbf{wv}$ is a ray in $C_v^+ \cap \mathcal{K}$ with initial point w and final point v, then f_0 decreases along $\mathbf{l}$.*

We can now prove H^2_{loc} estimates on minimizers using appropriate variations $u + t\phi$ and $u + tw$.

Lemma 2 *Assume that Ω' is an open connected set and $\Omega' \subset\subset \Omega$. If $u \in H^1_{loc}(\Omega; \mathcal{K})$ is a finite energy local minimizer of J in Ω, then $u \in H^2(\Omega')$ and*

$$\|u\|_{H^2(\Omega')} \leq C_o,$$

where C_o depends only on $J(u)$, Ω, $dist(\Omega', \partial\Omega)$, M_2, M and λ.

Proof Let $\eta_0 > 0$ satisfy $3\sqrt{2}\eta_0 < \frac{1}{4}$ dist $(\Omega', \partial\Omega)$. Define

$$\Omega'' = \{x \in \Omega : \text{ dist } (x, \Omega') < \frac{1}{2} \text{ dist } (\Omega', \partial\Omega)\} \tag{14}$$

and let $\omega(d) = Cd^\sigma$ be a modulus of continuity for u on Ω''. Given $\eta > 0$ with $\eta < \eta_0$, let $\mathcal{D}_0 = \{D_m : m \in \mathbb{N}\}$ be a tiling of $\mathbb{R}^2$ by closed squares of side length η, so that $\mathbb{R}^2 = \cup_{m=1}^{\infty} D_m$ and distinct squares in this tiling have at most one edge in common. For each $m \in \mathbb{N}$ let E_m be the union of D_m and its eight neighbors. Thus E_m and D_m have the same centers and E_m has side length 3η. Set

$$\mathcal{D}_1 = \{D_m \in \mathcal{D}_0 : D_m \cap \Omega' \neq \emptyset\}.$$

Then $\mathcal{D}_1 = \{D_{m_l} : 1 \leq l \leq L\}$. For ease of notation, let $\tilde{D}_l$ and $\tilde{E}_l$ denote D_{m_l} and E_{m_l}, respectively, for $1 \leq l \leq L$. Note that by our definition of η_0, $\tilde{E}_l \subset \Omega''$ for $1 \leq l \leq L$.

We first work through the case $q \geq 2$. Let $\mu \in [\mu_1, 1)$. Thus $\mu_1 \leq \mu \leq \frac{\mu+3}{4} < 1$. (Later we will also require that $(1 - \mu)$ is sufficiently small.) We partition $\mathcal{D}_1 = \mathcal{D}_2 \cup \mathcal{D}_3$ as follows:

$$\tilde{D}_l \in \mathcal{D}_2 \text{ if } u(\tilde{E}_l) \cap (\overline{\mathcal{K}} \setminus \mathcal{Y}_{\frac{\mu+3}{4}}) \neq \emptyset,$$

$$\tilde{D}_l \in \mathcal{D}_3 \text{ if } u(\tilde{E}_l) \subset \mathcal{Y}_{\frac{\mu+3}{4}}.$$

We first assume $\tilde{D}_l \in \mathcal{D}_2$ and show that $u \in H^2(\tilde{D}_l)$. Note that in this case, $u(\tilde{E}_l)$ is not a subset of $\partial\mathcal{K}$ because if so, $\tilde{E}_l \subset \Lambda$ (by (1) and (2) since $J(u) < \infty$) and this would imply that $|\tilde{E}_l| = 0$, a contradiction. From this and the definition of $\mathcal{D}_2$ it follows that there exists $\overline{y} \in \tilde{E}_l$ such that $u(\overline{y}) \in \mathcal{K} \setminus \mathcal{Y}_{\frac{\mu+3}{4}}$. Thus

$$u(\overline{y}) = \tilde{\mu} g(\overline{\nu})\overline{\nu} \text{ for } \overline{\nu} = \frac{u(\overline{y})}{|u(\overline{y})|} \in S^{q-1}$$

and some constant $\tilde{\mu}$ depending on $\overline{y}$ such that $\frac{\mu+3}{4} < \tilde{\mu} < 1$. Set $v_l = \mu g(\overline{\nu})\overline{\nu}$. Recall that $\mu_0 < \mu_1 \leq \mu < \frac{\mu+3}{4}$; hence v_l is on the segment between 0 and $u(\overline{y})$ and is contained in $\mathcal{K}$ by convexity. Also $v_l \in \mathcal{K} \setminus \mathcal{Y}_{\mu_1} \subset \mathcal{K} \setminus \overline{B_{r_0}(0)}$. Thus the cone $C^+_{v_l}$ is in our family of cones C and $u(\overline{y})$ is on the axis of $C^+_{v_l}$. Since $m_2 \geq g(\overline{\nu}) \geq m_1$ and $1 - \mu > \tilde{\mu} - \mu > \frac{\mu+3}{4} - \mu = \frac{3}{4}(1-\mu)$, using the definition of $u(\overline{y})$ and v_l, we have

$$\begin{aligned} m_2(1-\mu) \geq g(\overline{\nu})(\tilde{\mu} - \mu) &= |u(\overline{y}) - v_l| \\ &> m_1(\frac{\mu+3}{4} - \mu) = \frac{3m_1}{4}(1-\mu) \equiv \gamma_1 > 0. \end{aligned} \tag{15}$$

Next assume further that η satisfies

$$\omega(3\sqrt{2}\eta) < \frac{3m_1}{4}(1-\mu)\sin\alpha_0 \equiv r_1 = \gamma_1 \sin\alpha_0. \tag{16}$$

Since $\overline{y} \in \tilde{E}_l$ and the diameter of $\tilde{E}_l$ is $3\sqrt{2}\eta$, $u(\tilde{E}_l) \subset B_{r_1}(u(\overline{y}))$. By (15) and (16),

$$0 < r_1 = \gamma_1 \sin\alpha_0 < |u(\overline{y}) - v_l| \sin\alpha_0.$$

From this and Proposition 2, we have $\overline{B_{r_1}(u(\overline{y}))} \subset C^+_{v_l}$. Thus

$$u(\tilde{E}_l) \subset B_{r_1}(u(\overline{y})) \subset C^+_{v_l}. \tag{17}$$

Let $\zeta_l \in C^2_c((\tilde{E}_l)^o)$ such that $\zeta_l = 1$ on $\tilde{D}_l$ and $0 \leq \zeta_l \leq 1$. Define ζ_l to be zero in $\Omega \setminus (\tilde{E}_l)^o$. (For other values $1 \leq j \leq L$, we shall assume that the definition of ζ_j in $\tilde{E}_j$ differs only by a rigid translation that maps $\tilde{E}_j$ onto $\tilde{E}_l$). For $x \in \Omega$, and $h \neq 0$ sufficiently small, define

$$w_l(x) = \zeta_l^2(x)(v_l - u(x))|\nabla^h u(x)|^2 \tag{18}$$

for $1 \leq l \leq L$ where $\nabla^h u = (\nabla^h_1 u, \nabla^h_2 u)$ with $\nabla^h_k u(x) \equiv h^{-1}[u(x+he_k) - u(x)]$ for $k = 1, 2$. By (14) and Proposition 1, we can choose $\overline{t}(h) > 0$ depending only on h, $J(u)$, and $dist(\Omega', \partial\Omega)$ so that

$$\tau \equiv t\zeta^2(x)|\nabla^h u(x)|^2 \leq 1 \quad \text{for } x \in \Omega \quad \text{and } 0 \leq t \leq \overline{t}(h).$$

Note that if $w_l(x) \neq 0$, then $x \in \tilde{E}_l \subset \Omega''$ and by (17), $u(x) \in \overline{\mathcal{K}} \cap B_{r_1}(u(\overline{y})) \subset \overline{\mathcal{K}} \cap C^+_{v_l}$. Hence

$$\begin{aligned} u(x) + tw_l(x) &= u(x) + \tau(v_l - u(x)) \\ &= (1-\tau)u(x) + \tau v_l \in \overline{\mathcal{K}} \cap C_{v_l}^+ \end{aligned} \tag{19}$$

by convexity, and it is located on the ray from $u(x)$ to v_l in $\overline{\mathcal{K}} \cap C_{v_l}^+$. By Corollary 1, we have

$$f_0(u(x) + tw_l(x)) \le f_0(u(x)) \text{ for all } x \in \Omega \text{ and } 0 \le t \le \bar{t}(h).$$

Thus

$$\int\limits_\Omega f_0(u + tw_l)dx \le \int\limits_\Omega f_0(u)dx \quad \text{for } 0 \le t \le \bar{t}(h)$$

and since u is a minimizer and $f(u) = f_0(u) - \frac{M}{2}|u|^2$,

$$0 \le J(u+tw_l) - J(u) \le \int\limits_\Omega ((F(u+tw_l, D(u+tw_l)) - F(u, Du) + \frac{M}{2}(|u|^2 - |u+tw_l|^2))dx.$$

Dividing by t and letting $t \to 0$, we obtain

$$\begin{aligned} 0 &\le \int\limits_\Omega (F_P(u, Du) : Dw_l + [F_u(u, Du) - Mu] \cdot w_l)dx \\ &= \int\limits_{\tilde{E}_l} (F_P(u, Du) : Dw_l + [F_u(u, Du) - Mu] \cdot w_l)dx. \end{aligned} \tag{20}$$

We now proceed as in the proof of [7], Ch. II, Thm 1.2. Let C_i denote constants that are independent of h. Consider $\tilde{D}_l \in \mathcal{D}_2$ and $\zeta \equiv \zeta_l$ as above. The first step is to consider the test function $\phi \equiv \phi_l = \phi_{l,k}$ defined by

$$\begin{aligned} \phi_{l,k}(x) &= [\zeta^2(x)u(x + he_k) + \zeta^2(x - he_k)u(x - he_k) - (\zeta^2(x) + \zeta^2(x - he_k))u(x)]h^{-2} \\ &= \nabla_k^{-h}(\zeta^2 \nabla_k^h u) \quad \text{for } 1 \le \mathrm{k} \le 2, \end{aligned}$$

where $\nabla_k^h u(x) = h^{-1}[u(x + he_k) - u(x)]$ for $h \neq 0$ sufficiently small so that ϕ is compactly supported in $(\tilde{E}_l)^\circ$ for $k = 1, 2$. Recall that $\tilde{E}_l \subset \Omega''$ and $u(\tilde{E}_l) \subset B_{r_1}(u(\overline{y})) \subset \overline{\mathcal{K}}$. Thus by (17), $u(\tilde{E}_l) \subset \mathcal{K} \cap C_{v_l}^+$. Since $f_0(v)$ is convex on $\mathcal{K}$, it was proved in [5] that for h sufficiently small and $0 < t < t(h)$ sufficiently small, we have

$$\int_{\tilde{E}_l} (f_0(u + t\phi_{l,k}) - f_0(u))dx \le 0.$$

As in (20), since u is a minimizer of J we have

$$0 \leq \int\limits_{\Omega} (F_P(u, Du) : D\phi_{l,k} + [F_u(u, Du) - Mu] \cdot \phi_{l,k})dx \tag{21}$$
$$= \int\limits_{\tilde{E}_l} (F_P(u, Du) : D\phi_{l,k} + [F_u(u, Du) - Mu] \cdot \phi_{l,k})dx.$$

Here

$$\int\limits_{\tilde{E}_l} (F_P(u, Du) : D\phi_{l,k})dx = \int\limits_{\tilde{E}_l} (F_{p^i_\alpha}(u, Du) \cdot \frac{\partial}{\partial x_\alpha}(\nabla_k^{-h}[\zeta^2 \nabla_k^h u^i])dx$$
$$= \int\limits_{\tilde{E}_l} \left[A^{\alpha\beta}_{ij}(u)\frac{\partial u^j}{\partial x_\beta} + B^\alpha_i(u)\right] \cdot \frac{\partial}{\partial x_\alpha}(\nabla_k^{-h}[\zeta^2 \nabla_k^h u^i])dx$$
$$= -\int\limits_{\tilde{E}_l} \left\{\nabla_k^h \left[A^{\alpha\beta}_{ij}(u)\frac{\partial u^j}{\partial x_\beta} + B^\alpha_i(u)\right]\right\} \cdot \frac{\partial}{\partial x_\alpha}(\zeta^2 \nabla_k^h u^i)dx$$
$$-\int\limits_{\tilde{E}_l} \left\{\nabla_k^h \left[A^{\alpha\beta}_{ij}(u) \cdot \frac{\partial u^j}{\partial x_\beta} + B^\alpha_i(u)\right]\right\} \cdot \{\zeta^2 \nabla_k^h (\frac{\partial u_i}{\partial x_\alpha}) + 2\zeta \left(\frac{\partial \zeta}{\partial x_\alpha}\right)(\nabla_k^h u_i)\}dx$$

and

$$\int\limits_{\tilde{E}_l} ([F_u(u, Du) - Mu] \cdot \phi_{l,k})dx$$
$$= \int\limits_{\tilde{E}_l} \left[D_u(A^{\alpha\beta}_{ij}(u)) \cdot \frac{\partial u^i}{\partial x_\alpha} \cdot \frac{\partial u^j}{\partial x_\beta} + D_u(B^\alpha_i(u)) \cdot \frac{\partial u^i}{\partial x_\alpha} - Mu\right] \cdot [\nabla_k^{-h}(\zeta^2 \nabla_k^h u)]dx$$
$$= -\int\limits_{\tilde{E}_l} \nabla_k^h \left\{\left[D_u(A^{\alpha\beta}_{ij}(u)) \cdot \frac{\partial u^i}{\partial x_\alpha} \cdot \frac{\partial u^j}{\partial x_\beta} + D_u(B^\alpha_i(u)) \cdot \frac{\partial u^i}{\partial x_\alpha} - Mu\right]\right\} \cdot (\zeta^2 \nabla_k^h u))dx$$

for $1 \leq k \leq 2$. It follows that

$$\frac{\lambda}{2} \int\limits_{\tilde{E}_l} |\nabla^h Du|^2 \zeta^2 \, dx \leq C_0 \int\limits_{\tilde{E}_l} |\nabla^h u|^2 |Du|^2 \zeta^2 \, dx + C_1 \tag{22}$$

for all η sufficiently small, where λ is the constant defined in our assumption (3). Next we use (20) and our definition of w_l to prove a second inequality that will provide an upper bound on the second derivatives of u in $L^2(\tilde{E}_l)$.

By (20) and our definition of w_l, we have

$$0 \leq \int_{\tilde{E}_l} (F_P(u, Du) : Dw_l + [F_u(u, Du) - Mu] \cdot w_l) dx \tag{23}$$

$$= \int_{\tilde{E}_l} \left[A_{ij}^{\alpha\beta}(u) \cdot \frac{\partial u^j}{\partial x_\beta} + B_i^\alpha(u) \right] \cdot \frac{\partial}{\partial x_\alpha} (\zeta^2 (v_l^i - u^i(x)) |\nabla^h u|^2) dx$$

$$+ \int_{\tilde{E}_l} \left[D_u(A_{ij}^{\alpha\beta}(u)) \cdot \frac{\partial u^i}{\partial x_\alpha} \cdot \frac{\partial u^j}{\partial x_\beta} + (D_u(B_i^\alpha(u))) \cdot \frac{\partial u^i}{\partial x_\alpha} - Mu \right] \cdot \zeta^2 (v_l - u(x)) |\nabla^h u|^2 dx.$$

Recall that $u(\tilde{E}_l) \subset B_{r_1}(u(\overline{y}))$ and by (15) and (16),

$$|v_l - u(x)| \leq |v_l - u(\overline{y})| + |u(\overline{y}) - u(x)| \leq |v_l - u(\overline{y})| + r_1$$
$$\leq m_2(1-\mu) + \frac{3m_1}{4} \sin \alpha_0 (1-\mu) \equiv C_2(1-\mu),$$

for all x in $\tilde{E}_l$. Using this we see from (23) that

$$\frac{\lambda}{2} \int_{\tilde{E}_l} |\nabla^h u|^2 |Du|^2 \zeta^2 \, dx \leq C_3(1-\mu) \int_{\tilde{E}_l} |\nabla^h Du|^2 \zeta^2 \, dx + C_4. \tag{24}$$

Note that C_0 and C_3 depend only on λ and M_2. Taking $(1-\mu)$ sufficiently small so that $C_3(1-\mu) \leq \frac{\lambda^2}{8C_0}$, it follows from (22) and (24) that

$$\int_{\tilde{D}_l} (|\nabla^h Du|^2 + |\nabla^h u|^2 |Du|^2) dx \leq C_5. \tag{25}$$

The constants C_j are uniform in h for $|h| \leq h_0$. Letting $h \to 0$ we get

$$\int_{\tilde{D}_l} (|D^2 u|^2 + |Du|^4) dx \leq C_5. \tag{26}$$

This inequality holds for all $\tilde{D}_l \in \mathcal{D}_2$. If $\tilde{D}_l \in \mathcal{D}_3$, then f_u is bounded on a neighborhood of the union of all such squares. It follows as in the proof from [7], Ch. II referred to above that we have (26) (with a possibly larger value of C_5) in this case as well.

In conclusion, we first fix $\mu \in (\mu_1, 1)$ sufficiently close to 1 so that (25) follows from (22) and (24). We then choose $\eta \in (0, \eta_0)$ sufficiently small so that (16), (22), and (24) hold. This fixes the covering $\mathcal{D}_1$ such that (26) holds for a fixed constant C_5 for all $1 \leq l \leq L$. Summing on l we have $\int_{\Omega'} |D^2 u| dx < \infty$.

We now comment on the case $q = 1$. In this instance $\mathcal{K}$ is a bounded interval (a, b) such that $f_0(v)$ is a convex function satisfying $\lim\limits_{v \uparrow b} f_0(v) = \infty = \lim\limits_{v \downarrow a} f_0(v)$. It follows that there exists $\delta > 0$ so that $f_0'(v) > 0$ for $b - \delta < v < b$. In the same way we have that $f_0'(v) < 0$ for $a < v < a + \delta$. With this information we can carry out the argument as above using half lines $R_v^-(R_v^+)$ in place of the cones $C_v^-(C_v^+)$. □

3 Proof of Theorem 1

Recall that $u \in C^2(\Omega_0)$ and $\Omega_0 \subset \Omega \setminus \Lambda$. The two facts that $u \in H^2_{\text{loc}}(\Omega)$ and u satisfies Eq. (4) on Ω_0 imply that u is a strong solution to the equilibrium equations (4) throughout Ω and that each term appearing in this equation is in $L^2_{\text{loc}}(\Omega)$. We use this to prove Theorem 1.

Proof Suppose that $B_{2R}(x_0) \subset\subset \Omega$. Let (ρ, θ) be polar coordinates centered at x_0 and consider the field of pure second derivatives $D^2_{\upsilon\upsilon} u(x)$ where $\upsilon = \upsilon(x) = e_\theta$ for $x \in B_R(x_0)$. We have $D^2_{\upsilon\upsilon} u(x) = \frac{u_\rho}{\rho} + \frac{u_{\theta\theta}}{\rho^2}$ is in $L^2(B_R(x_0))$. Let $\zeta = \zeta(\rho) \in C^2_c(B_R(x_0))$ such that $\zeta = 1$ on $B_{R/2}(x_0)$. Fix $0 < r \leq \frac{R}{2}$. Multiplying the Eq. (4) by $\zeta^2 D^2_{\upsilon\upsilon} u$ and integrating over $B_R(x_0) \setminus B_r(x_0)$, we obtain

$$\int\limits_{B_R(x_0)\setminus B_r(x_0)} (\text{div} F_P - F_u) \cdot \zeta^2 D^2_{\upsilon\upsilon} u \, dx = \int\limits_{B_R(x_0)\setminus B_r(x_0)} f_u(u) \cdot \zeta^2 \left(\frac{u_\rho}{\rho} + \frac{u_{\theta\theta}}{\rho^2} \right) dx. \tag{27}$$

Since $\text{div} F_P - F_u$ and the test function $D^2_{\upsilon\upsilon} u$ are in $L^2(B_R(x_0))$ we get

$$\left| \int\limits_{B_R(x_0)\setminus B_r(x_0)} (\text{div} F_P - F_u) \cdot \zeta^2 D^2_{\upsilon\upsilon} u \, dx \right| \leq C_1, \tag{28}$$

where C_1 is independent of r for r in $(0, \frac{R}{2}]$. By (4), f_u is also in $L^2(B_R(x_0))$. Consider

$$\int\limits_{B_R(x_0)\setminus B_r(x_0)} f_u(u) \cdot \zeta^2 \frac{u_{\theta\theta}}{\rho^2} \, dx = \int_r^R \int_0^{2\pi} f_u(u) \cdot \zeta^2 \frac{u_{\theta\theta}}{\rho} \, d\theta d\rho.$$

Since u is a finite energy minimizer of J in Ω, $f(u)$ is in $L^1(\Omega)$ and for almost every ρ satisfying $r \leq \rho \leq R$ we have

$$\int_0^{2\pi} (|f(u(\rho,\theta))| + |f_u(u(\rho,\theta))|^2 + |u_\theta(\rho,\theta)|^2)d\theta < \infty. \tag{29}$$

For such a value of ρ, if there is an interval $(\alpha,\beta) \subset [0,2\pi)$ so that $u(\rho,\beta) \in \Lambda$ and $u(\rho,\theta) \notin \Lambda$ for all θ in $[\alpha,\beta)$, then

$$\int_\alpha^\theta \partial_\phi f(u(\rho,\phi))d\phi = \int_\alpha^\theta f_u(u(\rho,\phi) \cdot u_\phi d\phi \leq C_1 < \infty.$$

Thus $f(u(\rho,\theta)) - f(u(\rho,\alpha)) \leq C_1 < \infty$. However, $\lim\limits_{\theta\to\beta} f(u(\rho,\theta)) = \infty$ and this is not possible. It follows that for each ρ for which (29) holds, we have $\partial B_\rho(x_0) \cap \Lambda = \emptyset$. Since $u \in C^2(\Omega \setminus \Lambda)$ it follows that the functions $u(\rho,\cdot)$ and $f(u(\rho,\cdot))$ are smooth. This allows us to integrate by parts and obtain

$$\begin{aligned}\frac{\zeta^2(\rho)}{\rho}\int_0^{2\pi} f_u \cdot u_{\theta\theta} d\theta &= -\frac{\zeta^2(\rho)}{\rho}\int_0^{2\pi} u_\theta \cdot D^2 f \cdot u_\theta \, d\theta \\ &\leq M \int\limits_{\partial B_\rho(x_0)} |Du|^2 ds.\end{aligned}$$

We conclude then that

$$\int\limits_{B_R(x_0)\setminus B_r(x_0)} f_u \cdot \frac{u_{\theta\theta}}{\rho^2}\zeta^2 dx \leq M \int_{B_R(x_0)} |Du|^2 dx \leq C_2. \tag{30}$$

Next using the estimate

$$\int_{B_R(x_0)} (|f_u(u(x))|^2 + |Du(x)|^2)dx < \infty$$

and the same argument as above on almost every line parallel to one of the coordinate axes it follows that $f(u(x)) \in W^{1,1}(B_R(x_0))$ and $Df(u) = f_u \cdot Du$ almost everywhere on $B_R(x_0)$. With this fact we see that

$$\begin{aligned}&\int_{B_R(x_0)\setminus B_r(x_0)} \zeta^2 f_u(u(x)) \cdot \frac{u_\rho}{\rho} dx \\ &= \int_0^{2\pi}\int_r^R \zeta^2(\rho)(f(u(\rho,\theta)))_\rho d\rho d\theta \\ &= -r^{-1}\int_{\partial B_r(x_0)} f(u)ds - \int_{B_R(x_0)\setminus B_{R/2}(x_0)} f(u)\frac{(\zeta^2)_\rho}{\rho}dx,\end{aligned}$$

where for the last term we used the fact that $\zeta^2 = 1$ on $B_{R/2}(x_0)$.

Thus

$$\int_{B_R(x_0)\setminus B_r(x_0)} \zeta^2 f_u \cdot \frac{u_\rho}{\rho} dx \leq -r^{-1} \int_{\partial B_r(x_0)} f ds + C_3. \tag{31}$$

Taking (28), (30), and (31) together with (27) we see that

$$r^{-1} \int_{\partial B_r(x_0)} f(u) ds \leq C_4 \quad \text{for } 0 < r < \frac{R}{2},$$

where C_4 is independent of r. Since $u(x)$ is continuous on Ω and $\lim_{u \to u_0} f(u) = f(u_0) \in (-\infty, \infty]$ for each $u_0 \in \overline{\mathcal{K}}$, we have

$$2\pi f(u(x_0)) \leq C_4,$$

where C_4 depends on R, $J(u)$, and $\|u\|_{H^2(B_R(x_0))}$. In particular $\Lambda = \emptyset$. Furthermore, applying Lemma 2 we see that $dist(u(x_0), \partial\mathcal{K}) \geq c > 0$ where c depends on $dist(x_0, \partial\Omega)$ and $J(u)$. □

4 Applications to Liquid Crystals

We briefly describe the liquid crystal model that motivates our constrained problem and state our result as it applies to this case. A more detailed overview of energies for liquid crystals is given in [2]. Let $\Omega \subset \mathbb{R}^3$ be a region filled with rod–like liquid crystal molecules. For $x \in \Omega$ and $p \in \mathbb{S}^2$ denote by $\rho(x, p)$ the probability distribution for the long axes of the molecules near x aligned with the direction p. We have

$$\rho(x, p) \geq 0, \qquad \int\limits_{\mathbb{S}^2} \rho(x, p) dp = 1. \tag{32}$$

If the directions are random so that no direction is preferred, then $\rho(x, p) = \rho_0 = \frac{1}{4}\pi$ and the liquid crystal is in the isotropic state at x. The de Gennes Q tensor is introduced as a macroscopic order parameter

$$Q(x) = \int\limits_{\mathbb{S}^2} (p \otimes p - \frac{1}{3} I) \rho(x, p) dp \tag{33}$$

representing the second moments of ρ normalized so that $Q = 0$ if $\rho = \rho_0$. Set $S_0 = \{A \in \mathbb{M}^{3\times 3} : A = A^t, trA = 0\}$. Then from (32) and (33) we see that Q takes on values in the open, bounded, and convex set

$$\mathcal{M} = \left\{ A \in S_0 : -\frac{1}{3} < \lambda_{\min}(A) \le \lambda_{\max}(A) < \frac{2}{3} \right\}, \tag{34}$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the minimum and maximum eigenvalues of A, respectively. This is the set of physically attainable states in S_0. A free energy for constant nematic liquid crystal states identified with $Q \in \mathcal{M}$ was developed by Katriel et al. [9] and Ball and Majumdar [3]. Assuming Maier–Saupe molecular interactions it takes the form

$$\begin{cases} \psi_b(Q) = T f_{ms}(Q) - \kappa |Q|^2 & \text{for } Q \in \mathcal{M}, \\ \qquad = \infty & \text{for } Q \in S_0 \setminus \mathcal{M}, \\ f_b(Q) = \inf_{\rho \in A_Q} (\int\limits_{\mathbb{S}^2} \rho \log \rho \, dp) & \end{cases} \tag{35}$$

where

$$A_Q = \{\rho \in L^1(\mathbb{S}^2) : \rho \ge 0, \int\limits_{\mathbb{S}^2} \rho(p) dp = 1, \quad Q = \int\limits_{\mathbb{S}^2} (p \otimes p - \frac{1}{3} I) \, \rho(p) \, dp\}$$

and $T, \kappa > 0$. It is shown in [3] that f_{ms} is convex on $\mathcal{M}$ with $\lim\limits_{Q \to \partial \mathcal{M}} f(Q) = \infty$ and it is shown in [6] that $f \in C^\infty(\mathcal{M})$. Ball and Majumdar then used ψ_b to define an energy functional to characterize stable spatially varying liquid crystal configurations. They considered local minimizers $Q \in H^1(\Omega, \overline{\mathcal{M}})$ to

$$I_{LdG}[Q] = \int\limits_{\Omega} (G(Q, DQ) + \psi_b(Q)) dx, \tag{36}$$

where following [10] the elastic energy density takes the form

$$G(Q, DQ) = \sum_{i=1}^{5} L_i I_i(Q, DQ), \tag{37}$$

such that

$$\begin{aligned} I_1 &= D_{x_k} Q_{ij} D_{x_k} Q_{ij} & I_2 &= D_{x_j} Q_{ij} D_{x_k} Q_{ik}, \\ I_3 &= D_{x_j} Q_{ik} D_{x_k} Q_{ij} & I_4 &= Q_{\ell k} D_{x_\ell} Q_{ij} D_{x_k} Q_{ij}, \\ I_5 &= \epsilon_{\ell j k} Q_{\ell i} D_{x_j} Q_{ki}. \end{aligned}$$

These are polynomial expressions in terms of Q and DQ that satisfy the principle of frame indifference and material symmetry. The expressions I_1, I_2, I_3, and I_4 satisfy both properties. Analytically this means that $I_1, \cdots, I_4$ are invariant under transformations over $O(3)$ while I_5 satisfies only frame indifference and is invariant under transformations over $SO(3)$. The first four terms are quadratic in DQ while the fifth is linear and is included in the elastic energy density when modeling chiral liquid crystals. Here $\epsilon_{\ell jk}$ is the Levi–Civita tensor.

Let $\mathcal{D} := \{D = [D_{ijk}] \quad 1 \leq i, j, k \leq 3 : D_{ijk} = D_{jik}$ and $\sum_{i=1}^{3} D_{\ell\ell k} = 0$ for each i, j, and $k\}$. The elasticity constants are to be chosen so that $\sum_{\ell=1}^{4} L_k I_\ell(Q, D) \geq c_0|D|^2$ for some $c_0 > 0$ for all $Q \in \overline{\mathcal{M}}$ and $D \in \mathcal{D}$. Inequalities that ensure the coercivity condition are

$$L_1' + \frac{5}{3}L_2 + \frac{1}{6}L_3 > 0, \quad L_1' - \frac{1}{2}L_3 > 0, \quad L_1' + L_3 > 0, \tag{38}$$

where

$$L_1' = \begin{cases} L_1 - \frac{1}{3}L_4 & \text{if } L_4 \geq 0 \\ L_1 + \frac{2}{3}L_4 & \text{if } L_4 \leq 0. \end{cases}$$

(See [2]).

The space S_0 has dimension five. If we take an orthonormal basis $\{E_1, \ldots, E_5\}$ we can parameterize the space with the isometry

$$Q(v) : \mathbb{R}^5 \to S_0, \quad Q(v) = \sum_{j=1}^{5} v_j E_j.$$

Set $\mathcal{K} = \{v \in \mathbb{R}^5 : Q(v) \in \mathcal{M}\}$, $f(v) = \psi_b(Q(v))$ and $F(v, Dv) = G(Q(v), DQ(v))$. Then $\mathcal{K}$ is an open, bounded, and convex region in $\mathbb{R}^5$, f satisfies (2) and F satisfies (3). For the case $n = 2$ we view the liquid crystal body on the infinite cylinder $\Omega \times \mathbb{R}$ where $\Omega \subset \mathbb{R}^2$ is the cylinder's cross section and the order parameter is $Q = Q(x_1, x_2)$. We can then apply our results from Theorem 1 to $J[v] = I_{LdG}[Q(v)]$ to obtain:

Theorem 2 *Let $\Omega \subset \mathbb{R}^2$ and let $Q \in H^1(\Omega; \overline{\mathcal{M}})$ be a finite energy local minimizer for $I_{LdG}[\cdot]$ satisfying (35)-(38). Then $Q \in C^2(\Omega)$. If $E \subset\subset \Omega$, then $Q(E) \subset\subset \mathcal{M}$ and Q satisfies the equilibrium equation*

$$[div\, G_D(G, DQ) - G_Q(Q, DQ) - \psi_{b,Q}(Q)]^{st} = 0 \; in \; \Omega,$$

where $[A]^{st}$ is the symmetric and traceless part of $A \in \mathbb{M}^{3\times3}$.

References

1. J.M. Ball, Analysis of liquid crystals and their defects, in *Lecture Notes of the Scuola Estiva GNFM*, Ravello 17–22 Sep. 2018
2. J.M. Ball, Mathematics and liquid crystals. Mol. Cryst. Liquid Cryst. **647**(1), 1–27 (2017)
3. J.M. Ball, A. Majumdar, Nematic liquid crystals: from Maier–Saupe to continuum theory. Mol. Cryst. Liquid Cryst. **525**, 1–11 (2010)
4. P. Bauman, D. Phillips, Regularity and behavior of eigenvalues for minimizers of a constrained Q–tensor energy for liquid crystals. Cals. Var. Part. Differ. Equa. **55**(4), 22 pp. (2016). Art. 81
5. L.C. Evans, O. Kneuss, H. Tran, Partial regularity for minimizers of singular energy functionals, with applications to nematic liquid crystal models. Trans. Am. Math. Soc. **368**(5), 3389–3413 (2016)
6. E. Feireisl, E. Rocca, G. Schimperna, A. Zarnescu, Nonisothermal nematic liquid crystal flows with the Ball–Majumdar free energy. Ann. Mat. Puva. Anal. (4) **194**(50), 1269–1299 (2015)
7. M. Giaquinta, *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*. Annals of Mathematics Studies, vol. 105 (Princeton University Press, Princeton, NJ, 1983)
8. Z. Geng, J. Tong, Regularity of a tensor-valued variational obstacle problem in three dimensions. Calc. Var. Part. Differ. Equa. **59** (2020). Art. 57
9. J. Katriel, G.F. Kventsel, G. Luckhurst, T. Sluckin, Free energies in the Landau and molecular field approaches. Liquid Crystals **1**, 337–355 (1986)
10. L. Longa, D. Monselesan, H. Trebin, An extension of Landau de Gennes theory for liquid crystals. Liquid Cryst. **2**(6), 769–796 (1987)

On Some Models in Radiation Hydrodynamics

Xavier Blanc, Bernard Ducomet, and Šárka Nečasová

1 Introduction

Modeling the *radiation hydrodynamics* means to find a way to incorporate effects of radiation in the classical hydrodynamics framework. There exist numerous applications ranging from combustion and high-temperature hydrodynamics to models of gaseous stars in astrophysics. The mathematical models have to reflect the effect of coupling between the macroscopic description of the fluid and the statistical character of the motion of the massless photons see, e.g., the monograph by Chandrasekhar [10]. The relativistic version of model of radiation hydrodynamics has been introduced by Pomraning [45] and Mihalas and Weibel-Mihalas [40] and investigated more recently in astrophysics and laser applications (in the inviscid case) by Lowrie, Morel and Hittinger [36] and Buet and Després [9], with a special attention to asymptotic regimes.

In our paper we follow the studies and modeling by Buet and Després [9], Golse and Perthame [25]. The motion of the fluid is governed by the standard field equations of classical continuum fluid mechanics describing the evolution of *the mass density* $\varrho = \varrho(t, x)$, the *velocity field* $\mathbf{u} = \mathbf{u}(t, x)$, and the *absolute*

X. Blanc
Université Paris Cité and Sorbonne Université, CNRS, Paris, France
e-mail: xavier.blanc@u-paris.fr

B. Ducomet
Université Paris-Est, LAMA (UMR 8050), UPEMLV, UPEC, CNRS, Créteil, France
e-mail: bernard.ducomet@u-pec.fr

Š. Nečasová (✉)
Institute of Mathematics of the Academy of Sciences of the Czech Republic, Praha, Czech Republic
e-mail: matus@math.cas.cz

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_4

temperature $\vartheta = \vartheta(t, x)$ as functions of the time t and the Eulerian spatial coordinate $x \in \Omega \subset \mathbb{R}^3$. The effect of radiation, represented by its quanta—massless particles called *photons* traveling at the speed of light c—is incorporated in the *radiative intensity* $I = I(t, x, \boldsymbol{\omega}, \nu)$, depending on the direction vector $\boldsymbol{\omega} \in \mathcal{S}^2$, where $\mathcal{S}^2 \subset \mathbb{R}^3$ denotes the unit sphere, and the frequency $\nu \geq 0$. The collective effect of radiation is then expressed in terms of integral means with respect to the variables $\boldsymbol{\omega}$ and ν of quantities depending on I. The radiation energy E_R is given as

$$E_R(t, x) = \frac{1}{c} \int_{\mathcal{S}^2} \int_0^\infty I(t, x, \boldsymbol{\omega}, \nu) \, \mathrm{d}\boldsymbol{\omega} \, \mathrm{d}\nu. \tag{1.1}$$

The time evolution of I is described by a transport equation with a source term depending on the absolute temperature, while the effect of radiation on the macroscopic motion of the fluid is represented by extra source terms in the momentum and energy equations evaluated in terms of I.

2 Compressible Viscous Radiation Fluid

Let us first consider the viscous case. The system of equations in $(0, T) \times \Omega$ describing the motion of viscous fluid consists of the continuity equation, momentum equations, energy balance:

$$\left.\begin{array}{l}
\partial_t \varrho + \mathrm{div}_x(\varrho \mathbf{u}) = 0; \\
\partial_t(\varrho \mathbf{u}) + \mathrm{div}_x(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla_x p(\varrho, \vartheta) = \mathrm{div}_x \mathbb{T} - \mathbf{S}_F; \\
\partial_t \left(\varrho \left(\frac{1}{2} |\mathbf{u}|^2 + e(\varrho, \vartheta) \right) \right) + \mathrm{div}_x \left(\varrho \left(\frac{1}{2} |\mathbf{u}|^2 + e(\varrho, \vartheta) \right) \mathbf{u} \right) + \mathrm{div}_x \Big(p\mathbf{u} + \mathbf{q} - \mathbb{T}\mathbf{u} \Big) \\
= -S_E.
\end{array}\right\} \tag{2.1}$$

The radiation is modeled by the radiation transport equation

$$\frac{1}{c} \partial_t I + \boldsymbol{\omega} \cdot \nabla_x I = S \text{ in } (0, T) \times \Omega \times (0, \infty) \times \mathcal{S}^2. \tag{2.2}$$

The symbol $p = p(\varrho, \vartheta)$ denotes the thermodynamic pressure and $e = e(\varrho, \vartheta)$ is the specific internal energy, interrelated through *Maxwell's equation*

$$\frac{\partial e}{\partial \varrho} = \frac{1}{\varrho^2} \left(p(\varrho, \vartheta) - \vartheta \frac{\partial p}{\partial \vartheta} \right). \tag{2.3}$$

Furthermore, $\mathbb{T}$ is the viscous stress tensor determined by *Newton's rheological law*

$$\mathbb{T} = \mu \left(\nabla_x \mathbf{u} + \nabla_x^t \mathbf{u} - \frac{2}{3} \mathrm{div}_x \mathbf{u} \right) + \eta \, \mathrm{div}_x \mathbf{u} \, \mathbb{I}, \tag{2.4}$$

where the shear viscosity coefficient $\mu = \mu(\vartheta) > 0$ and the bulk viscosity coefficient $\eta = \eta(\vartheta) \geq 0$ are effective functions of the absolute temperature. Similarly, $\mathbf{q}$ is the heat flux given by *Fourier's law*

$$\mathbf{q} = -\kappa \nabla_x \vartheta, \tag{2.5}$$

with the heat conductivity coefficient $\kappa = \kappa(\vartheta) > 0$.

Finally,

$$S = S_{a,e} + S_s, \tag{2.6}$$

where

$$S_{a,e} = \sigma_a \Big(B(\nu, \vartheta) - I \Big), \ S_s = \sigma_s \left(\frac{1}{4\pi} \int_{\mathcal{S}^2} I(\cdot, \boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega} - I \right), \tag{2.7}$$

with $B(\nu, \vartheta) = 2h\nu^3 c^{-2} \left(e^{\frac{h\nu}{k\vartheta}} - 1 \right)^{-1}$ which define the radiative equilibrium function, h and k are the Planck and Boltzmann constants and with the absorption coefficient $\sigma_a = \sigma_a(\nu, \vartheta) \geq 0$, and the scattering coefficient $\sigma_s = \sigma_s(\nu, \vartheta) \geq 0$. Moreover,

$$S_E = \int_{\mathcal{S}^2} \int_0^\infty S(\cdot, \nu, \boldsymbol{\omega}) \, \mathrm{d}\nu \, \mathrm{d}\boldsymbol{\omega}, \ \mathbf{S}_F = \frac{1}{c} \int_{\mathcal{S}^2} \int_0^\infty \boldsymbol{\omega} S(\cdot, \nu, \boldsymbol{\omega}) \, \mathrm{d}\nu \, \mathrm{d}\boldsymbol{\omega}. \tag{2.8}$$

More restrictions on the structural properties of constitutive relations will be imposed in subsection below.

System (2.1)–(2.2) is supplemented with the boundary conditions:

No-slip, no-flux:

$$\mathbf{u}|_{\partial\Omega} = 0, \ \mathbf{q} \cdot \mathbf{n}|_{\partial\Omega} = 0; \tag{2.9}$$

Transparency:

$$I(t, x, \nu, \boldsymbol{\omega}) = 0 \text{ for } x \in \partial\Omega, \ \boldsymbol{\omega} \cdot \mathbf{n} \leq 0, \tag{2.10}$$

where $\mathbf{n}$ denotes the outer normal vector to $\partial\Omega$.

Remark 1 Let us mention that recently the existence of weak solutions of the full system with Dirichlet boundary conditions for the temperature were solved, see [11]. Moreover, let us point out that instead of transparency condition for radiative intensity it can be considered the so-called *specular reflection*, see [15]. Further, the more general boundary conditions (nonhomogeneous Dirichlet boundary conditions for the velocity field and temperature, with given density on the input) were investigated, see [22].

System (2.1)–(2.10) can be viewed as a simplification of a model arising in radiation hydrodynamics. The physical descriptions can be found, see [40, 45] in the framework of special relativity, see also [41, 42] for a list of references and a review of related computational works in the relativistic framework. See also [9, 12, 34–36].

The *existence* of local-in-time solutions and sufficient conditions for blow up of classical solutions in the non-relativistic inviscid case were obtained by Zhong and Jiang [56], see also papers by Jiang and Wang [29, 30] for a related one-dimensional "Euler–Boltzmann" type models. Moreover, a simplified version of the system has been investigated by Golse and Perthame [25], where global existence was proved by means of the theory of nonlinear semi-groups.

2.1 Hypotheses and Main Results

The hypotheses imposed on constitutive relations are motivated by the general *existence theory* for the Navier–Stokes–Fourier system developed in [20, Chapter 3]. See also [43, 44].

2.2 Constitutive Equations

We consider the pressure in the form

$$p(\varrho,\vartheta) = \vartheta^{5/2} P\left(\frac{\varrho}{\vartheta^{3/2}}\right) + \frac{a}{3}\vartheta^4,\ a > 0, \tag{2.11}$$

where $P : [0,\infty) \to [0,\infty)$ is a given function with the following properties:

$$P \in C^1[0,\infty),\ P(0) = 0,\ P'(Z) > 0 \text{ for all } Z \geq 0, \tag{2.12}$$

$$0 < \frac{\frac{5}{3}P(Z) - P'(Z)Z}{Z} < c \text{ for all } Z \geq 0, \tag{2.13}$$

$$\lim_{Z\to\infty} \frac{P(Z)}{Z^{5/3}} = p_\infty > 0. \tag{2.14}$$

For the physical background of hypotheses (2.11)–(2.14) we refer to [13, 20].

In accordance with Maxwell's equation (2.3), the specific internal energy e can be taken in the form

$$e(\varrho,\vartheta) = \frac{3}{2}\vartheta\left(\frac{\vartheta^{3/2}}{\varrho}\right) P\left(\frac{\varrho}{\vartheta^{3/2}}\right) + a\frac{\vartheta^4}{\varrho}, \tag{2.15}$$

whereas the associated specific entropy reads

$$s(\varrho, \vartheta) = M\left(\frac{\varrho}{\vartheta^{3/2}}\right) + \frac{4a}{3}\frac{\vartheta^3}{\varrho}, \tag{2.16}$$

with

$$M'(Z) = -\frac{3}{2}\frac{\frac{5}{3}P(Z) - P'(Z)Z}{Z^2} < 0.$$

The transport coefficients μ, η, and κ are continuously differentiable functions of the absolute temperature such that

$$0 < c_1(1+\vartheta) \leq \mu(\vartheta),\ \mu'(\vartheta) < c_2,\ 0 \leq \eta(\vartheta) \leq c(1+\vartheta), \tag{2.17}$$

$$0 < c_1(1+\vartheta^3) \leq \kappa(\vartheta) \leq c_2(1+\vartheta^3) \tag{2.18}$$

for any $\vartheta \geq 0$.

Finally, we assume that σ_a, σ_s, B are continuous functions of ν, ϑ such that

$$0 \leq \sigma_a(\nu,\vartheta), \sigma_s(\nu,\vartheta) \leq c_1,\ 0 \leq \sigma_a(\nu,\vartheta)B(\nu,\vartheta) \leq c_2, \tag{2.19}$$

$$\sigma_a(\nu,\vartheta), \sigma_s(\nu,\vartheta), \sigma_a(\nu,\vartheta)B(\nu,\vartheta) \leq h(\nu),\ h \in L^1(0,\infty), \tag{2.20}$$

and

$$\sigma_a(\nu,\vartheta), \sigma_s(\nu,\vartheta) \leq c\vartheta \tag{2.21}$$

for all $\nu \geq 0$, $\vartheta \geq 0$. Relations (2.19–2.21) represent a rather crude "cutoff" hypotheses neglecting the effect of radiation at large frequencies ν and low values of the temperature ϑ. Note, however, that relations similar to (2.21) were derived by Ripoll et al. [47].

Remark 2 A prototype of the above pressure law reads

$$p(\rho, \vartheta) = c_1\rho^{\frac{5}{3}} + c_2\rho\vartheta + \frac{a}{3}\vartheta^4,$$

with the corresponding internal energy and specific entropy of the form

$$e(\rho, \vartheta) = \frac{3}{2}c_1\rho^{\frac{2}{3}} + c_v\vartheta + \frac{a}{\rho}\vartheta^4,$$

$$s(\rho, \vartheta) = c_v \ln\vartheta - c_2\ln\rho + \frac{4a}{3\rho}\vartheta^3,$$

where $a, c_1, c_2, c_v > 0$.

2.3 Weak Formulation

In the weak formulation of the Navier–Stokes–Fourier system, it is customary to replace the equation of continuity (2.1) by its (weak) *renormalized* version represented by a family of integral identities

$$\int_0^T \int_\Omega \left(\Big(\varrho + b(\varrho)\Big)\partial_t\varphi + \Big(\varrho + b(\varrho)\Big)\mathbf{u}\cdot\nabla_x\varphi + \Big(b(\varrho) - b'(\varrho)\varrho\Big)\mathrm{div}_x\mathbf{u}\varphi\right)\,\mathrm{d}x\,\mathrm{d}t \\ = -\int_\Omega \Big(\varrho_0 + b(\varrho_0)\Big)\varphi(0,\cdot)\,\mathrm{d}x \tag{2.22}$$

satisfied for any $\varphi \in C_c^\infty([0,\infty)\times\overline{\Omega})$, and any $b \in C^\infty[0,\infty)$, $b' \in C_c^\infty[0,\infty)$.[1]

Similarly, the momentum equation (2.1_2) is replaced by

$$\int_0^T \int_\Omega (\varrho\mathbf{u}\cdot\partial_t\varphi + \varrho\mathbf{u}\otimes\mathbf{u} : \nabla_x\varphi + p\mathrm{div}_x\varphi)\,\mathrm{d}x\,\mathrm{d}t \tag{2.23}$$

$$= \int_0^T \int_\Omega \mathbb{T} : \nabla_x\varphi + \mathbf{S}_F\cdot\varphi\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega (\varrho\mathbf{u})_0\cdot\varphi(0,\cdot)\,\mathrm{d}x$$

for any $\varphi \in C_c^\infty([0,T)\times\Omega)$.[2]

As a matter of fact, the total energy balance (2.1_3) is not suitable for the weak formulation since, at least according to the recent state of the art, the term $\mathbb{S}\mathbf{u}$ is not controlled on the (hypothetical) vacuum zones of vanishing density.

Finally, similarly to [13], we consider in the weak formulation by an entropy *inequality*, specifically,

$$\int_0^T \int_\Omega (\varrho s\partial_t\varphi + \varrho\mathbf{u}\cdot\nabla_x\varphi + \mathbf{q}\vartheta\cdot\nabla_x\varphi)\,\mathrm{d}x\,\mathrm{d}t \leq -\int_\Omega (\varrho s)_0\varphi(0,\cdot)\,\mathrm{d}x \tag{2.24}$$

$$-\int_0^T \int_\Omega \frac{1}{\vartheta}\left(\mathbb{S} : \nabla_x\mathbf{u} - \frac{\mathbf{q}\cdot\nabla_x\vartheta}{\vartheta}\right)\varphi\,\mathrm{d}x\,\mathrm{d}t - \int_0^T \int_\Omega \frac{1}{\vartheta}\Big(\mathbf{u}\cdot\mathbf{S}_F - S_E\Big)\varphi\,\mathrm{d}x\,\mathrm{d}t$$

for any $\varphi \in C_c^\infty([0,T)\times\overline{\Omega})$, $\varphi \geq 0$.

[1] Note that (2.22) implicitly includes the initial condition

$$\varrho(0,\cdot) = \varrho_0.$$

[2] As the viscous stress contains first derivatives of the velocity $\mathbf{u}$, for (2.23) to make sense, the field $\mathbf{u}$ must belong to a certain Sobolev space with respect to the spatial variable. Here, we require that $\mathbf{u} \in L^2(0,T;W_0^{1,2}(\Omega))$.

The system (2.22), (2.23), (2.24) must be supplemented with the total energy balance

$$\begin{aligned}&\textstyle\int_{\Omega}\left(\frac{1}{2}\varrho|\mathbf{u}|^2+\varrho e(\varrho,\vartheta)+E_R\right)(\tau,\cdot)\,\mathrm{d}x\\&\qquad+\textstyle\int_0^\tau\int\int_{\partial\Omega\times\mathcal{S}^2,\ \boldsymbol{\omega}\cdot\mathbf{n}\geq 0}\int_0^\infty \boldsymbol{\omega}\cdot\mathbf{n}I(t,x,\boldsymbol{\omega},\nu)\,\mathrm{d}\nu\,\mathrm{d}\boldsymbol{\omega}\,\mathrm{d}S_x\,\mathrm{d}t\\&=\int_{\Omega}\left(\frac{1}{2\varrho_0}|(\varrho\mathbf{u})_0|^2+(\varrho e)_0+E_{R,0}\right)\,\mathrm{d}x,\end{aligned}\tag{2.25}$$

where E_R is given by (1.1), and

$$E_{R,0}=\int_{\mathcal{S}^2}\int_0^\infty I_0(\cdot,\boldsymbol{\omega},\nu)\,\mathrm{d}\boldsymbol{\omega}\ \mathrm{d}\nu.$$

The transport equation (2.2) can be extended to the whole physical space $\mathbb{R}^3$ provided we set

$$\sigma_a(x,\nu,\vartheta)=1_{\Omega}\sigma_a(\nu,\vartheta),\ \sigma_s(x,\nu,\vartheta)=1_{\Omega}\sigma_s(\nu,\vartheta)$$

and take the initial distribution $I_0(x,\boldsymbol{\omega},\nu)$ to be zero for $x\in\mathbb{R}^3\setminus\Omega$. Accordingly, for any fixed $\boldsymbol{\omega}\in\mathcal{S}^2$, Eq. (2.2) can be viewed as a linear transport equation defined in $(0,T)\times\mathbb{R}^3$, with a right-hand side S. With the above mentioned convention, extending $\mathbf{u}$ to be zero outside Ω, we may, therefore, assume that both ϱ and I are defined on the whole physical space $\mathbb{R}^3$.

Remark 3 It is possible to introduce "weaker" solutions in the sense that we consider energy inequality rather than energy equality. Although it seems that we are losing a lot of information our definition of weak solution is still sufficient. Namely, if the above defined weak solution is smooth enough it will be a strong one. For a justification of this fact we refer to [46, Section 1.2].

Definition 1 We say that ϱ, $\mathbf{u}$, ϑ, I is a weak solution of problem (2.1)–(2.10) if

$$\varrho\geq 0,\ \vartheta>0 \text{ for a.a. } (t,x)\times\Omega,\ I\geq 0 \text{ a.a. in } (0,T)\times\Omega\times\mathcal{S}^2\times(0,\infty),$$

$$\varrho\in L^\infty(0,T;L^{5/3}(\Omega)),\ \vartheta\in L^\infty(0,T;L^4(\Omega)),$$

$$\mathbf{u}\in L^2(0,T;W_0^{1,2}(\Omega)),\ \vartheta\in L^2(0,T;W^{1,2}(\Omega)),$$

$$I\in L^\infty((0,T)\times\Omega\times\mathcal{S}^2\times(0,\infty)),\ I(t,\cdot)\in L^\infty(0,T;L^1(\Omega\times\mathcal{S}^2\times(0,\infty)),$$

and if ϱ, $\mathbf{u}$, ϑ, I satisfy the integral identities (2.22), (2.23), (2.24), (2.25), together with the transport equation (2.2).

2.4 Existence Result

We state the existence result of radiation hydrodynamics.

Theorem 1 *Let $\Omega \subset \mathbb{R}^3$ be a bounded Lipschitz domain. Assume that the thermodynamic functions p, e, s satisfy hypotheses (2.11)–(2.16), and that the transport coefficients μ, λ, κ, σ_a, and σ_s comply with (2.17)–(2.21).*

Let $\{\varrho_\varepsilon, \mathbf{u}_\varepsilon, \vartheta_\varepsilon, I_\varepsilon\}_{\varepsilon>0}$ be a family of weak solutions to problem (2.1)–(2.10) in the sense of Definition 2.1 such that

$$\varrho_\varepsilon(0,\cdot) \equiv \varrho_{\varepsilon,0} \to \varrho_0 \text{ in } L^{5/3}(\Omega), \tag{2.26}$$

$$\int_\Omega \left(\frac{1}{2}\varrho_\varepsilon |\mathbf{u}_\varepsilon|^2 + \varrho_\varepsilon e(\varrho_\varepsilon, \vartheta_\varepsilon) + E_{R,\varepsilon} \right)(0,\cdot)\, dx \tag{2.27}$$

$$\equiv \int_\Omega \left(\frac{1}{2\varrho_{0,\varepsilon}} |(\varrho\mathbf{u})_{0,\varepsilon}|^2 + (\varrho e)_{0,\varepsilon} + E_{R,0,\varepsilon} \right) dx \le E_0,$$

$$\int_\Omega \varrho_\varepsilon s(\varrho_\varepsilon, \vartheta_\varepsilon)(0,\cdot)\, dx \equiv \int_\Omega (\varrho s)_{0,\varepsilon}\, dx \ge S_0,$$

and

$$0 \le I_\varepsilon(0,\cdot) \equiv I_{0,\varepsilon}(\cdot) \le I_0,\ |I_{0,\varepsilon}(\cdot,\nu)| \le h(\nu) \text{ for a certain } h \in L^1(0,\infty).$$

Then

$$\varrho_\varepsilon \to \varrho \text{ in } C_{weak}([0,T]; L^{5/3}(\Omega)),$$

$$\mathbf{u}_\varepsilon \to \mathbf{u} \text{ weakly in } L^2(0,T; W_0^{1,2}(\Omega)),$$

$$\vartheta_\varepsilon \to \vartheta \text{ weakly in } L^2(0,T; W^{1,2}(\Omega)),$$

and

$$I_\varepsilon \to I \text{ weakly-(*) in } L^\infty((0,T) \times \Omega \times \mathcal{S}^2 \times (0,\infty)),$$

at least for suitable subsequences, where $\{\varrho, \mathbf{u}, \vartheta, I\}$ is a weak solution of problem (2.1)–(2.10).

Remark 4 In comparison with the standard Navier–Stokes–Fourier system studied in [20], problem (2.1)–(2.10) features a new principal difficulty due to the apparent discrepancy between the classical (non-relativistic) description of the fluid motion, and the behavior of photons traveling with the speed of light. In particular, in contrast with the Second law of thermodynamics, the associated entropy equation may contain a *negative* production term. This problem, related to the fact that, hypothetically, one might have $|\mathbf{u}| > c$, has already been observed by Buet and Després [9, Section 2.5]. On the other hand, non-negativity of the entropy production rate plays a crucial role in the approach developed in [13]; whence its adaptation to the present setting requires new ideas. Instead of introducing the radiation entropy, we keep the classical form of the entropy balance equation supplemented with the relevant "radiation" production term proportional to $\frac{1}{\vartheta}\Big(\mathbf{u} \cdot \mathbf{S}_F - S_E\Big)$. As pointed out, this term may change sign and, accordingly, we have to establish its "weak continuity" with respect to ϑ, $\mathbf{u}$, and I contained in $\mathbf{S}_F$, S_E. Note that this is quite delicate as the velocity field $\mathbf{u}$ may develop uncontrolled *time oscillations* on the hypothetical vacuum zones where ϱ vanishes. In order to overcome this difficulty, we use higher regularity of the $\omega-$averages of the radiative intensity discovered by Bardos et al. [2] and Golse et al. [26, 27]. For further generalizations and a more complete list of references, see Bournaveas and Perthame [8].

Sketch of Proof of Theorem 1 (For More Details, See [14]) Uniform (*a priori*) bounds follow from the total energy balance, entropy production equation, and other related physical principles.

From the total energy balance (2.25), combined with hypotheses of Theorem 1, we obtain

$$\operatorname*{ess\ sup}_{t\in(0,T)} \|\sqrt{\varrho_\varepsilon}\mathbf{u}_\varepsilon\|_{L^2(\Omega)} \le c, \tag{2.28}$$

$$\operatorname*{ess\ sup}_{t\in(0,T)} \|\varrho_\varepsilon e(\varrho_\varepsilon, \vartheta_\varepsilon)\|_{L^1(\Omega)} \le c, \tag{2.29}$$

and

$$\operatorname*{ess\ sup}_{t\in(0,T)} \|E_{R,\varepsilon}\|_{L^1(\Omega)} \le c. \tag{2.30}$$

Since the internal energy contains the radiation component proportional to ϑ^4, we deduce from (2.29) that

$$\operatorname*{ess\ sup}_{t\in(0,T)} \|\vartheta_\varepsilon\|_{L^4(\Omega)} \le c, \tag{2.31}$$

and, by virtue of hypotheses (2.11)–(2.14),

$$\operatorname*{ess\ sup}_{t\in(0,T)} \|\varrho_\varepsilon\|_{L^{5/3}(\Omega)} \le c. \tag{2.32}$$

From the transport equation (2.2), using that I_ε is non-negative and applying the "cutoff" hypothesis (2.19), we deduce a uniform bound

$$0 \le I_\varepsilon(t,x,\nu,\boldsymbol{\omega}) \le c(T)(1+ \sup_{x\in\Omega,\ \nu\ge 0,\boldsymbol{\omega}\in\mathcal{S}^2} I_{0,\varepsilon}) \le c(T)(1+I_0) \text{ for any } t \in [0,T]. \tag{2.33}$$

Finally, hypothesis (2.20), together with (2.33), yield

$$\|S_{E,\varepsilon}\|_{L^\infty((0,T)\times\Omega)} + \|\mathbf{S}_{F,\varepsilon}\|_{L^\infty((0,T)\times\Omega)} \le c, \tag{2.34}$$

which, combined with hypothesis (2.21), implies

$$\left\| \frac{1}{\vartheta_\varepsilon} S_{E,\varepsilon} \right\|_{L^\infty((0,T)\times\Omega)} + \left\| \frac{1}{\vartheta_\varepsilon} \mathbf{S}_{F,\varepsilon} \right\|_{L^\infty((0,T)\times\Omega)} \le c. \tag{2.35}$$

Since the viscosity coefficients satisfy (2.17), we get

$$\int_0^T \int_\Omega \frac{1}{\vartheta_\varepsilon} \mathbb{T}_\varepsilon : \nabla_x \mathbf{u}_\varepsilon \, \mathrm{d}x \, \mathrm{d}t \ge c_2 \|\mathbf{u}_\varepsilon\|^2_{L^2(0,T;W_0^{1,2}(\Omega))},$$

where we have used a variant of the standard Korn's inequality.
On the other hand, since Ω is a bounded, in accordance with (2.35),

$$\left| \int_0^T \int_\Omega \frac{1}{\vartheta_\varepsilon} \mathbf{u}_\varepsilon \cdot \mathbf{S}_{F,\varepsilon} \, \mathrm{d}x \, \mathrm{d}t \right| \le c \|\mathbf{u}_\varepsilon\|_{L^1((0,T)\times\Omega)};$$

whence the entropy inequality (2.24) yields the uniform bounds

$$\|\mathbf{u}_\varepsilon\|_{L^2(0,T;W_0^{1,2}(\Omega))} \le c, \tag{2.36}$$

$$\|\nabla_x \vartheta_\varepsilon\|_{L^2((0,T)\times\Omega)} \le c. \tag{2.37}$$

Following the same way as in [20, Chapter 3]

$$\int_0^T \int_\Omega p(\varrho_\varepsilon, \vartheta_\varepsilon)\varrho_\varepsilon^\omega \, \mathrm{d}x \, \mathrm{d}t < c, \text{ with } c \text{ independent of } \varepsilon, \tag{2.38}$$

in particular,

$$\{p(\varrho_\varepsilon, \vartheta_\varepsilon)\}_{\varepsilon>0} \text{ is bounded in } L^p((0,T)\times\Omega) \text{ for a certain } p > 1. \tag{2.39}$$

As a second step and the most important part of the proof is to show the weak sequential stability. Weak sequential stability of macroscopic thermodynamic quantities, pointwise convergence of temperature, pointwise convergence of density are similar to [20, Chapter 3] and in detail is described in [14]. Let us stress the convergence of the radiation intensities.

Our aim is to establish the convergence of the quantities

$$\frac{1}{\vartheta_\varepsilon}\mathbf{u}_\varepsilon \cdot \mathbf{S}_{F,\varepsilon} = \frac{1}{c\vartheta_\varepsilon}\mathbf{u}_\varepsilon \cdot \int_0^\infty \sigma_a(\nu,\vartheta_\varepsilon)\left(\int_{\mathcal{S}^2} \boldsymbol{\omega}\,(B(\nu,\vartheta_\varepsilon) - I_\varepsilon)\ \mathrm{d}\boldsymbol{\omega}\right)\mathrm{d}\nu$$

$$+\frac{1}{c\vartheta_\varepsilon}\mathbf{u}_\varepsilon \cdot \int_0^\infty \sigma_s(\nu,\vartheta_\varepsilon)\left(\int_{\mathcal{S}^2} \boldsymbol{\omega}\left(\left(\frac{1}{4\pi}\int_{\mathcal{S}^2} I_\varepsilon\ \mathrm{d}\boldsymbol{\omega}\right) - I_\varepsilon\right)\mathrm{d}\boldsymbol{\omega}\right)\mathrm{d}\nu$$

and

$$\frac{1}{\vartheta_\varepsilon}S_{E,\varepsilon} = \frac{1}{c\vartheta_\varepsilon}\int_0^\infty \sigma_a(\nu,\vartheta_\varepsilon)\left(\int_{\mathcal{S}^2} (B(\nu,\vartheta_\varepsilon) - I_\varepsilon)\ \mathrm{d}\boldsymbol{\omega}\right)\mathrm{d}\nu.$$

Since $\vartheta_\varepsilon \to \vartheta$ a.a. in $(0,T)\times\Omega$, and

$$\mathbf{u}_\varepsilon \to \mathbf{u} \text{ weakly in } L^2(0,T;W_0^{1,2}(\Omega;\mathbb{R}^3))$$

the desired result follows from compactness of the velocity averages over the sphere $\mathcal{S}^2$ established by Golse et al. [26, 27], see also Bournevas and Perthame [8], and hypothesis (2.20). Specifically, we use the following result (see [26]):

Proposition 1 *Let $I \in L^q([0,T]\times\mathbb{R}^{n+1}\times\mathcal{S}^2)$, $\partial_t I + \omega\cdot\nabla_x I \in L^q([0,T]\times\mathbb{R}^{n+1}\times\mathcal{S}^2)$ for a certain $q>1$. In addition, let $I_0 \equiv I(0,\cdot) \in L^\infty(\mathbb{R}^{n+1}\times\mathcal{S}^2)$. Then*

$$\tilde{I} \equiv \int_{\mathcal{S}^2} I(\cdot,\nu)\,d\boldsymbol{\omega}$$

belongs to the space $W^{s,q}([0,T]\times\mathbb{R}^{n+1})$ for any s, $0<s<\inf\{1/q, 1-1/q\}$, and

$$\|\tilde{I}\|_{W^{s,q}} \le c(I_0)(\|I\|_{L^q} + \|\partial_t I + \omega\cdot\nabla I\|_{L^q}).$$

As the radiation intensity I_ε satisfies the transport equation (2.2), by virtue of the cutoff hypothesis (2.9)–(2.11) where S is bounded in $L^q \cap L^\infty([0,T)\times\Omega\times\mathbb{R}^1\times S^2)$, a direct application of Proposition 1 yields the desired conclusion

$$\int_{\mathcal{S}^2} I_\varepsilon(\cdot,\nu)\,\mathrm{d}\boldsymbol{\omega} \to \int_{\mathcal{S}^2} I(\cdot,\nu)\,\mathrm{d}\boldsymbol{\omega} \text{ in } L^2((0,T)\times\Omega)$$

and

$$\int_{\mathcal{S}^2} \boldsymbol{\omega} I_\varepsilon(\cdot,\nu)\,\mathrm{d}\boldsymbol{\omega} \to \int_{\mathcal{S}^2} \boldsymbol{\omega} I(\cdot,\nu)\,\mathrm{d}\boldsymbol{\omega} \text{ in } L^2((0,T)\times\Omega)$$

for any fixed ν.

Consequently,

$$\frac{1}{\vartheta_\varepsilon}\mathbf{u}_\varepsilon \cdot F_{s,\varepsilon} \to \frac{1}{\vartheta}\mathbf{u} \cdot F_s$$

and, similarly,

$$\frac{1}{\vartheta_\varepsilon} S_{E,\varepsilon} \to \frac{1}{\vartheta} S_E$$

as required. Note that strong (a.a. pointwise) convergence of the $\omega-$averages is needed as $\mathbf{u}_\varepsilon$ may fail to converge strongly on hypothetical vacuum zones, meaning on the part of Ω where the limit density ϱ vanishes.

Remark 5 For steady variant of radiative hydrodynamics we refer to [32].

2.5 Semi-Relativistic Models

We consider a "semi-relativistic" model of radiative viscous compressible Navier–Stokes–Fourier system coupled to the radiative transfer equation extending the classical model introduced in the previous subsection. The effect of radiation is still incorporated in the radiative intensity $I = I(t, x, \boldsymbol{\omega}, \nu)$, depending on the direction $\boldsymbol{\omega} \in \mathcal{S}^2$, where $\mathcal{S}^2 \subset \mathbb{R}^3$ denotes the unit sphere, and the frequency $\nu \geq 0$, but we take into account their relativistic corrections. The evolution of I is described by a transport equation with a source term and the fluid-radiation coupling is expressed through radiative sources in the momentum and energy equations. As usual we suppose that the radiative source S is splitted into two terms

$$\begin{aligned} S = \sigma_a\Big[B(\nu, \boldsymbol{\omega}, \mathbf{u}, \vartheta) - I(t, x, \nu, \boldsymbol{\omega})\Big] + \sigma_s \left(\tfrac{1}{4\pi}\int_{\mathcal{S}^2} I(t, x, \nu, \boldsymbol{\omega}')\, d\boldsymbol{\omega}' - I(t, x, \nu, \boldsymbol{\omega})\right) \\ =: S_{a,e} + S_s. \end{aligned} \tag{2.40}$$

In the right-hand side the first is the emission-absorption contribution where $\sigma_a > 0$ is the absorption coefficient and B is a perturbation of the equilibrium Planck's function given by

$$B(\nu, \boldsymbol{\omega}, \mathbf{u}, \vartheta) = \frac{2h}{c^2}\frac{\nu^3}{e^{\frac{h\nu}{k\vartheta}\left(1-\alpha\frac{\boldsymbol{\omega}\cdot\mathbf{u}}{c}\right)} - 1}, \tag{2.41}$$

where h is the Planck's constant, k is the Boltzmann's constant, and $0 \leq \alpha(\vartheta) \leq 1$ is a smooth function, to be determined below. One observes that for $\frac{|\mathbf{u}|}{c} << 1$ one recovers the standard equilibrium Planck's function $B(\nu, \vartheta) = \frac{2h}{c^2}\frac{\nu^3}{e^{\frac{h\nu}{k\vartheta}} - 1}$.

Note that the idea of this kind of perturbation is not new and has been extensively used in recent works on radiative transfer [7, 9, 12, 19], for example, in the $M1$ Levermore model [37, 38].

We first suppose that the transport coefficients are smooth functions satisfying $\sigma_a(\vartheta, \mathbf{u}) = \chi(|\mathbf{u}|)\tilde{\sigma}_a(\vartheta) \geq 0$ and $\sigma_s(\vartheta) \geq 0$ and that both depend neither on angular variable (2.1 - 2.2) (isotropy of radiation) nor on frequency (the so-called grey hypothesis).

The function χ appearing in the emission-absorption coefficient is a C^∞ cutoff satisfying

$$\chi(s) = \begin{cases} 1 \ \text{if} \ s \leq c, \\ 0 \ \text{if} \ s \geq c + \beta, \end{cases}$$

for an arbitrary $\beta > 0$. The role of this cutoff is to deal with the singularity of B and its meaning is the following: in the "over-relativistic" regime ($|\mathbf{u}| \geq c$) where special relativity would be violated, we decide to decouple matter and radiation. Of course this is an arbitrary choice but only a meaningless region with respect to physics is concerned (recall that in the relativistic setting [9], Lorentz factors of the type $\left(1 - \frac{\mathbf{u}^2}{c^2}\right)^{1/2}$ become singular for $|\mathbf{u}| = c$).

In comparison with the weak formulation 2.3 the entropy inequality is replaced by

$$\begin{aligned} &\int_0^T \int_\Omega \Big([\varrho s + s^R]\partial_t \varphi + \varrho s \mathbf{u} \cdot \nabla_x \varphi + [\frac{\mathbf{q}}{\vartheta} + \mathbf{q}^R] \cdot \nabla_x \varphi \Big) \, \mathrm{d}x \, \mathrm{d}t \\ &\leq - \int_\Omega (\varrho s + s^R)_0 \varphi(0, \cdot) \, \mathrm{d}x - \int_0^T \int_\Omega \frac{1}{\vartheta} \left(\mathbb{S} : \nabla_x \mathbf{u} - \frac{\mathbf{q} \cdot \nabla_x \vartheta}{\vartheta} \right) \varphi \, \mathrm{d}x \, \mathrm{d}t \\ &- \frac{k}{h} \int_0^T \int_\Omega \Bigg[\int_0^\infty \int_{\mathcal{S}^2} \frac{1}{\nu} \left[\log \frac{n(I)}{n(I)+1} - \log \frac{n(B)}{n(B)+1} \right] \sigma_a (B - I) \, d\boldsymbol{\omega} d\nu \\ &+ \int_0^\infty \int_{\mathcal{S}^2} \frac{1}{\nu} \left[\log \frac{n(I)}{n(I)+1} - \log \frac{n(\tilde{I})}{n(\tilde{I})+1} \right] \sigma_s (\tilde{I} - I) \, d\boldsymbol{\omega} d\nu \Bigg] \varphi \, \mathrm{d}x \mathrm{d}x \, \mathrm{d}t \end{aligned} \tag{2.42}$$

for any $\varphi \in C_c^\infty([0, T) \times \overline{\Omega})$, $\varphi \geq 0$, where the sign of all the terms in the right-hand side may be controlled, where s_R is the radiative entropy and q_R is the radiative entropy flux.[3]

[3] Let us recall [1] the formula for the entropy of a photon gas

$$s^R = -\frac{2k}{c^3} \int_0^\infty \int_{\mathcal{S}^2} \nu^2 \left[n \log n - (n+1) \log(n+1) \right] d\boldsymbol{\omega} d\nu, \tag{2.43}$$

where $n = n(I) = \frac{c^2 I}{2h\alpha^3 \nu^3}$ is the occupation number. Defining the radiative entropy flux

Definition 2 We say that $\varrho, \mathbf{u}, \vartheta, I$ is a weak solution of problem (2.1)–(2.10), (2.40), (2.41) if

$$\varrho \geq 0,\ \vartheta > 0 \text{ for a.a. } (t,x) \times \Omega,\ I \geq 0 \text{ a.a. in } (0,T) \times \Omega \times \mathcal{S}^2 \times (0,\infty),$$

$$\varrho \in L^\infty(0,T; L^{5/3}(\Omega)),\ \vartheta \in L^\infty(0,T; L^4(\Omega)),$$

$$\mathbf{u} \in L^2(0,T; W_0^{1,2}(\Omega)),\ \vartheta \in L^2(0,T; W^{1,2}(\Omega)),$$

$$I \in L^\infty((0,T) \times \Omega \times \mathcal{S}^2 \times (0,\infty)),\ I \in L^\infty(0,T; L^1(\Omega \times \mathcal{S}^2 \times (0,\infty)),$$

and if ϱ, $\mathbf{u}$, ϑ, I satisfy the integral identities (2.22), (2.23), (2.42), (2.25), together with the transport equation (2.2).

The existence result reads now

Theorem 2 *Let $\Omega \subset \mathbb{R}^3$ be a bounded Lipschitz domain. Assume that the thermodynamic functions p, e, s satisfy hypotheses (2.11)–(2.16), that B satisfies (2.41) and (2.46), and that the transport coefficients μ, λ, κ, σ_a, and σ_s comply with (2.17)–(2.20). Let $\{\varrho_\varepsilon, \mathbf{u}_\varepsilon, \vartheta_\varepsilon, I_\varepsilon\}_{\varepsilon>0}$ be a family of weak solutions to problem (2.1)–(2.10) in the sense of Definition 2 such that*

$$\varrho_\varepsilon(0,\cdot) \equiv \varrho_{\varepsilon,0} \to \varrho_0 \text{ in } L^{5/3}(\Omega), \tag{2.47}$$

$$\int_\Omega \left(\frac{1}{2}\varrho_\varepsilon|\mathbf{u}_\varepsilon|^2 + \varrho_\varepsilon e(\varrho_\varepsilon, \vartheta_\varepsilon) + E_{R,\varepsilon}\right)(0,\cdot)\,dx \equiv \int_\Omega \left(\frac{1}{2\varrho_{0,\varepsilon}}|(\varrho\mathbf{u})_{0,\varepsilon}|^2 + (\varrho e)_{0,\varepsilon} + E_{R,0,\varepsilon}\right)dx \leq E_0, \tag{2.48}$$

$$\int_\Omega [\varrho_\varepsilon s(\varrho_\varepsilon, \vartheta_\varepsilon) + s^R(I_\varepsilon)](0,\cdot)\,dx \equiv \int_\Omega (\varrho s + s^R)_{0,\varepsilon}\,dx \geq S_0,$$

$$\mathbf{q}^R = -\frac{2k}{c^2}\int_0^\infty \int_{\mathcal{S}^2} \nu^2 \left[n \log n - (n+1)\log(n+1)\right] \boldsymbol{\omega}\, d\boldsymbol{\omega} d\nu, \tag{2.44}$$

and using the radiative transfer equation, we get the equation

$$\partial_t s^R + \operatorname{div}_x \mathbf{q}^R = -\frac{k}{h}\int_0^\infty \int_{\mathcal{S}^2} \frac{1}{\nu} \log \frac{n}{n+1}\, S\, d\boldsymbol{\omega} d\nu =: \varsigma^R. \tag{2.45}$$

Moreover, $\log \frac{n(B)}{n(B)+1} = -\frac{h\nu}{k\vartheta}\left(1 - \alpha \frac{\boldsymbol{\omega}\cdot\mathbf{u}}{c}\right)$ and

$$\alpha = \frac{\sigma_a + \sigma_s}{\sigma_a + 2\sigma_s}, \tag{2.46}$$

For more details, see [18].

and

$$0 \leq I_\varepsilon(0, \cdot) \equiv I_{0,\varepsilon}(\cdot) \leq I_0, \ |I_{0,\varepsilon}(\cdot, \nu)| \leq h(\nu) \textit{ for a certain } h \in L^1(0, \infty).$$

Then

$$\varrho_\varepsilon \to \varrho \textit{ in } C_{weak}([0, T]; L^{5/3}(\Omega)),$$

$$\mathbf{u}_\varepsilon \to \mathbf{u} \textit{ weakly in } L^2(0, T; W_0^{1,2}(\Omega)),$$

$$\vartheta_\varepsilon \to \vartheta \textit{ weakly in } L^2(0, T; W^{1,2}(\Omega)),$$

and

$$I_\varepsilon \to I \textit{ weakly-(*) in } L^\infty((0, T) \times \Omega \times \mathcal{S}^2 \times (0, \infty)),$$

at least for suitable subsequences, where $\{\varrho, \mathbf{u}, \vartheta, I\}$ *is a weak solution of problem (2.1)–(2.10).*

Remark 6 The reason for introducing such type of model is to recover the crucial positivity property for the production rate of total entropy which is missing in the previous model. Using such model the singular limits were investigated, see [18]. Before the singular limits (see [16, 17]) were studied for a simplified model of radiation hydrodynamics introduced by Teleaga, Seaïd, Gasser, Klar, and Struckmeier in [52]. The main idea in the modeling is to introduce in the complete model of [14] a perturbed Planck's function and a suitable (relativistic) velocity cutoff (this is the meaning we give to "semi-relativistic" model) allowing to recover this crucial positivity property for the production rate of total entropy. As the perturbation will be small (going formally to zero as $c \to \infty$), one can expect to obtain the correct limit regimes.

3 Inviscid Case

3.1 Euler System with Damping Term

In this part we consider two models which can be seen as target systems of two singular limits. Precisely, we consider a compressible inviscid radiative flow where the motion of the fluid is given by the Euler system with damping for the evolution of the density $\varrho = \varrho(t, x)$, the velocity field $\mathbf{u} = \mathbf{u}(t, x)$, and the absolute temperature $\vartheta = \vartheta(t, x)$ as functions of the time t and the Eulerian spatial coordinate $x \in \mathbb{R}^3$.

In the first regime (equilibrium diffusion), the effect of radiation is incorporated in the state functions p (pressure) and e (internal energy). In the second regime (non-equilibrium diffusion), the radiation appears through an extra equation of parabolic type for the radiative temperature which is *a priori* different from the matter temperature.

More specifically, in the equilibrium case, the system of equations to be studied for the three unknowns $(\varrho, \mathbf{u}, \vartheta)$ reads

$$\partial_t \varrho + \mathrm{div}_x(\varrho \mathbf{u}) = 0, \tag{3.1}$$

$$\partial_t(\varrho \mathbf{u}) + \mathrm{div}_x(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla_x (p + p_r) + \nu \mathbf{u} = 0, \tag{3.2}$$

$$\partial_t (\varrho E + E_r) + \mathrm{div}_x [(\varrho E + E_r)\, \mathbf{u} + (p + p_r)\mathbf{u}] = \mathrm{div}_x (\kappa \nabla_x \vartheta) + \mathrm{div}_x \left(\frac{1}{3\sigma_a} \nabla_x E_r \right), \tag{3.3}$$

where $E = \frac{1}{2}|\mathbf{u}|^2 + e(\varrho, \vartheta)$, $E_r = a\vartheta^4$, and $p_r = \frac{a}{3}\, \vartheta^4$.

In the non-equilibrium case, the system of equations for the four unknowns $(\varrho, \mathbf{u}, \vartheta, E_r)$ is

$$\partial_t \varrho + \mathrm{div}_x(\varrho \mathbf{u}) = 0, \tag{3.4}$$

$$\partial_t(\varrho \mathbf{u}) + \mathrm{div}_x(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla_x(p + p_r) + \nu \mathbf{u} = 0, \tag{3.5}$$

$$\partial_t (\varrho E) + \mathrm{div}_x ((\varrho E + p)\mathbf{u}) + \mathbf{u} \cdot \nabla_x p_r = \mathrm{div}_x (\kappa \nabla_x \vartheta) - \sigma_a \left(a\vartheta^4 - E_r \right), \tag{3.6}$$

$$\partial_t E_r + \mathrm{div}_x (E_r \mathbf{u}) + p_r \mathrm{div}_x \mathbf{u} = \mathrm{div}_x \left(\frac{1}{3\sigma_s}\, \nabla_x E_r \right) - \sigma_a \left(E_r - a\vartheta^4 \right), \tag{3.7}$$

where $E = \frac{1}{2}|\mathbf{u}|^2 + e(\varrho, \vartheta)$, E_r is the radiative energy related to the temperature of radiation T_r by $E_r = aT_r^4$, and p_r is the radiative pressure given by $p_r = \frac{1}{3}aT_r^4 = \frac{1}{3}\, E_r$, with $a > 0$.

Systems (3.1)–(3.3) and (3.4)–(3.7) can be viewed as singular limits in radiation hydrodynamics in two limit diffusion regimes. Such systems (when damping is absent) have been investigated by Lowrie, Morel, and Hittinger [36] and more recently by Buet and Després [9].

3.1.1 Hypotheses

Hypotheses imposed on constitutive relations and transport coefficients are motivated by the general (local) existence theory for the Euler–Fourier system developed in [48, 49] (see also [20, Chapter 3] for the Navier–Stokes–Fourier framework) and reasonable physical assumptions for the radiative part [40, 45]. In our simplified setting, transport coefficients $\kappa, \sigma_a, \sigma_s$ and the Planck's coefficient are supposed to be fixed positive numbers. The damping with coefficient $\nu > 0$ of Darcy type can be interpreted here as a diffusion of a light gas into a heavy one.

We consider the pressure in the form (2.11) with $a = 0$

$$p(\varrho, \vartheta) = \vartheta^{5/2} P\left(\frac{\varrho}{\vartheta^{3/2}}\right), \tag{3.8}$$

where $P : [0, \infty) \to [0, \infty)$ is a given function with the properties (2.12)–(2.14).

After Maxwell's relations, the specific internal energy e and the associated specific entropy have form (2.15)–(2.16) with $a = 0$ and with

$$M'(Z) = -\frac{3}{2} \frac{\frac{5}{3} P(Z) - P'(Z)Z}{Z^2} < 0.$$

Theorem 3 *Let $\left(\overline{\varrho}, 0, \overline{\vartheta}\right)$ be a constant state with $\overline{\varrho} > 0$, $\overline{\vartheta} > 0$. Consider $d > 7/2$. There exists $\varepsilon > 0$ such that, for any initial state $(\varrho_0, \mathbf{u}_0, \vartheta_0)$ satisfying*

$$\left\| (\varrho_0, \mathbf{u}_0, \vartheta_0) - \left(\overline{\varrho}, 0, \overline{\vartheta}\right) \right\|_{W^{d,2}(\mathbb{R}^3)} \leq \varepsilon, \tag{3.9}$$

there exists a unique global solution $(\varrho, \mathbf{u}, \vartheta)$ to (3.1)–(3.3), such that $\left(\varrho - \overline{\varrho}, \mathbf{u}, \vartheta - \overline{\vartheta}\right) \in C\left([0, +\infty); W^{d,2}(\mathbb{R}^3)\right)$. In addition, this solution satisfies the following energy inequality:

$$\begin{aligned} &\left\| (\varrho(t) - \overline{\varrho}, \mathbf{u}(t), \vartheta(t) - \overline{\vartheta}) \right\|_{W^{d,2}\left(\mathbb{R}^3\right)} + \int_0^t \left(\| \nabla_x \left(\varrho, \mathbf{u}, \vartheta\right)(s) \|^2_{W^{d-1,2}\left(\mathbb{R}^3\right)} + \| \nabla_x \vartheta(s) \|^2_{W^{d,2}\left(\mathbb{R}^3\right)} \right) ds \\ &\leq C \left\| (\varrho_0 - \overline{\varrho}, 0, \vartheta_0 - \overline{\vartheta}) \right\|^2_{W^{d,2}\left(\mathbb{R}^3\right)}, \end{aligned} \tag{3.10}$$

for some constant $C > 0$ which does not depend on t.

The same result holds in the case of system (3.4)–(3.7):

Theorem 4 *Let $\left(\overline{\varrho}, 0, \overline{\vartheta}, \overline{E_r}\right)$ be a constant state with $\overline{\varrho} > 0$, $\overline{\vartheta} > 0$, $\overline{E_r} > 0$. Consider $d > 7/2$. There exists $\varepsilon > 0$ such that, for any initial state $\left(\varrho_0, \mathbf{u}_0, \vartheta_0, E_r^0\right)$ satisfying*

$$\left\| \left(\varrho_0, \mathbf{u}_0, \vartheta_0, E_r^0\right) - \left(\overline{\varrho}, 0, \overline{\vartheta}, \overline{E_r}\right) \right\|_{W^{d,2}(\mathbb{R}^3)} \leq \varepsilon, \tag{3.11}$$

there exists a unique global solution $(\varrho, \mathbf{u}, \vartheta, E_r)$ to (3.4)–(3.7), such that $\left(\varrho - \overline{\varrho}, \mathbf{u}, \vartheta - \overline{\vartheta}, \overline{E_r} - E_r\right) \in C\left([0, +\infty); W^{d,2}\left(\mathbb{R}^3\right)\right)$. In addition, this solution satisfies the following energy inequality:

$$
\begin{aligned}
&\left\|(\varrho(t)-\overline{\varrho}, \mathbf{u}(t), \vartheta(t)-\overline{\vartheta}, E_r(t)-\overline{E_r})\right\|_{W^{d,2}\left(\mathbb{R}^3\right)} + \int_0^t \|\nabla_x (\varrho, \mathbf{u}, \vartheta, E_r)(s)\|^2_{W^{d-1,2}\left(\mathbb{R}^3\right)} \, ds \\
&+ \int_0^t \left(\|\nabla_x \vartheta(s)\|^2_{W^{d,2}\left(\mathbb{R}^3\right)} + \|\nabla_x E_r(s)\|^2_{W^{d,2}\left(\mathbb{R}^3\right)} \right) ds \\
&\leq C \left\|\left(\varrho_0-\overline{\varrho}, 0, \vartheta_0-\overline{\vartheta}, E_r^0-\overline{E_r}\right)\right\|^2_{W^{d,2}\left(\mathbb{R}^3\right)},
\end{aligned}
\tag{3.12}
$$

for some constant $C > 0$ which does not depend on t.

Sketch of Proof of Theorems 3 and 4 We will focus on the non-equilibrium problem. Equation (3.6) rewrites

$$
\varrho C_v \left(\partial_t \vartheta + \mathbf{u} \cdot \nabla_x \vartheta\right) + \vartheta p_\vartheta \operatorname{div}_x \mathbf{u} = -p_r \operatorname{div}_x \mathbf{u} + \operatorname{div}_x \left(\kappa \nabla_x \vartheta\right) - \sigma_a \left(a\vartheta^4 - E_r\right) - \nu \left|\mathbf{u}\right|^2 . \tag{3.13}
$$

We linearize the system (3.4)–(3.6) around the constant state $(\overline{\varrho}, 0, \overline{\vartheta}, \overline{E}_r)$ with the compatibility condition $\overline{E}_r = a\overline{\vartheta}^4$ and putting $\varrho = r + \overline{\varrho}$, $\vartheta = T + \overline{\vartheta}$ and $E_r = e_r + \overline{E}_r$ we get

$$
\partial_t r + \overline{\varrho} \operatorname{div}_x \mathbf{u} = 0, \tag{3.14}
$$

$$
\partial_t \mathbf{u} + \frac{\overline{p}_\varrho}{\overline{\varrho}} \nabla_x r + \frac{\overline{p}_\vartheta}{\overline{\varrho}} \nabla_x T + \frac{1}{3\overline{\varrho}} \nabla_x e_r + \nu \mathbf{u} = 0, \tag{3.15}
$$

$$
\partial_t T + \frac{\overline{\vartheta}\,\overline{p}_\vartheta}{\overline{\varrho}\,\overline{C}_v} \operatorname{div}_x \mathbf{u} = \operatorname{div}_x \left(\frac{\kappa}{\overline{\varrho}\,\overline{C}_v} \nabla_x T \right) - \frac{\sigma_a}{\overline{\varrho}\,\overline{C}_v} \left(4a\overline{\vartheta}^3 T - e_r \right), \tag{3.16}
$$

$$
\partial_t e_r + \frac{4}{3} \overline{E}_r \operatorname{div}_x \mathbf{u} = \operatorname{div}_x \left(\frac{1}{3\sigma_s} \nabla_x e_r \right) - \sigma_a \left(e_r - 4a\overline{\vartheta}^3 T \right), \tag{3.17}
$$

using the vector notation $U := \begin{pmatrix} r \\ u_1 \\ u_2 \\ u_3 \\ T \\ e_r \end{pmatrix}$, the linearized system (3.14)–(3.17) rewrites

$$
\partial_t U + \sum_{j=1}^{3} A_j \partial_j U = D \Delta U - BU, \tag{3.18}
$$

for details, see below. [4,5]

Applying the symmetrization and the Kreiss theorem and the Shizuta–Kawashima condition (SK) we get the existence of the linearized model. Finally the fixed point argument is applied.

For more details, see [4].

3.2 Non-isentropic Euler–Maxwell's System Coupled with Transport of Radiation

We considered a compressible electro-magnetic inviscid radiative flow coupled where the motion of the fluid is given by the Euler system for the evolution of

[4]

$$A_1 := \begin{pmatrix} 0 & \overline{\varrho} & 0 & 0 & 0 & 0 \\ \alpha & 0 & 0 & 0 & \beta & \frac{1}{3\overline{\varrho}} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 & 0 & 0 \\ 0 & \delta & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_2 := \begin{pmatrix} 0 & 0 & \overline{\varrho} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \alpha & 0 & 0 & 0 & \beta & \frac{1}{3\overline{\varrho}} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma & 0 & 0 & 0 \\ 0 & 0 & \delta & 0 & 0 & 0 \end{pmatrix},$$

$$A_3 := \begin{pmatrix} 0 & 0 & 0 & \overline{\varrho} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \alpha & 0 & 0 & 0 & \beta & \frac{1}{3\overline{\varrho}} \\ 0 & 0 & 0 & \gamma & 0 & 0 \\ 0 & 0 & 0 & \delta & 0 & 0 \end{pmatrix},$$

and

$$D := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \nu \end{pmatrix}, \quad B := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \nu & 0 & 0 & 0 & 0 \\ 0 & 0 & \nu & 0 & 0 & 0 \\ 0 & 0 & 0 & \nu & 0 & 0 \\ 0 & 0 & 0 & 0 & \zeta & -\eta \\ 0 & 0 & 0 & 0 & -\pi & \sigma_a \end{pmatrix},$$

[5]

$$\alpha = \frac{\overline{p}_\varrho}{\overline{\varrho}}, \quad \beta = \frac{\overline{p}_\vartheta}{\overline{\varrho}}, \quad \gamma = \frac{\overline{\vartheta}\,\overline{p}_\vartheta}{\overline{\varrho}\,\overline{C}_v}, \quad \delta = \frac{4}{3}\overline{E}_r, \quad \mu = \frac{\kappa}{\overline{\varrho}\,\overline{C}_v},$$

$$\tau = \frac{1}{3\sigma_s}, \quad \zeta = \frac{4a\sigma_a\overline{\vartheta}^3}{\overline{\varrho}\,\overline{C}_v}, \quad \eta = \frac{\sigma_a}{\overline{\varrho}\,\overline{C}_v}, \quad \pi = 4a\sigma_a\overline{\vartheta}^3.$$

the density $\varrho = \varrho(t,x)$, the velocity field $\mathbf{u} = \mathbf{u}(t,x)$, and the absolute temperature $\vartheta = \vartheta(t,x)$, and where radiation is described in the limit by an extra temperature $T_r = T_r(t,x)$. All of these quantities are functions of the time t and the Eulerian spatial coordinate $x \in \mathbb{R}^3$.

More specifically the system of equations to be studied for the unknowns $(\varrho, \mathbf{u}, \vartheta, E_r, \mathbf{B}, \mathbf{E})$ reads

$$\partial_t \varrho + \mathrm{div}_x(\varrho \mathbf{u}) = 0, \tag{3.19}$$

$$\partial_t(\varrho \mathbf{u}) + \mathrm{div}_x(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla_x(p + p_r) = -\rho\,(\mathbf{E} + \mathbf{u} \times \mathbf{B}) - \nu\rho\mathbf{u}, \tag{3.20}$$

$$\partial_t\,(\varrho E) + \mathrm{div}_x\,((\varrho E + p)\mathbf{u}) + \mathbf{u}\cdot\nabla_x p_r = -\sigma_a\left(a\vartheta^4 - E_r\right) - \rho\mathbf{E}\cdot\mathbf{u}, \tag{3.21}$$

$$\partial_t E_r + \mathrm{div}_x\,(E_r\mathbf{u}) + p_r \mathrm{div}_x \mathbf{u} = -\sigma_a\left(E_r - a\vartheta^4\right), \tag{3.22}$$

$$\partial_t \mathbf{B} + \mathrm{curl}_x \mathbf{E} = 0, \tag{3.23}$$

$$\partial_t \mathbf{E} - \mathrm{curl}_x \mathbf{B} = \varrho \mathbf{u}, \tag{3.24}$$

$$\mathrm{div}_x \mathbf{B} = 0, \tag{3.25}$$

$$\mathrm{div}_x \mathbf{E} = \overline{\varrho} - \varrho, \tag{3.26}$$

where $\mathbf{E}$ is the electric field and $\mathbf{B}$ is the magnetic induction,

We assume that the pressure $p(\varrho, \vartheta)$ and the internal energy $e(\varrho, \vartheta)$ are positive smooth functions of their arguments with

$$C_v := \frac{\partial e}{\partial \vartheta} > 0, \quad \frac{\partial p}{\partial \varrho} > 0,$$

and we also suppose for simplicity that $\nu = \frac{1}{\tau}$ (where $\tau > 0$ is a momentum-relaxation time), μ, σ_a, and a are positive constants.

A simplification appears if one observes that, provided that Eqs. (3.25) and (3.26) are satisfied at $t = 0$, they are satisfied for any time $t > 0$ and consequently they can be discarded from the analysis below.

Notice that the reduced system (3.19)–(3.22) is the non-equilibrium regime of radiation hydrodynamics introduced by Lowrie, Morel, and Hittinger [36] and more recently by Buet and Després [9] and studied mathematically by Blanc, Ducomet, and Nečasová [4]. Extending this last analysis, our goal in this part is to prove global existence of solutions for the system (3.19)–(3.26) when data are sufficiently close to an equilibrium state and study their large time behavior.

We mention for completeness that related non-isentropic Euler–Maxwell systems have been the object of a number of studies in the recent past. see, [23, 24, 28, 33, 51, 55].

In the following we show that the ideas used by Y. Ueda, S. Wang, and S. Kawashima in [53, 54] in the isentropic case can be extended to the (radiative) non-isentropic system (3.19)–(3.24).

We state the result of this subsection that system (3.19)–(3.26) has a global smooth solution close to any equilibrium state.

Theorem 5 *Let* $\left(\overline{\varrho}, 0, \overline{\vartheta}, \overline{E_r}, \overline{\mathbf{B}}, 0\right)$ *be a constant state with* $\overline{\varrho} > 0$, $\overline{\vartheta} > 0$ *and* $\overline{E_r} > 0$ *with compatibility condition* $\overline{E_r} = a\overline{\vartheta}^4$ *and suppose that* $d \geq 3$.

There exists $\varepsilon > 0$ *such that, for any initial state* $\left(\varrho_0, \mathbf{u}_0, \vartheta_0, E_r^0, \mathbf{B}_0, \mathbf{E}_0\right)$ *satisfying*

$$div_x \mathbf{E}_0 = \varrho_0 - \overline{\varrho}, \quad div_x \mathbf{B}_0 = 0,$$

$$\left(\varrho_0 - \overline{\varrho}, \mathbf{u}_0, \vartheta_0 - \overline{\vartheta}, E_{r0} - \overline{E_r}, \mathbf{B}_0 - \overline{\mathbf{B}}, \mathbf{E}_0\right) \in W^{d,2}(\mathbb{R}^3),$$

and

$$\left\| \left(\varrho_0, \mathbf{u}_0, \vartheta_0, E_r^0, \mathbf{B}_0, \mathbf{E}_0\right) - \left(\overline{\varrho}, 0, \overline{\vartheta}, \overline{E_r}, \overline{\mathbf{B}}, 0\right) \right\|_{W^{d,2}(\mathbb{R}^3)} \leq \varepsilon, \tag{3.27}$$

there exists a unique global solution $(\varrho, \mathbf{u}, \vartheta, E_r, \mathbf{B}, \mathbf{E})$ *to (3.19)–(3.26), such that*

$$\left(\varrho - \overline{\varrho}, \mathbf{u}, \vartheta - \overline{\vartheta}, E_r - \overline{E_r}, \mathbf{B} - \overline{\mathbf{B}}, \mathbf{E}\right) \in C\left([0, +\infty); W^{d,2}(\mathbb{R}^3)\right) \cap C^1\left([0, +\infty); W^{d-1,2}(\mathbb{R}^3)\right).$$

In addition, this solution satisfies the following energy inequality:

$$\left\| (\varrho - \overline{\varrho}, \mathbf{u}, \vartheta - \overline{\vartheta}, E_r - \overline{E_r}, \mathbf{B} - \overline{\mathbf{B}}, \mathbf{E})(t) \right\|_{W^{d,2}(\mathbb{R}^3)}$$

$$+ \int_0^t \left(\left\| \left(\varrho - \overline{\varrho}, \mathbf{u}, \vartheta - \overline{\vartheta}, E_r - \overline{E_r}\right)(\tau) \right\|^2_{W^{d,2}(\mathbb{R}^3)} + \|\nabla_x \mathbf{B}(\tau)\|^2_{W^{d-2,2}(\mathbb{R}^3)} + \|\mathbf{E}(\tau)\|^2_{W^{d-1,2}(\mathbb{R}^3)} \right) d\tau$$

$$\leq C \left\| \left(\varrho_0 - \overline{\varrho}, 0, \vartheta_0 - \overline{\vartheta}, E_r^0 - \overline{E_r}, \mathbf{B}_0 - \overline{\mathbf{B}}, \mathbf{E}_0\right) \right\|^2_{W^{d,2}(\mathbb{R}^3)}, \tag{3.28}$$

for some constant $C > 0$ *which does not depend on* t.

The large time behavior of the solution is described as follows

Theorem 6 *Let* $d \geq 3$. *The unique global solution* $(\varrho, \mathbf{u}, \vartheta, E_r, \mathbf{B}, \mathbf{E})$ *to (3.19)–(3.26) defined in Theorem 5 converges to the constant state* $(\overline{\varrho}, \mathbf{0}, \overline{\vartheta}, \overline{E_r}, \overline{\mathbf{B}}, \mathbf{0})$ *uniformly in* $x \in \mathbb{R}^3$ *as* $t \to \infty$. *More precisely*

$$\left\| (\varrho - \overline{\varrho}, \mathbf{u}, \vartheta - \overline{\vartheta}, E_r - \overline{E_r}, \mathbf{E})(t) \right\|_{W^{d-2,\infty}(\mathbb{R}^3)} \to 0 \quad as\ t \to \infty. \tag{3.29}$$

Moreover if $d \geq 4$

$$\left\|(\mathbf{B}-\overline{\mathbf{B}})(t)\right\|_{W^{d-4,\infty}(\mathbb{R}^3)} \to 0 \quad as\ t \to \infty. \tag{3.30}$$

Remark 7 Note that, due to lack of dissipation by viscous, thermal, and radiative fluxes, the Kawashima–Shizuta stability criterion (see [50] and [6]) is not satisfied for the system under study and techniques of [31] relying on the existence of a compensating matrix do not apply. However, it was shown that radiative sources play the role of relaxation terms for temperature and radiative energy and it leads to global existence for the system.

For more details, see [5].

Acknowledgments The work of Š.N. were supported by the Czech Science Foundation grant GA19-04243S and by the programme of the Academy of Sciences of the Czech Republic, Institutional Research Plan 67985840.

References

1. R. Balian, *From Microphysics to Macrophysics*. Methods and Applications of Statistical Physics, vol. II (Springer, Berlin, Heidelberg, New York, 1992)
2. C. Bardos, F. Golse, B. Perthame, R. Sentis, The nonaccretive radiative transfer equations: existence of solutions and Rosseland approximation. J. Funct. Anal. **77**, 434–460 (1988)
3. X. Blanc, B. Ducomet, Weak and strong solutions of equations of compressible magnetohydrodynamics, in *Handbook of Mathematical Analysis in Mechanics of Viscous Fluids* (Springer, Cham, 2018), pp. 2869–2925
4. X. Blanc, B. Ducomet, Š. Nečasová, On some singular limits in damped radiation hydrodynamics. J. Hyperbolic Differ. Equ. **13**(2), 249–271 (2016)
5. X. Blanc, B. Ducomet, Š. Nečasová, Global existence of a radiative Euler system coupled to an electromagnetic field. Adv. Nonlinear Anal. **8**(1), 1158–1170 (2019)
6. C. Beauchard, E. Zuazua, Large time asymptotics for partially dissipative hyperbolic systems. Arch. Rational Mech. Anal. **199**, 177–227 (2011)
7. C. Berthon, C. Buet, J.-F Coulombel, B. Desprès, J. Dubois, T. Goudon, J.E. Morel, R. Turpault, *Mathematical Models and Numerical Methods for Radiative Transfer*. Panoramas et Synthèses, vol. 28 (Société Mathématique de France, Paris, 2009)
8. N. Bournaveas, B. Perthame, Averages over spheres for kinetic transport equations; hyperbolic Sobolev spaces and Strichartz inequalities. J. Math. Pures Appl. **80**(9), 517–534 (2001)
9. C. Buet, B. Després, Asymptotic analysis of fluid models for the coupling of radiation and hydrodynamics. J. Quant. Spectroscopy Rad. Transf. **85**, 385–480 (2004)
10. S. Chandrasekhar, *Radiative Transfer* (Dover Publications, New York, 1960)
11. N. Chaudhuri, E. Feireisl, Navier-Stokes-Fourier system with Dirichlet boundary conditions. arXiv:2106.05315
12. B. Dubroca, J.-L. Feugeas, Etude théorique et numérique d'une hiérarchie de modéles aux moments pour le transfert radiatif. C. R. Acad. Sci. Paris **329**, 915–920 (1999)
13. B. Ducomet, E. Feireisl, The equations of magnetohydrodynamics: on the interaction between matter and radiation in the evolution of gaseous stars. Commun. Math. Phys. **266**, 595–629 (2006)

14. B. Ducomet, E. Feireisl, Š. Nečasová, On a model of radiation hydrodynamics. Ann. I. H. Poincaré-AN **28**, 797–812 (2011)
15. B. Ducomet, Š. Nečasová, M. Pokorný, Milan, M.-A. Rodríguez-Bellido, Derivation of the Navier-Stokes-Poisson system with radiation for an accretion disk. J. Math. Fluid Mech. **20**(2), 697–719 (2018)
16. B. Ducomet, Š. Nečasová, Low Mach number limit for a model of radiative flow, J. Evol. Eq. **14**(2), 357–385 (2014)
17. B. Ducomet, Š. Nečasová, Diffusion limits in a model of radiative flow. Annali dell Universita di Ferrara. VII Sci. Mat. **61**(1), 17–59 (2015)
18. B. Ducomet, Š. Nečasová, Singular limits in a model of radiative flow. J. Math. Fluid Mech. **17**(2), 341–380 (2015)
19. B. Dubroca, M. Seaïd, J.-L. Feugeas, A consistent approach for the coupling of radiation and hydrodynamics at low Mach number. J. Comput. Phys. **225**, 1039–1065 (2007)
20. E. Feireisl, A. Novotný, *Singular Limits in Thermodynamics of Viscous Fluids* (Birkhauser, Basel, 2009)
21. E. Feireisl, On the motion of a viscous, compressible, and heat conducting fluid. Indiana Univ. Math. J. **53**, 1707–1740 (2004)
22. E. Feireisl, A. Novotný, Navier-Stokes-Fourier system with general boundary conditions. Commun. Math. Phys. **386**(2), 975–1010 (2021)
23. Y. Feng, S. Wang, S. Kawashima, Global existence and asymptotic decay of solutions of the non-isentropic Euler-Maxwell system. Math. Models Methods Appl. Sci. **24**, 2851–2884 (2014)
24. Y-H. Feng, S. Wang, X. Li, Stability of non-constant steady-state solutions for non-isentropic Euler-Maxwell system with a temperature damping term. Math. Meth. Appl. Sci. **39**, 2514–2528 (2016)
25. F. Golse, B. Perthame, Generalized solutions of the radiative transfer equations in a singular case. Commun. Math. Phys. **106**(2), 211–239 (1986)
26. F. Golse, P.L. Lions, B. Perthame, R. Sentis, Regularity of the moments of the solution of a transport equation. J. Funct. Anal. **16**, 110–125 (1988)
27. F. Golse, B. Perthame, R. Sentis, Un résultat de compacité pour les équations de transport et application au calcul de la limite de la valeur propre principale d'un opérateur de transport. C. R. Acad. Sci. Paris **301**, 341–344 (1985)
28. J.W. Jerome, The Cauchy problem for compressible hydrodynamic-Maxwell systems: a local theory for smooth solutions. Differ. Int. Equa. **16**, 1345–1368 (2003)
29. P. Jiang, D. Wang, Formation of singularities of solutions of the radiative transfer equations in a singular case. Nonlinearity **23**(4), 809–821 (2010)
30. P. Jiang, D. Wang, Global weak solutions to the Euler-Boltzmann equations in radiation hydrodynamics. Quart. Appl. Math. **70**(1), 25–44 (2012)
31. S. Kawashima, Systems of a hyperbolic-parabolic composite type, with applications to the equations of magnetohydrodynamics. Doctoral Thesis, Kyoto University, 1984
32. O. Kreml, Š. Nečasová, M. Pokorný, On the steady equations for compressible radiative gas. Z. Angew. Math. Phys. **64**(3), 539–571 (2013)
33. C. Lin, T. Goudon, Global existence of the equilibrium diffusion model in radiative hydrodynamics. Chin. Ann. Math. **32B**, 549–568 (2011)
34. C. Lin, Mathematical analysis of radiative transfer models, PhD Thesis, 2007
35. C. Lin, J.F. Coulombel, T. Goudon, Shock profiles for non-equilibrium radiative gases. Physica D **218**, 83–94 (2006)
36. R.B. Lowrie, J.E. Morel, J.A. Hittinger, The coupling of radiation and hydrodynamics. Astrophys. J. **521**, 432–450 (1999)
37. D. Levermore, Relating Eddington factors to flux limiters. J. Quant. Spectrosc. Rad. Transf. **31**, 149–160 (1984)
38. D. Levermore, Moment closure hierarchies for kinetic theories, J. Stat. Phys. **83**, 1021–1076 (1996)
39. B. Mihalas, *Stellar Atmospheres* (W.H. Freeman and Cie, San Francisco, 1978)

40. B. Mihalas, B. Weibel-Mihalas, *Foundations of Radiation Hydrodynamics* (Dover Publications, Dover, 1984)
41. A. Munier, R. Weaver, Radiation transfer in the fluid frame: a covariant formulation Part I: Radiation hydrodynamics. Comput. Phys. Rep. **3**, 125–164 (1986)
42. A. Munier, R. Weaver. Radiation transfer in the fluid frame: a covariant formulation Part II: Radiation transfer equation, Computer Phys. Rep. **3**, 165–208 (1986)
43. A. Novotný, Lecture notes on the Navier-Stokes-Fourier system: weak solutions, relative entropy inequality, weak strong uniqueness, in *Topics on Compressible Navier-Stokes Equations*. Panor. Synthèses, vol. 50 (Soc. Math. France, Paris, 2016), pp. 1–42
44. E. Feireisl, Concepts of solutions in the thermodynamics of compressible fluids, in *Handbook of Mathematical Analysis in Mechanics of Viscous Fluids* (Springer, Cham, 2018), pp. 1353–1379
45. G.C. Pomraning, *Radiation Hydrodynamics* (Dover Publications, New York, 2005)
46. L. Poul, On dynamics of fluids in astrophysics. J. Evol. Equ. **9**, 37–66 (2009)
47. J.F. Ripoll, B. Dubroca, G. Duffa, Modelling radiative mean absorption coefficients. Combust. Theory Modell. **5**, 261–274 (2001)
48. D. Serre, Systèmes de lois de conservation I, II. Diderot Editeur (Arts et Sciences, Paris, New-York, Amsterdam, 1996)
49. D. Serre, Systems of conservation laws with dissipation, in *Lecture Notes SISSA* (2007)
50. Y. Shizuta, S. Kawashima, Systems of equation of hyperbolic-parabolic type with application to the discrete Boltzmann equation. Hokkaido Math. J. **14**, 249–275 (1985)
51. Z. Tan, Y. Wang, Large time behavior of solutions to the compressible non-isentropic Euler-Maxwell system in $\mathbb{R}^3$. Nonlinear Anal. Real World Appl. **15**, 187–204 (2014)
52. I. Teleaga, M. Seaïd, I. Gasser, A. Klar, J. Struckmeier, Radiation models for thermal flows at low Mach number. J. Comput. Phys. **215**, 506–525 (2006)
53. Y. Ueda, S. Wang, S. Kawashima, Dissipative structure of the regularity-loss type and time asymptotic decay of solutions for the Euler-Maxwell system. SIAM J. Math. Anal. **44**, 2002–2017 (2012)
54. Y. Ueda, S. Kawashima, Decay properties of regularity-loss type for the Euler-Maxwell system. Methods Appl. Anal. **18**, 245–268 (2011)
55. J. Xu, J. Xiong, Global existence of classical solutions of full Euler-Maxwell equations. J. Math. Anal. Appl. **402**, 545–557 (2013)
56. X.Zhong, J. Jiang, Local existence and finite-time blow up in multidimensional radiation hydrodynamics. J. Math.Fluid Mech. **9**, 543–564 (2007)

Poro-Visco-Elasticity in Biomechanics: Optimal Control

Lorena Bociu and Sarah Strikwerda

1 Introduction

Poro-visco-elastic models provide a good representation of the mechanical characteristics of biological tissues and consequently have many important applications in biology, medicine, and bioengineering. Examples include fluid flow inside cartilages, bones, and engineered tissue scaffolds [17, 23, 24, 31, 37, 40], and blood flow through tissues in the human body, like the brain, the liver, and the eye [15, 29, 36, 41]. Moreover, poro-visco-elastic models have been widely used to analyze in vitro creep and stress relaxation experiments on the tissues under confined and unconfined compression tests [37]. The models couple the visco-elastic solid matrix with the interstitial fluid flow through the permeability tensor.

The study presented in this paper is pertinent to the tissue perfusion in the eye and its relationship to the development of glaucoma [15]. Glaucoma is a group of eye diseases that leads to damage to the optic nerve (through retinal ganglion cell loss) and ultimately, vision loss. It is strongly believed that the biomechanics of the Lamina Cribrosa (LC) inside the optic nerve head plays an important role in the development and progression of glaucoma [15, 21, 25]. The LC is a thin, porous tissue at the base of the optic nerve head in the eye, formed by a multilayered network of collagen fibers that insert into the scleral canal wall. It allows passage of the central retinal vessels, and the retinal ganglion cell axons, which relay visual information from the retina to the brain. One of the main functions of the LC is to stabilize the pressure difference between the intraocular pressure (IOP) in the intraocular space and the retrolaminar tissue pressure (RLTp) in the optic nerve canal.

L. Bociu (✉) · S. Strikwerda
North Carolina State University, Department of Mathematics, Raleigh, NC, USA
e-mail: lvbociu@ncsu.edu; slstrikw@ncsu.edu

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_5

Many studies have indicated that chronical IOP elevation induces significant structural changes in the LC [6, 13, 18–20]. To date, elevated IOP is the only treatable risk factor for glaucoma, but there is significant evidence that other factors might be involved in the disease. In fact, many individuals with elevated IOP never develop glaucoma [22], while many patients continue to progress to blindness despite IOP within target levels [32]. Hence, to further understand the pathogenesis of glaucoma, improve diagnosis, and enable novel means for preventing or treating glaucoma, it is of interest to understand the effects of IOP, RLTp, and blood pressure on the optic nerve head, and, in particular, on the biomechanics-hemodynamics of the LC. *The LC can be modeled as a fluid-solid mixture problem*, and it belongs in the category of fluid flowing through a poro-visco-elastic material. Most biological tissues are composed by both elastin and collagen, and, therefore, the deformable matrix through which the fluid flows exhibits both elastic and visco-elastic behaviors. As material properties and volume fractions of elastic and collagen vary in age, health and disease, their influence on the physical system is a crucial part in the investigation of these biological fluid-mixture problems.

Deformable Porous Medium We consider a poro-visco-elastic Biot model which, under the assumptions of full saturation, negligible inertia, small deformations, and incompressible mixture components, is described by the following PDE system holding in $\Omega \times (0, T)$:

$$\frac{\partial \zeta}{\partial t} + \nabla \cdot \mathbf{v} = S, \qquad \nabla \cdot \mathbf{T} + \mathbf{F} = \mathbf{0}, \tag{1a}$$

$$\zeta = \nabla \cdot \mathbf{u}, \qquad \mathbf{v} = -\mathbf{K}\nabla p, \tag{1b}$$

$$\mathbf{T} = \mathbf{T}_e + \delta \mathbf{T}_v - p\mathbf{I}, \quad \mathbf{T}_e = 2\mu_e \epsilon(\mathbf{u}) + \lambda_e \mathrm{tr}(\epsilon(\mathbf{u}))\mathbf{I}, \quad \mathbf{T}_v = 2\mu_v \epsilon(\mathbf{u}_t) + \lambda_v \mathrm{tr}(\epsilon(\mathbf{u}_t))\mathbf{I}, \tag{1c}$$

where ζ is the fluid content, $\mathbf{v}$ is the discharge (or Darcy) velocity, p is the Darcy pressure, $\mathbf{u}$ is the solid displacement, $\mathbf{T}$ is the total stress tensor, $\epsilon(\mathbf{w}) = (\nabla \mathbf{w} + \nabla \mathbf{w}^T)/2$ is the linearized strain tensor, and S and $\mathbf{F}$ are given functions of time and space [10, 12, 30].

- Equation (1a) express the balance of mass and linear momentum, whereas Eqs. (1b)–(1c) are the constitutive equations that are necessary to close the system.
- In the general Biot model, the fluid content is given by $\zeta = c_0 p + \alpha \nabla \cdot \mathbf{u}$, c_0 being the constrained specific storage coefficient and α the Biot-Willis coefficient. In the case of incompressible mixture components, as often assumed in biological tissue modeling, we have that $c_0 = 0$ and $\alpha = 1$ [16], and, therefore, the fluid content equals the solid dilation, i.e., $\zeta = \nabla \cdot \mathbf{u}$, see Eq. (1b).
- The discharge velocity $\mathbf{v}$ and the Darcy pressure p are related via the permeability tensor $\mathbf{K} = k(x, t)\mathbf{I}$, representing the fluid and pore properties within the porous medium, see Eq. (1b). We assume that there exists constants κ_* and κ^* such that $0 < \kappa_* \leq k(x, t) \leq \kappa^* < \infty \ \forall (x, t) \in \Omega \times [0, T]$.

- In the constitutive equation for the total stress tensor $\mathbf{T}$ given in Eq. (1c), the behavior of the solid component within the medium is characterized by the Lamé elastic parameters λ_e and μ_e and visco-elastic parameters λ_v and μ_v as in [10]. The extent to which the model includes visco-elastic effects is represented by the parameter $\delta > 0$.
- Boundary conditions: We write the boundary of Ω as $\Gamma = \Gamma_D \cup \Gamma_N$, with $\Gamma_D = \Gamma_{D,v} \cup \Gamma_{D,p}$. For $\mathbf{g}$ and ψ given functions of space and time, we assume

$$\mathbf{T}\,\mathbf{n} = \mathbf{g} \;\text{ and }\; \mathbf{v}\cdot\mathbf{n} = 0 \;\text{ on }\; \Gamma_N \times (0, T), \tag{2}$$

$$\mathbf{u} = \mathbf{0} \;\text{ and }\; p = 0 \;\text{ on }\; \Gamma_{D,p} \times (0, T),\;\; \mathbf{u} = \mathbf{0} \;\text{ and }\; \mathbf{v}\cdot\mathbf{n} = \psi \;\text{ on }\; \Gamma_{D,v} \times (0, T), \tag{3}$$

  where $\mathbf{n}$ is the outward unit normal vector to the surface boundary. Note that (i) Γ_N is assumed to be an impermeable surface where the total stress is prescribed. In the eye study, we have $\mathbf{g} = -\text{IOP}\mathbf{n}$ on the surface of the LC facing the intraocular space; and (ii) $\Gamma_{D,p}$ and $\Gamma_{D,v}$ are permeable clamped surfaces, on which either the pressure or the normal velocity are prescribed. The intersection $\bar{\Gamma}_D \cap \bar{\Gamma}_N$ could be potentially non-empty.
- Initial conditions:

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}),\; \text{for } \mathbf{x} \in \Omega, \quad \text{with } \;\mathbf{u}_0(\mathbf{x})\; \text{ given. We also define } \; d_0(\mathbf{x}) = \nabla \cdot \mathbf{u}_0(\mathbf{x}). \tag{4}$$

Well-Posedness Analysis Mathematically, the study of poroelasticity ($\delta = 0$) was initiated by the 1D work of Terzaghi in the 1920s [42] and the groundbreaking consolidation theory developed by Biot in the 1940–1950s [7]. Subsequently, the poroelastic system has inspired many theoretical investigations due to its applications in geophysics and petroleum engineering, as well as medicine and bioengineering. Most of the work has been done under the assumption of *constant permeability* ($\mathbf{K} = k\mathbf{I}$, with k = constant), which yields a linear coupled system [1, 27, 35, 45]. A *monotone, nonlinear permeability* depending on pressure is studied in [39] by means of semigroup theory for implicit evolution [33, 34]. The study presented in [14] is the first to consider a *nonlinear Biot* model with permeability depending nonlinearly on dilation, i.e., $k(\nabla\cdot\mathbf{u})$. The analysis is performed *in the case of null Dirichlet boundary conditions* for both pressure and elastic displacement, and with *compressible constituents* (i.e., $c_0 > 0$). The strategy in [14] relies on Rothe's method, uses the simplified structure of a pressure-to-dilation operator introduced in [35], and takes advantage of compressibility and full elliptic regularity for the solid displacement. In [10] the authors provide the first result for *existence of weak solutions* for nonlinear poroelastic ($\delta = 0$) and poro-visco-elastic ($\delta > 0$) models with incompressible constituents and with permeability depending nonlinearly on solid dilation, and no simplifying assumptions on the domain boundary (i.e., the case when $\overline{\Gamma}_D \cap \overline{\Gamma}_N \neq \emptyset$ is included) and the associated boundary conditions (i.e., incorporating mixed boundary conditions). This work extends all previous results on analysis of poroelastic models [14, 27, 35, 39, 45]. The theoretical results in

[10] are complemented with *numerical simulations* based on a novel dual mixed hybridized finite element discretization. When the data are sufficiently regular, the simulations show that the solutions satisfy the energy estimates predicted by the theoretical analysis. Interestingly, the simulations also show that, in the purely elastic case ($\delta = 0$), the Darcy velocity and the related fluid energy might become unbounded if the data do not enjoy the time regularity required by the theory. *Relevance in the case of the lamina cribrosa (LC)*: [10] identified the presence of visco-elasticity in the solid phase as a major determinant in the behavior of the solutions, suggesting that the lack of visco-elasticity may increase the susceptibility of the tissue to localized damage as volumetric sources of linear momentum and/or boundary sources of traction experience sudden changes in time. Sudden changes in IOP (which acts as a boundary datum in the system) physiologically occur even with changes between day and night. Therefore, the novel hypothesis in [10] on the etiology of ischemic damage in the LC tissue is that [even these *physiological* changes in IOP might induce *pathological* changes in the hemodynamics of the LC tissue if the visco-elasticity provided by the collagen fibers is absent]. The finding is supported by experimental and clinical evidence showing that loss of visco-elastic tissue property has been associated with various pathological conditions, including glaucoma, osteoporosis, atherosclerosis, and Alzheimer's disease. The new hypothesis for tissue microstructural damage introduced in [10] was further investigated in [44]. The response of deformable porous media with incompressible constituents to external applied loads (either step or trapezoidal pulses) and the role that the structural viscosity plays in this response were analyzed for the 1D prototype. The analysis in [44] showed that *the fluid velocity within the medium could increase tremendously, should the external applied load experience sudden changes in time and the structural visco-elasticity be too small.* Moreover, [44] provided a dimensional analysis of the system and identified some dimensionless parameters that could be used in the design of structural properties and experimental conditions in order to maintain the fluid velocity within the medium below a desired threshold and implicitly prevent potential damage to the tissue. A similar analysis focused on poro-visco-elastic systems with compressible coefficients is provided in [9].

Sensitivity Analysis Sensitivity analysis is the first step towards optimization and control problems associated with these fluid-solid mixture models. Numerical sensitivity analysis on the 1D poro-visco-elastic models from [44] with respect to boundary data, which are the main drivers of the system, was performed in [2, 3], using the complex-step method [4, 5]. The authors compared the results obtained in the elastic case vs. visco-elastic case, as it is known that structural viscosity of biological tissues decreases with age and disease. *Key observations:* (1) Solution $(\mathbf{u}, p, \mathbf{v})$ is more sensitive to boundary traction $\mathbf{g}$ in the elastic case ($\delta = 0$) than in the visco-elastic scenario ($\delta > 0$). This could explain why in the theoretical results provided in [10], the boundary source was required to have higher time regularity in order to obtain solution $(\mathbf{u}, p) \in L^2(L^2)$ in space and time, and with appropriate energy estimate in terms of data, in the purely elastic case ($\delta = 0$); (2) Effects of the boundary source $\mathbf{g}$ are most significant for the discharge velocity $\mathbf{v}$, especially in the purely elastic case. This agrees with the numerical investigation in [10] which hinted that the fluid energy (dependent on the discharge velocity) becomes unbounded as the boundary source of traction loses H^1-smoothness in time, and visco-elasticity is no longer present; (3) The sensitivity of the solution with respect to the boundary source ψ was computed as well. Interestingly, the fluid-dynamical variables $\mathbf{v}$ and p appear to be more sensitive to changes in $\mathbf{g}$ than to changes in ψ. This suggests that, in order to control fluid velocity and pressure, it would be much more effective to act on the boundary conditions for the solid structure, namely the traction $\mathbf{g}$, rather than on the boundary conditions for the velocity itself, namely ψ; (4) Solid displacement $\mathbf{u}$ is the least sensitive to changes in $\mathbf{g}$ and ψ. This finding shows that small changes in the solid displacement may actually correspond to big changes in fluid velocity and pressure, thereby suggesting that monitoring the sole solid displacement might not be indicative of the fluid-dynamical state inside the medium.

In addition, in [11] the authors performed sensitivity analysis on the 1D dimensionless poroelastic ($\delta = 0$) and poro-visco-elastic ($\delta > 0$) solutions with respect to the boundary traction and their dependence on the dimensionless parameters identified in [44]. The results in [11] consistently show that the maximum magnitude of the sensitivities (for all three variables: solid displacement, fluid pressure, and fluid velocity) is largest when visco-elasticity is not present and gets smaller monotonically with respect to the dimensionless parameter η associated with visco-elasticity (but also dependent on the value of the permeability and the length of the domain). However, the magnitudes of the sensitivities were not always monotonic with respect to η. The numerical results in [11] correspond closely with results typically observed in creep tests for poro-visco-elastic materials. The lag in the solid displacement due to the inclusion of structural viscosity is indicative of the increase in time that it takes for the solid to reach an equilibrium after an applied load.

Optimal Control Problems In this paper we address optimal control problems subject to fluid flows through deformable, porous media described by (1a)–(4), using the well-posedness and sensitivity analysis results described above. In particular, we focus on the case of given permeability $k(x, t)$, which translates into a convex control problem, with both distributed and boundary controls. We focus on investigating the problem of maintaining the solid displacement and Darcy pressure close to desired values (motivated by the application below) using the sources present in the system as control variables. The results provided in this paper include existence and uniqueness of optimal control, as well as the characterization of the optimal control through the first order necessary optimality conditions, based on the adjoint system. *Relevance to Applications:* So far, we know that there are two factors that contribute to the onset and progression of glaucoma: Intraocular Pressure (IOP) and LC tissue perfusion. However, up to date, it is not clear whether the biomechanical pinching of the RGC axons or the hemodynamic deficiencies in the blood vessels within the LC are the major contributors to the microstructural damage in the LC leading to the development of glaucoma. Thus our goal in this article is to study the relation between the IOP and LC tissue perfusion via an optimal control problem for the poro-visco-elastic model describing the LC. For the control variables, we use the interior and boundary data in the system which are the main drivers of the equations (recall that the Neumann boundary data represents the IOP). Variables of interest are the structural displacement **u** within the LC, which is indicative of the strains experienced by the RGC axons, and the blood flow pressure p in the LC, which is indicative of tissue perfusion.

The remaining sections of the paper are organized as follows: In Sect. 2, we discuss the well-posedness analysis of the poro-visco-elastic problem. Section 3 sets up the optimal control problem under consideration and provides the results on existence and uniqueness of both distributed and boundary controls. In Sect. 4 we present necessary optimality conditions based on the adjoint system, which is found using the formal Lagrangian method [43].

2 Poro-Visco-Elasticity: Well-posedness Analysis

As usual, $H^s(D)$ represents the Sobolev space of order s defined on a domain D, while $H_0^s(D)$ is the closure of $C_0^\infty(D)$ in the $H^s(D)$ norm, denoted by $\|\cdot\|_{H^s(D)}$. We use $\mathbf{H}^s(D)$ to denote $(H^s(D))^3$ and $\mathbf{L}^2(D)$ to denote $(L^2(D))^3$. Unless otherwise specified, $\|\cdot\|$ and $(\cdot,\cdot)$ denote the norm and inner product, respectively, taken in $(L^2(D))^n$ where n is clear by the context. Additionally, $\langle\cdot,\cdot\rangle$ will be used to denote the $L^2(\partial D)$ inner product on a portion of the boundary which will be denoted in a subscript, e.g., $\langle\mathbf{u},\mathbf{w}\rangle_{\Gamma_N}$.

The primary spaces in our analysis are

$$\mathbf{V} \equiv (H^1_{\Gamma_D}(\Omega))^3, \qquad V \equiv H^1_{\Gamma_{D,p}}(\Omega), \qquad \mathbb{V} \equiv \mathbf{V} \times V,$$

for displacement **u** and the pressure p, respectively. Note that the functional spaces are of the form $H^1_{\Gamma_*}(\Omega) = \{f \in H^1(\Omega) \ : \ \gamma[f]\Big|_{\Gamma_*} = 0\}$. For sake of exposition we take the Lamé parameters in the elasticity term normalized to unity (the analysis follows similarly if the Lamé parameters were not normalized to unity), and we define the bilinear form associated with the elasticity operator by

$$a(\mathbf{u},\mathbf{w}) = (\nabla\cdot\mathbf{u},\nabla\cdot\mathbf{w}) + (\nabla\mathbf{u} : \nabla\mathbf{w}) + (\nabla\mathbf{u} : (\nabla\mathbf{w})^T), \tag{5}$$

where $\nabla\mathbf{u}$ stands for the Jacobian matrix of **u** and the Frobenius inner product of two matrices is given by

$$(\mathbf{A} : \mathbf{B}) = \int_\Omega (A_{ij}B_{ij})d\Omega.$$

We note that the bilinear form $a(\cdot,\cdot)$ defines an inner product on $\mathbf{V}$, due to Assumption 1 on the domain described below. The inner product for V is inherited from $H^1(\Omega)$, and the norm on $\mathbb{V}$ is given by $\|(\mathbf{w},w)\| = \sqrt{\|\mathbf{w}\|^2_{\mathbf{V}} + \|w\|^2_V}$.

We define a weak solution as follows:

Definition 1 (Weak Solution) A weak solution to (1)–(4) is represented by the pair of functions $\mathbf{u} \in H^1(0,T;\mathbf{V})$ and $p \in L^2(0,T;V)$ such that:

(a) For any $\mathbf{w} \in L^2(0,T;\mathbf{V})$ and $w \in L^2(0,T;V)$, the following variational formulations are satisfied:

$$\int_0^T \delta a(\mathbf{u}_t,\mathbf{w}) + a(\mathbf{u},\mathbf{w}) - (p,\nabla\cdot\mathbf{w})\,dt = \int_0^T \langle \mathbf{g},\mathbf{w}\rangle_{\Gamma_N} + (\mathbf{F},\mathbf{w})\,dt \quad (6)$$

$$\int_0^T (k(x,t)\nabla p,\nabla w) + (\nabla\cdot\mathbf{u}_t, w)\,dt = -\int_0^T \langle \psi, w\rangle_{\Gamma_{D,\mathbf{v}}} + (S,w)\,dt. \quad (7)$$

(b) For every $w \in V$, the term $(\nabla\cdot\mathbf{u}(t), w)$ uniquely defines an absolutely continuous function on $[0,T]$ and the initial condition $(\nabla\cdot\mathbf{u}(0), w) = (\nabla\cdot\mathbf{u}_0, w)$ is satisfied.

Definition 2 (Energy and Data) Energy functionals for solutions and data are defined as follows:

$$E(\mathbf{u}(t)) \equiv \frac{1}{2}a(u,u) = \frac{1}{2}\left[||\nabla\cdot\mathbf{u}(t)||^2 + ||\nabla\mathbf{u}||^2 + (\nabla\mathbf{u} : \nabla\mathbf{u}^T)\right],$$

$$E(p(t)) \equiv (k(x,t)\nabla p,\nabla p) \geq k_*\|\nabla p\|^2,$$

$$\mathcal{F}(T) \equiv \int_0^T \left[||\mathbf{g}(t)||^2_{\mathbf{L}^2(\Gamma_N)} + ||\psi(t)||^2_{L^2(\Gamma_{D,\mathbf{v}})} + ||S(t)||^2_{L^2(\Omega)} + ||\mathbf{F}(t)||^2_{\mathbf{L}^2(\Omega)}\right]dt. \quad (8)$$

Assumption 1 (Assumptions on the Domain) *We assume that Γ_D and $\Gamma_{D,p}$ are sets of positive measure. Thus Korn's Inequality for $\mathbf{u} \in \mathbf{V}$ and Poincaré's Inequality for $p \in V$ are satisfied:*

$$E(\mathbf{u}(t)) \geq c||\mathbf{u}(t)||^2_{\mathbf{H}^1(\Omega)}, \qquad ||p||_{L^2(\Omega)} \leq C_P||\nabla p||_{\mathbf{L}^2(\Omega)}.$$

Theorem 1 (Existence and Uniqueness of Weak Solution) *Consider (1a)–(4) with $\mathbf{u}_0 \in \mathbf{V}$ and under Assumption 1. Let the data have the following regularity:*

$$\mathbf{F} \in L^2\left(0,T;\mathbf{L}^2(\Omega)\right),\ S \in L^2\left(0,T;L^2(\Omega)\right),$$
$$\mathbf{g} \in L^2\left(0,T;\mathbf{L}^2(\Gamma_N)\right),\ \psi \in L^2\left(0,T;L^2(\Gamma_{D,\mathbf{v}})\right). \quad (9)$$

Then there exists a unique weak solution $(\mathbf{u},p)$ satisfying the following energy estimate:

$$\sup_{t\in[0,T]} E(\mathbf{u}(t)) + \int_0^T \Big[E(p(t)) + E(\mathbf{u}(t)) + \delta E(\mathbf{u}_t(t))\Big] dt \le C(T)\Big[E(\mathbf{u}(0)) + \mathcal{F}(T)\Big]. \tag{10}$$

Proof In [10] one can find a proof for existence of weak solutions in the case of nonlinear permeability (dependent on solid dilation) under stricter assumptions on the domain (related to obtaining enough elliptic regularity for the displacement **u** in order to use compactness arguments in the passing with the limit in the nonlinear term). The proof is based on Rothe's method (discretizations in time and space) and could be invoked here. However, we provide a simplified, straightforward proof for existence and uniqueness in the case of given permeability as a function of time and space, following [8]. Our proof is based on Lax–Milgram theorem and formal energy estimates.

Solving the Discretized in Time Problem We partition the time interval $[0, T]$ into $r \ge 1$ subintervals of length $\Delta t = T/r$. The intervals are given by $(t_{i-1}, t_i]$, where $t_i = i\Delta t$, for $0 \le i \le r$. We define

$$\mathbf{F}^i := \frac{1}{\Delta t}\int_{t_{i-1}}^{t_i} \mathbf{F}(x,t)dt, \quad S^i := \frac{1}{\Delta t}\int_{t_{i-1}}^{t_i} S(x,t)dt, \quad \mathbf{g}^i := \frac{1}{\Delta t}\int_{t_{i-1}}^{t_i} \mathbf{g}(x,t)dt,$$

$$\psi^i := \frac{1}{\Delta t}\int_{t_{i-1}}^{t_i} \psi(x,t)dt \quad \text{and } k^i := \frac{1}{\Delta t}\int_{t_{i-1}}^{t_i} k(x,t)dt.$$

We seek solutions $(\mathbf{u}^i, p^i) \in \mathbb{V}$ for $1 \le i \le r$ which satisfies the weak formulation for the semi-discrete problem:

$$\delta a(\mathbf{u}^i, \mathbf{w}) + \Delta t a(\mathbf{u}^i, \mathbf{w}) - \Delta t(p^i, \nabla\cdot\mathbf{w}) = \Delta t\langle \mathbf{g}^i, \mathbf{w}\rangle_{\Gamma_N} + \Delta t(\mathbf{F}^i, \mathbf{w}) + \delta a(\mathbf{u}^{i-1}, \mathbf{w}) \tag{11}$$

$$[\Delta t]^2(k^i(x)\nabla p^i, \nabla w) + \Delta t(\nabla\cdot\mathbf{u}^i, w) = -[\Delta t]^2\langle\psi^i, w\rangle_{\Gamma_{D,p}} + [\Delta t]^2(S^i, w) + \Delta t(\nabla\cdot\mathbf{u}^{i-1}, w) \tag{12}$$

for all $\mathbf{w} \in \mathbf{V}$ and $w \in V$. We find a solution inductively, by initially setting $\mathbf{u}^0$ equal to the given initial condition $\mathbf{u}_0$.

We define the bilinear form $\mathscr{F} : \mathbb{V}\times\mathbb{V} \to \mathbb{R}$ as follows:

$$\begin{aligned}\mathscr{F}\Big((\mathbf{u}^i, p^i), (\mathbf{w}, w)\Big) :=&(\delta + \Delta t)a(\mathbf{u}^i, \mathbf{w}) - \Delta t(p^i, \nabla\cdot\mathbf{w})\\ &+ [\Delta t]^2(k^i(x)\nabla p^i, \nabla w) + \Delta t(\nabla\cdot\mathbf{u}^i, w).\end{aligned} \tag{13}$$

Given data $\mathbf{F}^i \in \mathbf{L}^2(\Omega)$, $\mathbf{g}^i \in \mathbf{L}^2(\Gamma_N)$, $S^i \in L^2(\Omega)$, $\psi^i \in L^2(\Gamma_{D,v})$, and solution $\mathbf{u}^{i-1} \in \mathbf{V}$ at time t^{i-1}, we define the linear functional $\mathcal{G} : \mathbb{V} \to \mathbb{R}$ as follows:

$$\begin{aligned}\mathcal{G}(\mathbf{w}, w) :=&\Delta t\langle\mathbf{g}^i, \mathbf{w}\rangle_{\Gamma_N} + \Delta t(\mathbf{F}^i, \mathbf{w}) + \delta a(\mathbf{u}^{i-1}, \mathbf{w}) - [\Delta t]^2\langle\psi^i, w\rangle_{\Gamma_{D,p}} + [\Delta t]^2(S^i, w)\\ &+ \Delta t(\nabla\cdot\mathbf{u}^{i-1}, w).\end{aligned}$$

Using these linear operators, we can write (11) and (12) equivalently as: find a solution $(\mathbf{u}^i, p^i) \in \mathbb{V}$ such that for every $(\mathbf{w}, w) \in \mathbb{V}$, we have

$$\mathscr{F}((\mathbf{u}^i, p^i), (\mathbf{w}, w)) = \mathcal{G}(\mathbf{w}, w). \tag{14}$$

The bilinear form $\mathscr{F}$ is continuous on $\mathbb{V}$. This can be seen by applying Cauchy–Schwarz inequality and the upper-bound assumption on $k(x, t)$:

$$\begin{aligned}
&\mathscr{F}\left((\mathbf{u}^i, p^i), (\mathbf{w}, w)\right) \\
&\le (\delta + \Delta t)\|\mathbf{u}^i\|_{\mathbf{V}}\|\mathbf{w}\|_{\mathbf{V}} + \Delta t\|p^i\|_V\|\mathbf{w}\|_{\mathbf{V}} + [\Delta t]^2\kappa^*\|p^i\|_V\|w\|_V + \Delta t\|\mathbf{u}^i\|_{\mathbf{V}}\|w\|_V \\
&= C_{\Delta t}\left(\sqrt{\|\mathbf{u}^i\|^2_{\mathbf{V}}\|\mathbf{w}\|^2_{\mathbf{V}}} + \sqrt{\|p^i\|^2_V\|\mathbf{w}\|^2_{\mathbf{V}}} + \sqrt{\|p^i\|^2\|w\|^2} + \sqrt{\|\mathbf{u}^i\|^2_{\mathbf{V}}\|w\|^2_V}\right) \\
&\le 2C_{\Delta t}\left(\sqrt{\|\mathbf{u}^i\|^2_{\mathbf{V}}\|\mathbf{w}\|^2_{\mathbf{V}} + \|p^i\|^2_V\|\mathbf{w}\|^2_{\mathbf{V}} + \|p^i\|^2_V\|w\|^2_V + \|\mathbf{u}^i\|^2_{\mathbf{V}}\|w\|^2_V}\right) \\
&= 2C_{\Delta t}\left\|(\mathbf{u}^i, p^i)\right\|_{\mathbb{V}}\|(\mathbf{w}, w)\|_{\mathbb{V}}\,.
\end{aligned} \tag{15}$$

Additionally, using the lower bound on $k(x, t)$ and Assumption 1, we obtain that

$$\begin{aligned}
&\left(\mathscr{F}\begin{bmatrix}\mathbf{u}^i \\ p^i\end{bmatrix}, \begin{bmatrix}\mathbf{u}^i \\ p^i\end{bmatrix}\right) \\
&= (\delta + \Delta t)a(\mathbf{u}^i, \mathbf{u}^i) - \Delta t(p^i, \nabla\cdot\mathbf{u}^i) + [\Delta t]^2(k^i(x)\nabla p^i, \nabla p^i) + \Delta t(\nabla\cdot\mathbf{u}^i, p^i) \\
&\ge C_{\Delta t}(\|\mathbf{u}^i\|^2_{\mathbf{V}} + \|p^i\|^2_V),
\end{aligned} \tag{16}$$

which implies that $\mathscr{F}$ is coercive on $\mathbb{V}$.

Similarly, we have that $\mathcal{G}$ is continuous on $\mathbb{V}$ (this can be seen using Cauchy–Schwarz and Trace Theorem). Hence, $\mathcal{G} \in \mathbb{V}'$. Therefore, application of Lax–Milgram Theorem provides existence of a unique solution $(\mathbf{u}^i, p^i) \in \mathbb{V}$ to (14), which consequently satisfies (11) and (12).

Note that the continuity and coercivity coefficients are dependent on Δt in a singular way. Therefore, we only invoke Lax–Milgram Theorem to show the existence of a semi-discrete solution, for each value of Δt, and then we derive a uniform bound using energy estimates.

Deriving Uniform Bounds on the Discretized Solution $(\mathbf{u}^i, p^i)$ We define

$$\mathbf{u}^{[r]} = \mathbf{u}^i \quad \text{in} \quad (t_{i-1}, t_i], \quad i = 1, \ldots, r \tag{17}$$

$$p^{[r]} = p^i \quad \text{in} \quad (t_{i-1}, t_i], \quad i = 1, \ldots, r. \tag{18}$$

$$(\mathbf{u}^{[r]})_{\Delta t} := \frac{\mathbf{u}^i - \mathbf{u}^{i-1}}{\Delta t} \quad \text{in} \quad (t_{i-1}, t_i] \quad i = 1, \ldots, r. \tag{19}$$

Let $\mathbf{u}^i - \mathbf{u}^{i-1}$ be the test function in (11) and let p^i be the test function in (12). Adding these two equations together, and summing from $i = 1$ to l, we have

$$\begin{aligned}
&(\delta + \Delta t)\sum_{i=1}^{l} a(\mathbf{u}^i, \mathbf{u}^i - \mathbf{u}^{i-1}) - \Delta t \sum_{i=1}^{l}(p^i, \nabla\cdot\mathbf{u}^i - \nabla\cdot\mathbf{u}^{i-1}) + [\Delta t]^2\sum_{i=1}^{l}(k^i(x)\nabla p^i, \nabla p^i)\\
&+ \Delta t\sum_{i=1}^{l}(\nabla\cdot\mathbf{u}^i - \nabla\cdot\mathbf{u}^{i-1}, p^i) = \Delta t\sum_{i=1}^{l}\langle\mathbf{g}^i, \mathbf{u}^i - \mathbf{u}^{i-1}\rangle_{\Gamma_N} + \Delta t\sum_{i=1}^{l}(\mathbf{F}^i, \mathbf{u}^i - \mathbf{u}^{i-1})\\
&+ \delta\sum_{i=1}^{l} a(\mathbf{u}^{i-1}, \mathbf{u}^i - \mathbf{u}^{i-1}) - [\Delta t]^2\sum_{i=1}^{l}\langle\psi^i, p^i\rangle + [\Delta t]^2\sum_{i=1}^{l}(S^i, p^i).
\end{aligned} \tag{20}$$

Using the definition of $\mathbf{g}^i$, Cauchy–Schwarz Inequality, Trace Theorem, and Young's Inequality, we have the following estimate:

$$\begin{aligned}
\Delta t\sum_{i=1}^{l}\langle\mathbf{g}^i, \mathbf{u}^i - \mathbf{u}^{i-1}\rangle_{\Gamma_N} &= \Delta t\sum_{i=1}^{l}\left\langle\frac{1}{\Delta t}\int_{t_{i-1}}^{t_i}\mathbf{g}(x,t)dt, \mathbf{u}^i - \mathbf{u}^{i-1}\right\rangle_{\Gamma_N}\\
&\le \sum_{i=1}^{l}\left\|\int_{t_{i-1}}^{t_i}\mathbf{g}(x,t)\right\|_{\mathbf{L}^2(\Gamma_N)}\|\mathbf{u}^i - \mathbf{u}^{i-1}\|_{\mathbf{L}^2(\Gamma_N)}\\
&\le C[\Delta t]^{1/2}\sum_{i=1}^{l}\|\mathbf{g}\|_{L^2((t_{i-1},t_i);\mathbf{L}^2(\Gamma_N))}\|\mathbf{u}^i - \mathbf{u}^{i-1}\|_{\mathbf{V}}\\
&\le \frac{C}{2\epsilon}\Delta t\|\mathbf{g}\|^2_{L^2(0,T;\mathbf{L}^2(\Gamma_N))} + \frac{\epsilon}{2}\sum_{i=1}^{l}\|\mathbf{u}^i - \mathbf{u}^{i-1}\|^2_{\mathbf{V}}.
\end{aligned} \tag{21}$$

Similarly, we have

$$\begin{aligned}
[\Delta t]^2\sum_{i=1}^{l}(S^i, p^i) &\le C[\Delta t]^{3/2}\sum_{i=1}^{l}\|S\|_{L^2((t_{i-1},t_i);L^2(\Omega))}\|p^i\|_V\\
&\le \frac{C}{2\epsilon}\Delta t\|S\|^2_{L^2(0,T;L^2(\Omega))} + \frac{\epsilon}{2}\sum_{i=1}^{l}[\Delta t]^2\|p^i\|^2_V.
\end{aligned} \tag{22}$$

Applying (21), (22), and similar estimates on $\mathbf{F}$ and ψ, we obtain

$$\begin{aligned}&\delta\sum_{i=1}^{l}a(\mathbf{u}^i-\mathbf{u}^{i-1},\mathbf{u}^i-\mathbf{u}^{i-1})+\Delta t\sum_{i=1}^{l}a(\mathbf{u}^i,\mathbf{u}^i-\mathbf{u}^{i-1})+[\Delta t]^2\kappa_*\sum_{i=1}^{l}(\nabla p^i,\nabla p^i)\\&\leq\frac{C\Delta t}{2\epsilon}\left(\|\mathbf{g}\|^2_{L^2(0,T;\mathbf{L}^2(\Gamma_N))}+\|\mathbf{F}\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))}\right)+\frac{C\Delta t}{2\epsilon}\Big(\|S\|^2_{L^2(0,T;L^2(\Omega))}\\&+\|\psi\|^2_{L^2(0,T;L^2(\Gamma_{D,v}))}\Big)+\frac{\epsilon}{2}\sum_{i=1}^{l}\|\mathbf{u}^i-\mathbf{u}^{i-1}\|_{\mathbf{V}}+\frac{\epsilon}{2}[\Delta t]^2\sum_{i=1}^{l}\|\nabla p^i\|^2_{L^2(\Omega)}.\end{aligned}\tag{23}$$

Using the following identity

$$a(\mathbf{w}^i,\mathbf{w}^i-\mathbf{w}^{i-1})=\frac{1}{2}a(\mathbf{w}^i,\mathbf{w}^i)-\frac{1}{2}a(\mathbf{w}^{i-1},\mathbf{w}^{i-1})+\frac{1}{2}a(\mathbf{w}^i-\mathbf{w}^{i-1},\mathbf{w}^i-\mathbf{w}^{i-1}),\tag{24}$$

we obtain that

$$\Delta t\sum_{i=1}^{l}a(\mathbf{u}^i,\mathbf{u}^i-\mathbf{u}^{i-1})=\frac{\Delta t}{2}a(\mathbf{u}^l,\mathbf{u}^l)-\frac{\Delta t}{2}a(\mathbf{u}^0,\mathbf{u}^0)+\Delta t\sum_{i=1}^{l}a(\mathbf{u}^i-\mathbf{u}^{i-1},\mathbf{u}^i-\mathbf{u}^{i-1}).\tag{25}$$

Letting $\epsilon<\min\{\delta,\kappa_*\}$ and applying (25) to (23), we have that

$$\begin{aligned}&\sum_{i=1}^{l}\|\mathbf{u}^i-\mathbf{u}^{i-1}\|^2_{\mathbf{V}}+\frac{\Delta t}{2}\|\bar{\mathbf{u}}^l\|^2_{\mathbf{V}}+\sum_{i=1}^{l}\frac{\Delta t}{2}\|\mathbf{u}^i-\mathbf{u}^{i-1}\|^2_{\mathbf{V}}+[\Delta t]^2\sum_{i=1}^{l}(\nabla p^i,\nabla p^i)\\&\leq C(\delta,\kappa_*)\Delta t\mathcal{F}(T)+\frac{\Delta t\|\bar{\mathbf{u}}^0\|^2_{\mathbf{V}}}{2}.\end{aligned}\tag{26}$$

Letting $l=r$, and dividing by Δt we obtain

$$\Delta t\frac{1}{[\Delta t]^2}\sum_{i=1}^{r}\|\mathbf{u}^i-\mathbf{u}^{i-1}\|^2_{\mathbf{V}}\leq C(\delta,\kappa^*)(\mathcal{F}(T)+\|\mathbf{u}_0\|^2_{\mathbf{V}}).\tag{27}$$

Identifying the sum as an integral, we obtain the following estimate:

$$\begin{aligned}&\int_0^T\|(\mathbf{u}^{[r]})_{\Delta t}\|^2_{\mathbf{V}}dt\\&\leq C(\delta,\kappa_*)\left(\int_0^T(\|\mathbf{g}\|^2_{\mathbf{L}^2(\Gamma_N)}+\|\mathbf{F}\|^2_{\mathbf{L}^2(\Omega)}+\|\psi\|^2_{L^2(\Gamma_{D,v})}+\|S\|^2_{L^2(\Omega)})dt+\|\mathbf{u}_0\|^2_{\mathbf{V}}\right).\end{aligned}\tag{28}$$

Similarly, using Poincarè's inequality in (26), we obtain that

$$\int_0^T \|p^{[r]}\|_V^2 dt$$
$$\leq C(\delta,\kappa_*)\left(\int_0^T (\|\mathbf{g}\|^2_{\mathbf{L}^2(\Gamma_N)} + \|\mathbf{F}\|^2_{\mathbf{L}^2(\Omega)} + \|\psi\|^2_{L^2(\Gamma_{D,v})} + \|S\|^2_{L^2(\Omega)})dt + \|\mathbf{u}_0\|^2_{\mathbf{V}}\right). \tag{29}$$

Finally, from (26), we also have that

$$\max_{0\leq t\leq T} \|\mathbf{u}^{[r]}(t)\|^2_{\mathbf{V}} \leq C(\delta,\kappa_*)(\mathcal{F}(T) + \|\mathbf{u}_0\|^2_{\mathbf{V}}), \tag{30}$$

which provides the following estimate:

$$\int_0^T \|\mathbf{u}^{[r]}\|^2_{\mathbf{V}} dt$$
$$\leq C(\delta,\kappa_*,T)\left(\int_0^T (\|\mathbf{g}\|^2_{\mathbf{L}^2(\Gamma_N)} + \|\mathbf{F}\|^2_{\mathbf{L}^2(\Omega)} + \|\psi\|^2_{L^2(\Gamma_{D,v})} + \|S\|^2_{L^2(\Omega)})dt + \|\mathbf{u}_0\|^2_{\mathbf{V}}\right). \tag{31}$$

Passing with the Limit From (28), (29), and (31), we have that there exists weakly convergent subsequences of $\{(\mathbf{u}^{[r]})_{\Delta t}\}$, $\{\mathbf{u}^{[r]}\}$ in $L^2(0,T;\mathbf{V})$ and of $\{p^{[r]}\}$ in $L^2(0,T;V)$. We denote the weak limits as follows:

$$\lim_{r\to\infty} (\mathbf{u}^{[r]})_{\Delta t} \rightharpoonup \mathbf{u}^\sharp, \quad \lim_{r\to\infty} \mathbf{u}^{[r]} \rightharpoonup \mathbf{u}, \quad \lim_{r\to\infty} p^{[r]} \rightharpoonup p.$$

Now following [45], we let $f \in C^\infty([0,T])$. We define

$$f^i := f(t_i), \;\; i = 1,\ldots,r, \quad f^{r+1} := f^r = f(T),$$

$$f^{[r]} := f^i \quad \text{in} \quad (t_{i-1},t_i], \;\; (f^{[r]})_{\Delta t} := \frac{f^{i+1} - f^i}{\Delta t} \quad \text{in} \quad (t_{i-1},t_i], \quad i = 1,\ldots,r.$$

By Taylor's theorem, we have for all $f \in C^\infty([0,T])$,

$$\|f^{[r]} - f\|_{L^2(0,T)} \leq C[\Delta t] \quad \text{and} \quad \|(f^{[r]})^+_{\Delta t} - f'\|_{L^2(0,T)} \leq C[\Delta t]. \tag{32}$$

Multiplying (11) and (12) by f^i and summing from $i = 1$ to r, we obtain

$$\delta\Delta t\sum_{i=1}^r a\left(\frac{\mathbf{u}^i - \mathbf{u}^{i-1}}{\Delta t}, \mathbf{w}\right) f^i + \Delta t\sum_{i=1}^r a(\mathbf{u}^i,\mathbf{w}) f^i - \Delta t\sum_{i=1}^r (p^i, \nabla\cdot\mathbf{w}) f^i$$
$$= \Delta t\sum_{i=1}^r \langle \mathbf{g}^i, \mathbf{w}\rangle_{\Gamma_N} f^i + \Delta t\sum_{i=1}^r (\mathbf{F}^i,\mathbf{w}) f^i \tag{33}$$

$$
\begin{aligned}
&\Delta t \sum_{i=1}^{r} (k^i \nabla p^i, w) f^i + \Delta t \sum_{i=1}^{r} \left(\frac{\nabla \cdot \mathbf{u}^i - \nabla \cdot \mathbf{u}^{i-1}}{\Delta t}, w \right) f^i \\
&= -\Delta t \sum_{i=1}^{r} \langle \psi^i, w \rangle_{\Gamma_{D,p}} f^i + \Delta t \sum_{i=1}^{r} (S^i, w) f^i .
\end{aligned}
\tag{34}
$$

We now identify these sums as integrals. Due to the fact that p^i and f^i are constant on each interval (t_{i-1}, t_i) for $1 \le i \le r$ and using Bochner's Theorem, we have that

$$
\begin{aligned}
\Delta t \sum_{i=1}^{r} (k^i \nabla p^i, \nabla w) f^i &= \sum_{i=1}^{r} \left(\int_{t_{i-1}}^{t_i} k(x,t) dt \nabla p^i, \nabla w \right) f^i \\
&= \sum_{i=1}^{r} \left(\int_{t_{i-1}}^{t_i} k(x,t) \nabla p^i dt, \nabla w \right) f^i \\
&= \sum_{i=1}^{r} \int_{t_{i-1}}^{t_i} (k(x,t) \nabla p^i, \nabla w) f^i dt = \int_0^T (k(x,t) \nabla p^{[r]}, \nabla w) f^{[r]} dt .
\end{aligned}
\tag{35}
$$

The other terms in (33) and (34) can similarly be written as integrals. Therefore, we have

$$
\begin{aligned}
&\delta \int_0^T a((\mathbf{u}^{[r]})_{\Delta t}, \mathbf{w}) f^{[r]} dt + \int_0^T a(\mathbf{u}^{[r]}, \mathbf{w}) f^{[r]} dt - \int_0^T (p^{[r]}, \nabla \cdot \mathbf{w}) f^{[r]} dt \\
&= \int_0^T \langle \mathbf{g}, \mathbf{w} \rangle_{\Gamma_N} f^{[r]} dt + \int_0^T (\mathbf{F}, \mathbf{w}) f^{[r]} dt
\end{aligned}
\tag{36}
$$

$$
\begin{aligned}
&\int_0^T (k(x,t) \nabla p^{[r]}, \nabla w) f^{[r]} dt + \int_0^T (\nabla \cdot (\mathbf{u}^{[r]})_{\Delta t}, w) f^{[r]} dt \\
&= - \int_0^T \langle \psi, w \rangle_{\Gamma_{D,v}} f^{[r]} dt + \int_0^T (S, w) f^{[r]} dt .
\end{aligned}
\tag{37}
$$

We will now take the limit as $r \to \infty$ of (36) and (37). First we observe that

$$
\begin{aligned}
\int_0^T a(\mathbf{u}^{[r]}, \mathbf{w}) f^{[r]} dt =& \left\{ \int_0^T a(\mathbf{u}^{[r]}, \mathbf{w}) f^{[r]} dt - \int_0^T a(\mathbf{u}, \mathbf{w}) f^{[r]} dt \right\} \\
&+ \left\{ \int_0^T a(\mathbf{u}, \mathbf{w}) f^{[r]} dt - \int_0^T a(\mathbf{u}, \mathbf{w}) f dt \right\} + \int_0^T a(\mathbf{u}, \mathbf{w}) f dt .
\end{aligned}
\tag{38}
$$

Let $\mathcal{H}_1 : L^2(0,T) \to \mathbb{R}$ and $\mathcal{H}_2 : L^2(0,T; \mathbf{V}) \to \mathbb{R}$ be defined, respectively, as

$$\mathcal{H}_1(f) := \int_0^T a(\mathbf{m}, \mathbf{w}) f dt \quad \text{for fixed } \mathbf{m}, \mathbf{w} \in L^2(0, T; \mathbf{V})$$

$$\mathcal{H}_2(\mathbf{v}) := \int_0^T a(\mathbf{v}, \mathbf{w}) f dt \quad \text{for fixed } \mathbf{w} \in L^2(0, T; \mathbf{V}),\ \ f \in L^2(0, T).$$

Then the continuity of both the linear functionals follows from Cauchy–Schwarz. Therefore,

$$\lim_{r\to\infty} \int_0^T a(\mathbf{m}^{[r]}, \mathbf{w}) f^{[r]} dt - \int_0^T a(\mathbf{m}, \mathbf{w}) f^{[r]} dt = 0 \quad \text{and}$$

$$\lim_{r\to\infty} \int_0^T a(\mathbf{m}, \mathbf{w}) f^{[r]} dt - \int_0^T a(\mathbf{m}, \mathbf{w}) f dt = 0.$$

Combined with (38), this implies

$$\lim_{r\to\infty} \int_0^T a(\mathbf{u}^{[r]}, \mathbf{w}) f^{[r]} dt = \int_0^T a(\mathbf{u}, \mathbf{w}) f dt \quad \text{and} \tag{39}$$

$$\lim_{r\to\infty} \int_0^T a((\mathbf{u}^{[r]})_{\Delta t}, \mathbf{w}) f^{[r]} = \int_0^T a(\mathbf{u}^{\sharp}, \mathbf{w}) f dt. \tag{40}$$

Similarly, passing the limit as $r \to \infty$ on each term in (36) and (37) yields

$$\int_0^T \delta a(\mathbf{u}^{\sharp}, \mathbf{w}) f + a(\mathbf{u}, \mathbf{w}) f - (p, \nabla \cdot \mathbf{w}) f \; dt = \int_0^T \langle \mathbf{g}, \mathbf{w} \rangle_{\Gamma_N} f + (\mathbf{F}, \mathbf{w}) f \; dt \tag{41}$$

$$\int_0^T (k(x,t)\nabla p, \nabla p) f + (\nabla \cdot \mathbf{u}^{\sharp}, w) f \; dt = \int_0^T -\langle \psi, w \rangle_{\Gamma_{D,v}} f + (S, w) f \; dt. \tag{42}$$

Finally, we show that $\mathbf{u}^{\sharp} = \mathbf{u}_t$. Let $f \in C^\infty([0,T])$. Then using (32), (40), and (39), we have

$$\begin{aligned}\int_0^T a(\mathbf{u}^{\sharp}, \mathbf{w}) f dt &= \lim_{r\to\infty} \int_0^T a((\mathbf{u}^{[r]})_{\Delta t}, \mathbf{w}) f^{[r]} dt = \lim_{r\to\infty} \Delta t \sum_{i=1}^{r} a\left(\frac{\mathbf{u}^i - \mathbf{u}^{i-1}}{\Delta t}, \mathbf{w}\right) f^i \\ &= \lim_{r\to\infty} \left[a(\mathbf{u}^r, \mathbf{w}) f^r - a(\mathbf{u}^0, \mathbf{w}) f^1 - \sum_{i=1}^{r-1} a(\mathbf{u}^i, \mathbf{w})(f^{i+1} - f^i) \right] \\ &= a(\mathbf{u}(T), \mathbf{w}) f(T) - a(\mathbf{u}^0, \mathbf{w}) f(0) - \int_0^T a(\mathbf{u}, \mathbf{w}) f' dt.\end{aligned} \tag{43}$$

If we further restrict $f \in C_0^\infty(0,T)$, then $\int_0^T a(\mathbf{u}^\sharp, \mathbf{w}) f dt = \int_0^T a(\mathbf{u}, \mathbf{w}) f dt$ is true for all $\mathbf{w} \in \mathbf{V}$ and $f \in C_0^\infty(0,T)$. Since $\{\mathbf{w} f(t) | \mathbf{w} \in \mathbf{V},\ f \in C_0^\infty(0,T)\}$ is dense in $L^2(0,T;\mathbf{V})$, we have $\mathbf{u}^\sharp = \mathbf{u}_t \in L^2(0,T;\mathbf{V})$. Therefore, $\mathbf{u}$ and p satisfy (6) and (7).

Recovering the Initial Condition We must show $\mathbf{u}$ satisfies the initial condition. For any $\mathbf{w} \in \mathbf{V}$, we define

$$G(t) := \delta a(\mathbf{u}(t), \mathbf{w}) \tag{44}$$

$$H(t) := -(a(\mathbf{u}(t), \mathbf{w}) + (p(t), \nabla \cdot \mathbf{w}) + \langle \mathbf{g}(t), \mathbf{w} \rangle_{\Gamma_N} + (\mathbf{F}(t), \mathbf{w}) \tag{45}$$

$$F(t) := \int_0^t H(\tau) d\tau. \tag{46}$$

Note that $F(t)$ is absolutely continuous on $[0,T]$ and satisfies $F'(t) = H(t)$ $a.e.$ $(0,T)$. Since $\mathbf{u}$ and p satisfy (7), we have that

$$\delta \int_0^T a(\mathbf{u}_t, \mathbf{w}) f + a(\mathbf{u}, \mathbf{w}) f - (p, \nabla \cdot \mathbf{w}) f \ dt = \int_0^T \langle \mathbf{g}, \mathbf{w} \rangle_{\Gamma_N} f + (\mathbf{F}, \mathbf{w}) f \ dt \tag{47}$$

for all $f \in C^\infty([0,T])$. Using the above definitions, we rewrite this equation as the following:

$$\int_0^T (G'(t) - F'(t)) f(t) dt = 0 \quad \forall f \in C_0^\infty(0,T). \tag{48}$$

Therefore,

$$G - F = c. \tag{49}$$

We now consider $f \in C^\infty([0,T])$ such that

$$f(0) = 1 \quad \text{and} \quad f(T) = 1. \tag{50}$$

Integrating (47) by parts in time and applying (43), we obtain

$$\begin{aligned} &- \delta \int_0^T a(\mathbf{u}, \mathbf{w}) f'(t) dt - \delta a(\mathbf{u}^0, \mathbf{w}) + \int_0^T a(\mathbf{u}, \mathbf{w}) f(t) dt - \int_0^T (p, \nabla \cdot \mathbf{w}) f(t) dt \\ &= \int_0^T \langle \mathbf{g}, \mathbf{w} \rangle_{\Gamma_N} f(t) dt + \int_0^T (\mathbf{F}, \mathbf{w}) f(t) dt. \end{aligned} \tag{51}$$

This can be rewritten as

$$-\int_0^T G(t)f'(t) - \delta a(\mathbf{u}^0, \mathbf{w}) = \int_0^T H(t)f(t)dt. \tag{52}$$

Using integration by parts, (49), and (50), we obtain

$$F(0) + c + \int_0^T F'(t)f(t)dt - \int_0^T H(t)f(t)dt = \delta a(\mathbf{u}^0, \mathbf{w}). \tag{53}$$

Recalling that $F'(t) = H(t)$ a.e., we have that $F(0) + c = \delta a(\mathbf{u}^0, \mathbf{w})$. Looking at the definition of F, it is easy to see that $F(0) = 0$. Therefore, $c = \delta(\mathbf{u}^0, \mathbf{w})$. From (49), we then obtain that

$$\delta a(\mathbf{u}(t), \mathbf{w}) + \int_0^t (a(\mathbf{u}(t), \mathbf{w}) + (p(t), \nabla\cdot\mathbf{w}) + \langle \mathbf{g}(t), \mathbf{w}\rangle_{\Gamma_N} + (\mathbf{F}(t), \mathbf{w})dt = \delta a(\mathbf{u}^0, \mathbf{w}). \tag{54}$$

The time integral vanishes when we set $t = 0$, and thus we have that

$$\delta a(\mathbf{u}(0), \mathbf{w}) = \delta a(\mathbf{u}^0, \mathbf{w}). \tag{55}$$

Since this holds for all $\mathbf{w} \in \mathbf{V}$, then $\mathbf{u}(0) = \mathbf{u}^0$.

Uniqueness of Weak Solution Assume $(\mathbf{u}, p)$ satisfies Definition 1. Using $\mathbf{u}_t$ and p as test functions in (6) and (7), respectively, and integrating from 0 to τ, we have

$$\begin{aligned} &\delta \int_0^\tau a(\mathbf{u}_t, \mathbf{u}_t)dt + \int_0^\tau a(\mathbf{u}, \mathbf{u}_t)dt + \int_0^\tau (k(x,t)\nabla p, \nabla p)dt \\ &= \int_0^\tau \langle \mathbf{g}, \mathbf{u}_t\rangle_{\Gamma_N} dt + \int_0^\tau (\mathbf{F}, \mathbf{u}_t)dt - \int_0^\tau \langle \psi, p\rangle_{\Gamma_{D,v}} dt + \int_0^\tau (S, p)dt \ \Rightarrow \end{aligned} \tag{56}$$

$$\begin{aligned} &\delta \int_0^\tau a(\mathbf{u}_t, \mathbf{u}_t)dt + \frac{1}{2}a(\mathbf{u}(\tau), \mathbf{u}(\tau)) + \int_0^\tau (k(x,t)\nabla p, \nabla p)dt \\ &\leq a(\mathbf{u}(0), \mathbf{u}(0)) + \int_0^\tau \langle \mathbf{g}, \mathbf{u}_t\rangle_{\Gamma_N} dt + \int_0^\tau (\mathbf{F}, \mathbf{u}_t)dt - \int_0^\tau \langle \psi, p\rangle_{\Gamma_{D,v}} dt + \int_0^\tau (S, p)dt. \end{aligned} \tag{57}$$

Therefore, using the fact that $\kappa_* \leq k(x,t) \leq \kappa^*$, Cauchy–Schwarz, and Young's Inequality, we have

$$\begin{aligned} &\delta \|\mathbf{u}_t\|^2_{L^2(0,\tau;\mathbf{V})} + \|\mathbf{u}(\tau)\|^2_{\mathbf{V}} + \kappa_* \|p\|^2_{L^2(0,\tau;V)} \leq C\Big(\|\mathbf{u}_0\|^2_{\mathbf{V}} + \frac{1}{2\epsilon}\|\mathbf{g}\|^2_{L^2(0,\tau;L^2(\Gamma_N))} \\ &+ \epsilon \|\mathbf{u}_t\|^2_{L^2(0,\tau;L^2(\Gamma_N))} + \frac{1}{2\epsilon}\|\mathbf{F}\|^2_{L^2(0,\tau;\mathbf{L}^2(\Omega))} + \frac{1}{2\epsilon}\|\boldsymbol{\psi}\|^2_{L^2(0,\tau;L^2(\Gamma_{D,v}))} \\ &+ \epsilon \|p\|^2_{L^2(0,\tau;L^2(\Gamma_{D,v}))} + \frac{1}{2\epsilon}\|S\|^2_{L^2(0,\tau;L^2(\Omega))}\Big). \end{aligned} \tag{58}$$

Then applying Trace Theorem, setting ϵ sufficiently small, and letting $\tau = T$, we obtain

$$\|\mathbf{u}_t\|^2_{L^2(0,T;\mathbf{V})} + \|p\|^2_{L^2(0,T;V)} \leq C\left(\|\mathbf{u}_0\|^2_{\mathbf{V}} + \mathcal{F}(T)\right), \tag{59}$$

where $\mathcal{F}(T)$ is as given in (8). Additionally, by integrating (58) from 0 to T, we see

$$\int_0^T \|\mathbf{u}(t)\|^2_{\mathbf{V}} dt \leq \int_0^T C\left(\|\mathbf{u}_0\|^2_{\mathbf{V}} + \mathcal{F}(T)\right) dt \tag{60}$$

which implies

$$\|\mathbf{u}\|^2_{L^2(0,T;\mathbf{V})} \leq C(T)\left(\|\mathbf{u}_0\|^2_{\mathbf{V}} + \mathcal{F}(T)\right). \tag{61}$$

Putting (58) and (61) together shows that any solution that satisfies Definition 1 satisfies (10). If the solution was not unique than the difference of solutions would satisfy the system with initial condition and sources set to zero. However, in this case, (10) would imply the difference of the solutions had norm zero. Therefore, the solution must be unique. □

3 Optimal Control Problems: Well-Posedness

Our goal is to “optimize" the solution of the poro-visco-elastic system (1a)–(4) represented by the solid displacement $\mathbf{u}$ and fluid pressure p using either distributed or boundary controls. More specifically, we want to find controls so that their corresponding solid displacement and fluid pressure are the best possible approximations to desired/target values (dictated by the applications described in the Introduction), denoted by $\mathbf{u}_d$ and p_d, respectively.

The sources $\mathbf{F}$, S, $\mathbf{g}$, or ψ will be used as controls. For sake of exposition, we denote the control with a generic q, to represent the different choices of control variables. The regularity of the control q, dictated by the well-posedness theory presented in Sect. 2, is given by

$$Q = L^2(0,T;\mathbf{L}^2(\Omega)),\ L^2(0,T;L^2(\Omega)),\ L^2(0,T;\mathbf{L}^2(\Gamma_N)),\ \text{or } L^2\left(0,T;L^2(\Gamma_{D,\mathbf{v}})\right),$$

depending on if $\mathbf{F}$, S, $\mathbf{g}$, or ψ is the control, respectively. Notation wise, we denote the inner product in the control space as $(\cdot,\cdot)_Q$ and the norm in the control space as $\|\cdot\|_Q$.

We denote by $\mathbf{u}[q]$ and $p[q]$ the solid displacement and fluid pressure corresponding to control q in (1a)–(4). Then the optimal control problem under consideration is

$$\min_{q \in Q_{ad}} J(\mathbf{u}[q], p[q], q), \tag{62}$$

where the cost functional has the following form:

$$J(\mathbf{u}, p, q) = \frac{1}{2}\|\mathbf{u} - \mathbf{u}_d\|^2_{L^2(0,T;\mathbf{L}^2(\Omega))} + \frac{1}{2}\|p - p_d\|^2_{L^2(0,T;L^2(\Omega))} + \frac{\lambda}{2}\|q\|^2_Q,$$

and $Q_{ad} := \{q \in Q | q_l \leq q \leq q_u\}$ is the set of admissible controls. Note that q_l could potentially be $-\infty$ and q_u could potentially be ∞.

Definition 3 [Control-to-State Operator] Let $G : Q_{ad} \to L^2(0, T; \mathbb{V})$ be the control-to-state operator that maps q to $y = (\mathbf{u}, p)$ where $(\mathbf{u}, p)$ is the weak solution to (1)–(4) with the initial condition, boundary and interior sources all set to zero except the control.

Using Theorem 1, we have that the control-to-state operator G is linear, continuous, and injective.

The optimal control problem (62) can now be written equivalently as

$$\min_{q \in Q_{ad}} J(q), \quad \text{with} \quad J(q) = \frac{1}{2}\|I \circ G(q) + y_0 - y_d\|^2_Y + \frac{\lambda}{2}\|q\|^2_Q,$$

where I is the continuous embedding $L^2(0, T; \mathbb{V}) \to Y = L^2(0, T; \mathbf{L}^2(\Omega) \times L^2(\Omega))$, $y_d = (\mathbf{u}_d, p_d)$, and y_0 is the weak solution to (1)–(4) with the control set to zero and the initial condition, boundary and interior sources set as desired. For the sake of exposition, we let $\tilde{y} = y_d - y_0$ and define $(\tilde{\mathbf{u}}, \tilde{p}) = \tilde{y}$. Then the optimal control problem becomes

$$\min_{q \in Q_{ad}} J(q), \quad \text{with} \quad J(q) = \frac{1}{2}\|I \circ G(q) - \tilde{y}\|^2_Y + \frac{\lambda}{2}\|q\|^2_Q. \tag{63}$$

The main result of this section, along with the proof, is presented below:

Theorem 2 *Assume that either* $[\lambda > 0]$ *or* $[\lambda = 0$ *and* $q_l \neq -\infty$, *and* $q_u \neq \infty]$.

1. *There exists a unique optimal distributed control* $\bar{\mathbf{F}} \in L^2(0, T; \mathbf{L}^2(\Omega))$ *that solves the minimization problem (63) subject to (1)–(4) with* $S \in L^2(0, T; L^2(\Omega))$, $\mathbf{g} \in L^2(0, T; \mathbf{L}^2(\Gamma_N))$, $\psi \in L^2(0, T; L^2(\Gamma_{D,\mathbf{v}}))$, *and initial condition* $\mathbf{u}_0 \in \mathbf{V}$.
2. *There exists a unique optimal distributed control* $\bar{S} \in L^2(0, T; L^2(\Omega))$ *that solves the minimization problem (63) subject to (1)–(4) with* $\mathbf{F} \in L^2(0, T; \mathbf{L}^2(\Omega))$, $\mathbf{g} \in L^2(0, T; \mathbf{L}^2(\Gamma_N))$, $\psi \in L^2(0, T; L^2(\Gamma_{D,\mathbf{v}}))$, *and initial condition* $\mathbf{u}_0 \in \mathbf{V}$.
3. *There exists a unique optimal boundary control* $\bar{\mathbf{g}} \in L^2(0, T; \mathbf{L}^2(\Gamma_N))$ *that solves the minimization problem (63) subject to (1)–(4) with* $\mathbf{F} \in L^2(0, T; \mathbf{L}^2(\Omega))$, $S \in L^2(0, T; L^2(\Omega))$, $\psi \in L^2(0, T; L^2(\Gamma_{D,\mathbf{v}}))$, *and initial condition* $\mathbf{u}_0 \in \mathbf{V}$.

4. *There exists a unique optimal boundary control* $\bar{\psi} \in L^2(0,T;L^2(\Gamma_{D,\mathbf{v}}))$ *that solves the minimization problem (63) subject to (1)–(4) with* $\mathbf{F} \in L^2(0,T;\mathbf{L}^2(\Omega))$, $S \in L^2(0,T;L^2(\Omega))$, $\mathbf{g} \in L^2(0,T;\mathbf{L}^2(\Gamma_N))$, *and initial condition* $\mathbf{u}_0 \in \mathbf{V}$.

Proof Using Theorem 1, we see that the control-to-state operator G is linear, continuous, and injective. As a consequence, the cost functional $J(q)$ is continuous and strictly convex, for all choices of controls q. We focus on the case when $q = \psi$ is the control (Part 4). We omit the proofs for the other parts, as they follow similarly.

If $\lambda = 0$, then the cost functional is minimized over a bounded set, as $q_l \neq -\infty$, and $q_u \neq \infty$. We denote the bounded set by Q_b. Note that the OC problem (63) can be reduced to a minimization problem over a bounded set if $\lambda > 0$, as well. We have the following argument. Let $\psi_0 \in L^2(0,T;L^2(\Gamma_{D,\mathbf{v}}))$. If $\|\psi\|^2 > \frac{2}{\lambda}J(\psi_0)$, then

$$J(\psi) = \frac{1}{2}\|G(\psi) - \tilde{y}\|_Y^2 + \frac{\lambda}{2}\|\psi\|^2_{L^2(0,T;L^2(\Gamma_{D,\mathbf{v}}))} > J(\psi_0). \tag{64}$$

Thus the search for an optimum can be restricted to the bounded set $Q_b := \{\psi \in L^2(0,T;L^2(\Omega)) \mid \|\psi\|^2 \leq \frac{1}{\lambda}\|\tilde{y}\|_Y^2\}$ if $0 \in Q_{ad}$ and a similar set otherwise.

Now we can show that a solution exists for (63). Since $J(q) \geq 0$, there exists $j := \inf_{q \in Q_b} J(q)$. As a consequence, there exists a sequence $\{q_n\}_{n=1}^{\infty} \subset Q_b$ such that $J(q_n) \to j$ as $n \to \infty$. Since Q is reflexive, and Q_b is closed, bounded, and convex, then Q_b is weakly sequentially compact. Therefore, there exists a subsequence of $\{q_n\}_{n=1}^{\infty}$, which for convenience we still denote by $\{q_n\}_{n=1}^{\infty}$, that converges weakly to some $\bar{q} \in Q_b$. The cost functional J is weakly lower semicontinuous and thus

$$\liminf_{n\to\infty} J(q_n) \geq J(\bar{q}) \Rightarrow j \geq J(\bar{q}).$$

Using the definition of j, we, therefore, obtain that $J(\bar{q}) = j$, which means that $\bar{q}$ is the optimal control.

To prove uniqueness of optimal control, we assume that there are two solutions $\bar{q}$ and $\bar{r}$ for (63). By the strict convexity of J, we have

$$J(\mu\bar{q} + (1-\mu)\bar{r}) < \mu J(\bar{q}) + (1-\mu)J(\bar{r}) = j. \tag{65}$$

Given the convexity of Q_b, $\mu\bar{q} + (1-\mu)\bar{r} \in Q_b$, so (65) implies that $\bar{q}$ and $\bar{r}$ are not optimal controls. This is a contradiction, and thus the optimal control is unique. □

4 Necessary Optimality Condition

We have that the cost functional $J : Q_{ad} \to \mathbb{R}$ is Frechét differentiable and its derivative is given by

$$J'(q) = G^*(G(q) - \tilde{y}) + \lambda q,$$

where G^* is the adjoint operator of G. Thus we can immediately characterize the optimal control $\bar{q}$ as follows:

Lemma 1 *Consider the minimization problem (63). Then $\bar{q} \in Q_{ad}$ is a solution for (63) if and only if $\bar{q}$ satisfies the following inequality:*

$$(G^*(G\bar{q} - \tilde{y}), q - \bar{q})_Q + (\lambda\bar{q}, q - \bar{q})_Q \geq 0 \quad \forall q \in Q_{ad}. \tag{66}$$

Since the adjoint operator of G can be difficult to work with computationally, we will characterize the optimal control using the adjoint system.

4.1 Adjoint System

We define, formally, the Lagrangian function

$$\begin{aligned}
&\mathcal{L}(\mathbf{u}, p, q, \hat{\mathbf{m}}, \hat{h}) = J(\mathbf{u}, p, q) + \int_0^T (\nabla \cdot \mathbf{T}(\mathbf{u}, p), \mathbf{m}_1)dt + \int_0^T (\mathbf{F}, \mathbf{m}_1)dt \\
&- \int_0^T (\nabla \cdot \mathbf{u}_t, h_1)dt + \int_0^T (\nabla \cdot (k(x,t)\nabla p), h_1)dt + \int_0^T (S, h_1)dt - (\nabla \cdot \mathbf{u}(0), h_2) \\
&+ (d_0, h_2) - \int_0^T \langle \mathbf{T}(\mathbf{u}, p)\mathbf{n}, \mathbf{m}_2\rangle_{\Gamma_N} dt + \int_0^T \langle \mathbf{g}, \mathbf{m}_2\rangle_{\Gamma_N} dt - \int_0^T \langle \mathbf{u}, \mathbf{m}_3\rangle_{\Gamma_D} dt \\
&- \int_0^T \langle k(x,t)\nabla p \cdot \mathbf{n}, h_3\rangle_{\Gamma_N} dt - \int_0^T \langle p, h_4\rangle_{\Gamma_{D,p}} dt - \int_0^T \langle k(x,t)\nabla p \cdot \mathbf{n}, h_5\rangle_{\Gamma_{D,v}} dt \\
&- \int_0^T \langle \psi, h_5\rangle_{\Gamma_{D,v}} dt,
\end{aligned}$$

where the Lagrange multipliers $\mathbf{m}_1$, h_1 are functions defined on $\Omega \times (0, T)$, h_2 is a function defined on Ω, and $\mathbf{m}_2$, $\mathbf{m}_3$, h_3, h_4, h_5 are functions defined on parts of $\Gamma \times (0, T)$, and they are expressed in $\mathscr{L}$ as the vectors $\hat{\mathbf{m}}$ and $\hat{h}$. Moreover, we recall that the control q will be taken to be one of the sources $\mathbf{F}$, S, $\mathbf{g}$, or ψ. Assuming sufficient smoothness on $\hat{\mathbf{m}}$ and $\hat{h}$ and integrating by parts, we obtain the following equivalent form for the Lagrangian function:

$$
\begin{aligned}
&\mathcal{L}(\mathbf{u}, p, q, \hat{\mathbf{m}}, \hat{h}) = J(\mathbf{u}, p, q) + \int_0^T (\mathbf{u}, \nabla \cdot \sigma(\mathbf{m}_1))dt - \int_0^T \langle \sigma(\mathbf{m}_1)\mathbf{n}, \mathbf{u}\rangle_\Gamma dt \\
&- \delta \int_0^T (\mathbf{u}, \nabla \cdot \sigma(\mathbf{m}_1)_t)dt + \delta(\mathbf{u}(T), \nabla \cdot \sigma(\mathbf{m}_1(T))) - \delta(\mathbf{u}(0), \nabla \cdot \sigma(\mathbf{m}_1(0))) \\
&+ \delta \int_0^T \langle \sigma(\mathbf{m}_1)_t \mathbf{n}, \mathbf{u}\rangle_\Gamma dt - \delta \langle \sigma(\mathbf{m}_1)(T)\mathbf{n}, \mathbf{u}(T)\rangle_\Gamma + \delta\langle \sigma(\mathbf{m}_1(0)), \mathbf{u}(0)\rangle_\Gamma \\
&+ \int_0^T (p, \nabla \cdot \mathbf{m}_1)dt + \int_0^T \langle \sigma(\mathbf{u})\mathbf{n}, \mathbf{m}_1\rangle_\Gamma dt - \delta \int_0^T \langle \sigma(\mathbf{u})\mathbf{n}, \mathbf{m}_{1t}\rangle_\Gamma dt \\
&+ \delta\langle \sigma(\mathbf{u}(T))\mathbf{n}, \mathbf{m}_1(T)\rangle_\Gamma - \delta\langle \sigma(\mathbf{u}(0))\mathbf{n}, \mathbf{m}_1(0)\rangle_\Gamma - \int_0^T \langle p, \mathbf{m}_1 \cdot \mathbf{n}\rangle_\Gamma dt \\
&+ \int_0^T (\mathbf{F}, \mathbf{m}_1)dt - \int_0^T (\mathbf{u}, \nabla h_{1t})dt + (\mathbf{u}(T), \nabla h_1(T)) - (\mathbf{u}(0), \nabla h_1(0)) \\
&+ \int_0^T \langle \mathbf{u} \cdot \mathbf{n}, h_{1t}\rangle_\Gamma dt - \langle \mathbf{u}(T) \cdot \mathbf{n}, h_1(T)\rangle_\Gamma + \langle \mathbf{u}(0) \cdot \mathbf{n}, h_1(0)\rangle_\Gamma \\
&+ \int_0^T (p, \nabla \cdot (k(x,t)\nabla h_1)dt + \int_0^T \langle k(x,t)\nabla p \cdot \mathbf{n}, h_1\rangle dt - \int_0^T \langle p, k(x,t)\nabla h_1 \cdot \mathbf{n}\rangle_\Gamma dt \\
&+ \int_0^T (S, h_1)dt + (\mathbf{u}(0), \nabla h_2) - \langle \mathbf{u}(0) \cdot \mathbf{n}, h_2\rangle_\Gamma - \int_0^T \langle \sigma(\mathbf{u}), \mathbf{m}_2\rangle_{\Gamma_N} dt \\
&+ \delta \int_0^T \langle \sigma(\mathbf{u}), \mathbf{m}_{2t}\rangle_{\Gamma_N} dt - \delta\langle \sigma(\mathbf{u}(T)), \mathbf{m}_2(T)\rangle_{\Gamma_N} + \delta\langle \sigma(\mathbf{u}(0)), \mathbf{m}_2(0)\rangle_{\Gamma_N} \\
&+ \int_0^T \langle p, \mathbf{m}_2 \cdot \mathbf{n}\rangle_{\Gamma_N} dt + \int_0^T \langle \mathbf{g}, \mathbf{m}_2\rangle_{\Gamma_N} dt - \int_0^T \langle \mathbf{u}, \mathbf{m}_3\rangle_{\Gamma_D} dt \\
&- \int_0^T \langle k(x,t)\nabla p \cdot \mathbf{n}, h_3\rangle_{\Gamma_N} dt - \int_0^T \langle p, h_4\rangle_{\Gamma_{D,p}} dt - \int_0^T \langle k(x,t)\nabla p \cdot \mathbf{n}, h_5\rangle_{\Gamma_{D,v}} dt \\
&- \int_0^T \langle \psi, h_5\rangle_{\Gamma_{D,v}} dt.
\end{aligned} \tag{67}
$$

By Lagrange principle, the optimal solution $(\bar{\mathbf{u}}, \bar{p}, \bar{q})$ and the multipliers $\hat{\mathbf{m}}$ and $\hat{h}$ should satisfy the optimality conditions associated with the unconstrained problem "min $\mathcal{L}(\mathbf{u}, p, q, \hat{\mathbf{m}}, \hat{h})$ with $q \in Q_{ad}$".

From (71), we obtain the following formulas for the derivatives of $\mathcal{L}$ w.r.t. the state variables:

$$\begin{aligned}
&D_{\mathbf{u}}\mathcal{L}(\mathbf{u}, p, q, \hat{\mathbf{m}}, \hat{h})\mathbf{v} = \int_0^T (\mathbf{u}-\mathbf{u}_d, \mathbf{v})dt + \int_0^T (\mathbf{v}, \nabla\cdot\sigma(\mathbf{m}_1))dt \\
&- \int_0^T \langle\sigma(\mathbf{m}_1)\mathbf{n}, \mathbf{v}\rangle_\Gamma dt - \delta\int_0^T (\mathbf{v}, \nabla\cdot\sigma(\mathbf{m}_1)_t)dt + \delta(\mathbf{v}(T), \nabla\cdot\sigma(\mathbf{m}_1(T))) \\
&- \delta(\mathbf{v}(0), \nabla\cdot\sigma(\mathbf{m}_1(0))) + \delta\int_0^T \langle\sigma(\mathbf{m}_1)_t\mathbf{n}, \mathbf{v}\rangle_\Gamma dt - \delta\langle\sigma(\mathbf{m}_1)(T)\mathbf{n}, \mathbf{v}(T)\rangle_\Gamma \\
&+ \delta\langle\sigma(\mathbf{m}_1(0)), \mathbf{v}(0)\rangle_\Gamma + \int_0^T \langle\sigma(\mathbf{v})\mathbf{n}, \mathbf{m}_1\rangle_\Gamma dt - \delta\int_0^T \langle\sigma(\mathbf{v})\mathbf{n}, \mathbf{m}_{1t}\rangle_\Gamma dt \\
&+ \delta\langle\sigma(\mathbf{v}(T))\mathbf{n}, \mathbf{m}_1(T)\rangle_\Gamma - \delta\langle\sigma(\mathbf{v}(0))\mathbf{n}, \mathbf{m}_1(0)\rangle_\Gamma - \int_0^T (\mathbf{v}, \nabla h_{1t})dt + (\mathbf{v}(T), \nabla h_1(T)) \\
&- (\mathbf{v}(0), \nabla h_1(0)) + \int_0^T \langle\mathbf{v}\cdot\mathbf{n}, h_{1t}\rangle_\Gamma dt - \langle\mathbf{v}(T)\cdot\mathbf{n}, h_1(T)\rangle_\Gamma + \langle\mathbf{v}(0)\cdot\mathbf{n}, h_1(0)\rangle_\Gamma \\
&+ (\mathbf{v}(0), \nabla h_2) - \langle\mathbf{v}(0)\cdot\mathbf{n}, h_2\rangle_\Gamma - \int_0^T \langle\sigma(\mathbf{v}), \mathbf{m}_2\rangle_{\Gamma_N} dt + \delta\int_0^T \langle\sigma(\mathbf{v}), \mathbf{m}_{2t}\rangle_{\Gamma_N} dt \\
&- \delta\langle\sigma(\mathbf{v}(T)), \mathbf{m}_2(T)\rangle_{\Gamma_N} + \delta\langle\sigma(\mathbf{v}(0)), \mathbf{m}_2(0)\rangle_{\Gamma_N} - \int_0^T \langle\mathbf{v}, \mathbf{m}_3\rangle_{\Gamma_D} dt
\end{aligned} \tag{68}$$

$$\begin{aligned}
&D_p\mathcal{L}(\mathbf{u}, p, q, \hat{\mathbf{m}}, \hat{h})r = \int_0^T (p-p_d, r)dt + \int_0^T (r, \nabla\cdot\mathbf{m}_1)dt - \int_0^T \langle r, \mathbf{m}_1\cdot\mathbf{n}\rangle dt \\
&+ \int_0^T (r, \nabla\cdot(k(x,t)\nabla h_1)dt + \int_0^T \langle k(x,t)\nabla r\cdot\mathbf{n}, h_1\rangle_\Gamma dt - \int_0^T \langle r, k(x,t)\nabla h_1\cdot\mathbf{n}\rangle_\Gamma dt \\
&+ \int_0^T \langle r, \mathbf{m}_2\cdot\mathbf{n}\rangle_{\Gamma_N} dt - \int_0^T \langle k(x,t)\nabla r\cdot\mathbf{n}, h_3\rangle_{\Gamma_N} dt - \int_0^T \langle r, h_4\rangle_{\Gamma_{D,p}} dt \\
&- \int_0^T \langle k(x,t)\nabla r\cdot\mathbf{n}, h_5\rangle_{\Gamma_{D,v}} dt.
\end{aligned} \tag{69}$$

Since the derivative of $\mathcal{L}$ with respect to $(\mathbf{u}, p)$ should vanish at the optimal point $(\bar{\mathbf{u}}, \bar{p}, \bar{q})$, and if we let

$$\mathbf{m} := \mathbf{m}_1,\ h := h_1,\ \mathbf{m}_{\mathbf{2}} = \mathbf{m}|_{\Gamma_N}\ \ \mathbf{m}_3 = (-\sigma(\mathbf{m}) + \delta\sigma(\mathbf{m}_1)_t + h_t\mathbf{I})\mathbf{n})|_{\Gamma_D},$$

$$\nabla h_2 = -\delta\nabla\cdot\sigma(\mathbf{m}_1(0)) + \nabla h(0)\ \ h_3 = h|_{\Gamma_N}\ \ h_4 = -k(x,t)\nabla h\cdot\mathbf{n}|_{\Gamma_{D,p}}\ \ h_5 = h|_{\Gamma_{D,v}},$$

we obtain the following adjoint system:

$$\begin{cases} -\delta\nabla\cdot\sigma(\mathbf{m}_t)+\nabla\cdot\sigma(\mathbf{m})-\nabla h_t=-(\bar{\mathbf{u}}-\tilde{\mathbf{u}}) & \text{in } \Omega\times(0,T)\\ -\nabla\cdot\mathbf{m}-\nabla\cdot(k(x,t)\nabla h)=\bar{p}-\tilde{p} & \text{in } \Omega\times(0,T)\\ -k(x,t)\nabla h\cdot n=0 & \text{on } (\Gamma_N\cup\Gamma_{D,v})\times(0,T)\\ (\delta\sigma(\mathbf{m}_t)-\sigma(\mathbf{m})+h_t\mathbf{I})n=0 & \text{on } \Gamma_N\times(0,T)\\ h=0 & \text{on } \Gamma_{D,p}\times(0,T)\\ \mathbf{m}=0 & \text{on } \Gamma_D\times(0,T)\\ \nabla h(T)+\delta\nabla\cdot\sigma(\mathbf{m}(T))=0 & \text{in } \Omega. \end{cases} \tag{70}$$

Remark 1 The proper definition of the Lagrangian function $\mathscr{L}$ is given below. For $\mathbf{m} \in L^2(0,T;\mathbf{V})$ and $h \in L^2(0,T;V)$, let the Lagrangian function $\mathscr{L}(\mathbf{u},p,q,\mathbf{m},h) : H^1(0,T;\mathbf{V})\times L^2(0,T;V)\times Q\times L^2(0,T;\mathbf{V})\times L^2(0,T;V) \to \mathbb{R}$ be given by

$$\begin{aligned} \mathscr{L}(\mathbf{u},p,q,\mathbf{m},h) &= J(\mathbf{u},p,q)-\Big(\delta\int_0^T a(\mathbf{u}_t,\mathbf{m})\,dt+\int_0^T a(\mathbf{u},\mathbf{m})\,dt\\ &-\int_0^T(p,\nabla\cdot\mathbf{m})\,dt-\int_0^T\langle\mathbf{g},\mathbf{m}\rangle_{\Gamma_N}\,dt-\int_0^T(\mathbf{F},\mathbf{m})\,dt\Big)\\ &-\Big(\int_0^T(k(x,t)\nabla p,\nabla h)\,dt+\int_0^T(\nabla\cdot\mathbf{u}_t,h)\,dt+\int_0^T\langle\psi,h\rangle_{\Gamma_{D,\mathbf{v}}}\,dt-\int_0^T(S,h)\,dt\Big). \end{aligned} \tag{71}$$

Now we consider the adjoint system (70) with given data as below:

$$\begin{cases} -\delta\nabla\cdot\sigma(\mathbf{m}_t)+\nabla\cdot\sigma(\mathbf{m})-\nabla h_t=-\mathbf{z}_1 & \text{in } \Omega\times(0,T)\\ -\nabla\cdot\mathbf{m}-\nabla\cdot(k(x,t)\nabla h)=z_2 & \text{in } \Omega\times(0,T)\\ -k(x,t)\nabla h\cdot n=0 & \text{on } (\Gamma_N\cup\Gamma_{D,v})\times(0,T)\\ (\delta\sigma(\mathbf{m}_t)-\sigma(\mathbf{m})+h_t\mathbf{I})n=0 & \text{on } \Gamma_N\times(0,T)\\ h=0 & \text{on } \Gamma_{D,p}\times(0,T)\\ \mathbf{m}=0 & \text{on } \Gamma_D\times(0,T)\\ \nabla h(T)+\delta\nabla\cdot\sigma(\mathbf{m}(T))=\nabla z_3+\delta\nabla\cdot\sigma(\mathbf{z}_4) & \text{in } \Omega. \end{cases} \tag{72}$$

We define a weak solution for the adjoint system (72) as follows.

Definition 4 [Weak Solution for Adjoint System] A weak solution to (72) is represented by the pair of functions $\mathbf{m} \in L^2(0,T;\mathbf{V})$ and $h \in L^2(0,T;V)$ such that:

(a) For any $\mathbf{w} \in \mathbf{V}$, $w \in V$, and $f \in C_0^\infty(0,T)$ the following variational formulations are satisfied:

$$\delta \int_0^T a(\mathbf{m}, \mathbf{w}) f'(t) + a(\mathbf{m}, \mathbf{w}) f(t) + (h, \nabla \cdot \mathbf{w}) f'(t)\, dt = \int_0^T (\mathbf{z}_1, \mathbf{w}) f(t)\, dt \quad (73)$$

$$\int_0^T (k(x,t)\nabla h, \nabla w) f(t) - (\nabla \cdot \mathbf{m}, w) f(t)\, dt = \int_0^T (z_2, w) f(t)\, dt. \quad (74)$$

(b) For every $\mathbf{w} \in \mathbf{V}$, the term $(h(t), \nabla \cdot \mathbf{w}) + \delta a(\mathbf{m}(t), \mathbf{w})$ uniquely defines an absolutely continuous function on $[0, T]$ and the terminal condition $(h(T), \nabla \cdot \mathbf{w}) + \delta a(\mathbf{m}(T), \mathbf{w}) = (z_3, \nabla \cdot \mathbf{w}) + \delta a(\mathbf{z}_4, \mathbf{w})$ is satisfied.

The adjoint system (72) is a linear, weakly coupled system. Once reversed in time,

$$\begin{cases} \delta \nabla \cdot \sigma(\mathbf{m}_t) + \nabla \cdot \sigma(\mathbf{m}) + \nabla h_t = -\mathbf{z}_1 & \text{in } \Omega \times (0,T) \\ -\nabla \cdot (k(x,t)\nabla h) - \nabla \cdot \mathbf{m} = z_2 & \text{in } \Omega \times (0,T) \\ -k(x,t)\nabla h \cdot n = 0 & \text{on } (\Gamma_N \cup \Gamma_{D,v}) \times (0,T) \\ (\sigma(\mathbf{m}_t) + \sigma + h_t I) n = 0 & \text{on } \Gamma_N \times (0,T) \\ h = 0 & \text{on } \Gamma_{D,p} \times (0,T) \\ \mathbf{m} = 0 & \text{on } \Gamma_D \times (0,T) \\ \nabla h(\cdot, 0) + \delta \nabla \cdot \sigma(\cdot, 0) = \nabla z_3 + \delta \nabla \cdot \sigma(\mathbf{z}_4) & \text{in } \Omega \end{cases} \quad (75)$$

the system is similar to the original poro-visco-elastic system described in (1)–(4). Existence and uniqueness of weak solution (in the sense of Definition 4) for given data $\mathbf{z}_1 \in L^2(0,T;\mathbf{V}')$, $z_2 \in L^2(0,T;V)$, $z_3 \in V$ and $\mathbf{z}_4 \in \mathbf{V}$ can be obtained using the strategy presented in [10, 45], namely Rothe's method.

4.2 First Order Necessary Optimality Conditions

First we provide an identity which is essential in the derivation of the first order necessary optimality conditions.

Lemma 2 *Let $(\mathbf{u}, p)$ be the weak solution to the poro-visco-elastic system (1)–(4) with $\mathbf{u}_0 = 0$ and $\mathbf{F} \in L^2\left(0,T;\mathbf{L}^2(\Omega)\right)$, $S \in L^2(0,T;L^2(\Omega))$, $\mathbf{g} \in L^2\left(0,T;\mathbf{L}^2(\Gamma_N)\right)$, $\psi \in L^2\left(0,T;L^2(\Gamma_{D,\mathbf{v}})\right)$. Let $(\mathbf{m}, h)$ be the weak solution to the adjoint equation (72) with $\mathbf{z}_1 \in L^2(0,T;\mathbf{V}')$, $z_2 \in L^2(0,T;V)$, $z_3 = 0$ and $\mathbf{z}_4 = 0$. Then the following identity holds:*

$$\int_0^T \langle \mathbf{g}, \mathbf{m} \rangle_{\Gamma_N} dt + \int_0^T (\mathbf{F}, \mathbf{m}) dt - \int_0^T \langle \psi, h \rangle_{\Gamma_{D,\mathbf{v}}} dt + \int_0^T (S, h) dt$$
$$= \int_0^T (\mathbf{z}_1, \mathbf{u}) dt + \int_0^T (z_2, p) dt.$$

Proof Let $(\mathbf{m}, h)$ be the weak solution to the adjoint equation (70). Since we have that

$$(h(t), \nabla \cdot \mathbf{w}) \leq \|h(t)\|_{L^2(\Omega)} \|\mathbf{w}\|_{\mathbf{V}}$$

then by Riesz Representation Theorem, there exits $R(h_\nabla)(t) \in \mathbf{V}$ such that for all $\mathbf{w} \in \mathbf{V}$

$$(h(t), \nabla \cdot \mathbf{w})_{L^2(\Omega)} = (R(h_\nabla)(t), \mathbf{w})_{\mathbf{V}} \text{ and } \|R(h_\nabla)(t)\|_{\mathbf{V}} \leq \|h(t)\|_{L^2(\Omega)}.$$

Therefore, $R(h_\nabla)(t) \in L^2(0, T; \mathbf{V})$.

Furthermore, for any $\mathbf{w} \in \mathbf{V}$ and $f \in C_0^\infty(0, T)$, we have

$$-\delta \int_0^T a(\mathbf{m}(t), \mathbf{w}) f'(t) dt + \int_0^T a(\mathbf{m}(t), \mathbf{w}) f(t) dt - \int_0^T (R(h_\nabla)(t), \mathbf{w})_{\mathbf{V}} f'(t) dt$$
$$= \int_0^T (\mathbf{z}_1(t), \mathbf{w}) f(t) dt.$$

Now we define the following linear and bounded functionals $F_i(t) : \mathbf{V} \to \mathbb{R}$:

$$F_1(t) : \mathbf{w} \mapsto -(\mathbf{z}_1(t), \mathbf{w})_{L^2(\Omega)}, \quad F_2(t) : \mathbf{w} \mapsto a(\mathbf{m}(t), \mathbf{w}),$$

with estimates given by

$$|F_1(t)\mathbf{w}| \leq \|\mathbf{z}_1(t)\|_{(L^2(\Omega))^3} \|\mathbf{w}\|_{(L^2(\Omega))^3} \leq \|\mathbf{z}_1(t)\|_{\mathbf{V}} \|\mathbf{w}\|_{\mathbf{V}}$$

$$|F_2(t)\mathbf{w}| \leq 2\|\nabla \mathbf{m}(t)\|_{(L^2(\Omega))^9} \|\nabla \mathbf{w}\|_{(L^2(\Omega))^9} \leq 2\|\mathbf{m}(t)\|_{\mathbf{V}} \|\mathbf{w}\|_{\mathbf{V}}.$$

This implies that $F_i(t) \in \mathbf{V}'$ for every t and there exists some constant $C > 0$ such that

$$\|F_1(t)\|_{\mathbf{V}'} + \|F_2(t)\|_{\mathbf{V}'} \leq C(\|z_1(t)\|_{\mathbf{V}} + \|\mathbf{m}(t)\|_{\mathbf{V}}).$$

Since $\mathbf{z}_1(t), \mathbf{m}(t) \in L^2(0, T; \mathbf{V})$, then we obtain that $F(t) = F_1(t) + F_2(t) \in L^2(0, T; \mathbf{V}')$. Now using Bochner's Theorem and (73) we have

$$\left(\int_0^T (\delta \mathbf{m}(t) + R(h_\nabla)(t)) f'(t) dt, \mathbf{w} \right)_{\mathbf{V}} = \int_0^T (\delta \mathbf{m}(t) + R(h_\nabla)(t)) f'(t), \mathbf{w})_{\mathbf{V}} dt$$
$$= \int_0^T (F(t) f(t), \mathbf{w})_{\mathbf{V}', \mathbf{V}} dt = \left(\int_0^T F(t) f(t) dt, \mathbf{w} \right)_{\mathbf{V}', \mathbf{V}}.$$

Hence, we have that in the space of $\mathbf{V}'$,

$$\int_0^T (\delta \mathbf{m}(t) + R(h_\nabla)(t)) f'(t) dt = \int_0^T F(t) f(t) dt, \quad \forall f \in C_0^\infty(0,T).$$

Therefore, by definition of distributional derivatives for vector valued functions, $\frac{d}{dt}(\delta \mathbf{m}(t) + R(h_\nabla)(t)) = F(t) \in L^2(0,T;\mathbf{V}')$.

Additionally, (73) can be written as

$$\int_0^T a(\mathbf{m},\mathbf{w}) f(t) dt - \int_0^T \left(\frac{d}{dt}(\delta \mathbf{m} + R(h_\nabla)), \mathbf{w} \right)_{\mathbf{V}',\mathbf{V}} f(t) dt = \int_0^T (\mathbf{z}_1, \mathbf{w}) f(t) dt. \tag{76}$$

However, since $\{\mathbf{w} f(t) | \mathbf{w} \in \mathbf{V},\ f \in C_0^\infty(0,T)\}$ is dense in $L^2(0,T;\mathbf{V})$, we can let $\mathbf{u}$ and p be test functions in (74) and (76) to obtain

$$\int_0^T a(\mathbf{m},\mathbf{u}) dt - \int_0^T \left(\frac{d}{dt}(\delta \mathbf{m} + R(h_\nabla)), \mathbf{u} \right)_{\mathbf{V}',\mathbf{V}} dt - \int_0^T (\nabla \cdot \mathbf{m}, p) dt$$
$$+ \int_0^T (k \nabla h, \nabla p) dt = \int_0^T (\mathbf{z}_1, \mathbf{u}) dt + \int_0^T (z_2, p) dt.$$

Since $\mathbf{u} \in H^1(0,T;\mathbf{V})$, we can integrate the second term by parts.

$$\int_0^T a(\mathbf{m},\mathbf{u}) dt + \int_0^T (\delta \mathbf{m} + R(h_\nabla), \mathbf{u}_t)_{\mathbf{V}} dt - (\delta \mathbf{m}(T) + R(h_\nabla)(T), \mathbf{u}(T))_{\mathbf{L}^2(\Omega)}$$
$$+ (\delta \mathbf{m}(0) + R(h_\nabla)(0), \mathbf{u}(0))_{\mathbf{L}^2(\Omega)} - \int_0^T (\nabla \cdot \mathbf{m}, p) dt + \int_0^T (k \nabla h, \nabla p) dt$$
$$= \int_0^T (\mathbf{z}_1, \mathbf{u}) dt + \int_0^T (z_2, p) dt. \tag{77}$$

Recall that for all $\mathbf{w} \in \mathbf{V}$, $\delta a(\mathbf{m}(T), \mathbf{w}) + (h(T), \nabla \cdot \mathbf{w}) = \delta a(\mathbf{z}_4, \mathbf{w}) + (z_3, \nabla \cdot \mathbf{w})$. Since $z_3 = 0$ and $\mathbf{z}_4 = 0$, $(z_3, \nabla \cdot \mathbf{w}) = 0$ and $\delta a(\mathbf{z}_4, \mathbf{w}) = 0$. Therefore, $(\delta \mathbf{m}(T) + R(h_\nabla)(T), \mathbf{w})_{\mathbf{V}} = \delta a(\mathbf{m}(T), \mathbf{w}) + (h, \nabla \cdot \mathbf{w}) = 0$. Hence, $\delta \mathbf{m}(T) + R(h_\nabla)(T) = 0$. Also, we assumed $\mathbf{u}_0 = 0$. Therefore, the temporal boundary terms in (77) are equal to 0, and using the definition of $R(h_\nabla)$, we have

$$\delta \int_0^T a(\mathbf{m}, \mathbf{u}_t) dt + \int_0^T a(\mathbf{m},\mathbf{u}) dt + \int_0^T (h, \nabla \cdot \mathbf{u}_t) dt - \int_0^T (\nabla \cdot \mathbf{m}, p) dt$$
$$+ \int_0^T (k \nabla h, \nabla p) dt = \int_0^T (\mathbf{z}_1, \mathbf{u}) dt + \int_0^T (z_2, p) dt. \tag{78}$$

Let $(\mathbf{m}, h)$ be the test functions used in the weak form of the poro-visco-elastic system. We obtain

$$\delta \int_0^T a(\mathbf{u}_t, \mathbf{m})dt + \int_0^T a(\mathbf{u}, \mathbf{m})dt - \int_0^T (p, \nabla \cdot \mathbf{m})dt + \int_0^T (k\nabla p, \nabla h)dt$$
$$+ \int_0^T (\nabla \cdot \mathbf{u}_t, h)dt$$
$$= \int_0^T \langle g, \mathbf{m}\rangle_{\Gamma_N} dt + \int_0^T (F, \mathbf{m})dt - \int_0^T \langle \psi, h\rangle_{\Gamma_{D,\mathbf{v}}} dt + \int_0^T (S, h)dt. \tag{79}$$

Combining (78) and (79), we obtain the desired equality:

$$\int_0^T \langle g, \mathbf{m}\rangle_{\Gamma_N} dt + \int_0^T (F, \mathbf{m})dt - \int_0^T \langle \psi, h\rangle_{\Gamma_{D,\mathbf{v}}} dt + \int_0^T (S, h)dt$$
$$= \int_0^T (\mathbf{z}_1, \mathbf{u})dt + \int_0^T (z_2, p)dt.$$

□

Finally, we can state and prove our theorem on first order necessary optimality conditions.

Theorem 3 *If $\bar{q} \in Q$ is the optimal control, then there exists a solution $(\mathbf{m}, h)$ to the adjoint system (70) which satisfies*

$$\begin{cases} \int_0^T (q - \bar{q}, h + \lambda\bar{q})_Q \geq 0 \ \forall q \in Q_{ad} & \text{when } S \text{ is used as the control } q \\ \int_0^T (q - \bar{q}, -h + \lambda\bar{q})_Q \geq 0 \ \forall q \in Q_{ad} & \text{when } \psi \text{ is used as the control } q \\ \int_0^T (q - \bar{q}, \mathbf{m} + \lambda\bar{q})_Q \geq 0 \ \forall q \in Q_{ad} & \text{when } \mathbf{F} \text{ or } \mathbf{g} \text{ is used as the control } q. \end{cases} \tag{80}$$

Conversely, let $\bar{q} \in Q_{ad}$ with associated state $(\bar{\mathbf{u}}, \bar{p})$. Let $(\mathbf{m}, h)$ be the solution to the adjoint system (70), and if (80) is satisfied, then $\bar{q}$ is a solution to the optimal control problem (62).

Proof Let $\bar{q} \in Q$ be the optimal control. Then $G(\bar{q}) = (\bar{\mathbf{u}}, \bar{p}) \in H^1(0, T; \mathbf{V}) \times L^2(0, T; V)$. Since $\bar{q}$ is optimal, (66) yields:

$$(G\bar{q} - \tilde{y}, Gq - G\bar{q})_Y + (\lambda\bar{q}, q - \bar{q})_Q \geq 0 \ \forall q \in Q_{ad},$$

which is equivalent to

$$\int_0^T (\bar{\mathbf{u}} - \tilde{\mathbf{u}}, \mathbf{u} - \bar{\mathbf{u}})dt + \int_0^T (\bar{p} - \tilde{p}, p - \bar{p})dt + (\lambda\bar{q}, q - \bar{q})_Q \geq 0 \ \forall q \in Q_{ad}. \tag{81}$$

Recall $G(q - \bar{q}) = (\mathbf{u} - \bar{\mathbf{u}}, p - \bar{p})$ corresponds to the solution of the poro-visco-elastic system where all sources and initial conditions are zero except the control $q - \bar{q}$. There exists $(\mathbf{m}, h)$ satisfying the adjoint system (70). Therefore, applying

Lemma 2, we obtain

$$(q - \bar{q}, h)_Q = \int_0^T (\bar{\mathbf{u}} - \tilde{\mathbf{u}}, \mathbf{u} - \bar{\mathbf{u}}) + \int_0^T (\bar{p} - \tilde{p}, p - \bar{p}) \ \forall q \in Q_{ad}, \tag{82}$$

if S is used as the control variable,

$$-(q - \bar{q}, h)_Q = \int_0^T (\bar{\mathbf{u}} - \tilde{\mathbf{u}}, \mathbf{u} - \bar{\mathbf{u}}) + \int_0^T (\bar{p} - \tilde{p}, p - \bar{p}) \ \forall q \in Q_{ad}, \tag{83}$$

if ψ is used as the control variable, and

$$(q - \bar{q}, \mathbf{m})_Q = \int_0^T (\bar{\mathbf{u}} - \tilde{\mathbf{u}}, \mathbf{u} - \bar{\mathbf{u}}) + \int_0^T (\bar{p} - \tilde{p}, p - \bar{p}) \ \forall q \in Q_{ad}, \tag{84}$$

if $\mathbf{F}$ or $\mathbf{g}$ is used as the control variable. Combining (82), (83), and (84) with inequality (81), we obtain the desired inequality (80).

Conversely, assume that $G(\bar{q}) = (\bar{\mathbf{u}}, \bar{p})$ and that there exists a solution $(\mathbf{m}, h)$ to the adjoint system (70) that satisfies (80). Then again using the identity from Lemma 2 we obtain

$$(G\bar{q} - \tilde{y}, Gq - G\bar{q})_Y + (\lambda\bar{q}, q - \bar{q})_Q \geq 0 \ \forall q \in Q_{ad},$$

which implies that $\bar{q}$ is the optimal control. □

Acknowledgments The authors were partially supported by L. Bociu's NSF-DMS 1555062.

References

1. J.L. Auriault, E. Sanchez-Palencia, Etude du comportement macroscopique d'un milieu poreux sature deformable. J. Mec. **16**(4), 575–603 (1977)
2. H.T. Banks, K. Bekele-Maxwell, L. Bociu, M. Noorman, G. Guidoboni, Local sensitivity via the complex-step derivative approximation for 1-D poro-elastic and poro-visco-elastic models. Math. Control Relat. Fields **9**(4), 623–642 (2019)
3. H.T. Banks, K. Bekele-Maxwell, L. Bociu, M. Noorman, G. Guidoboni, Sensitivity analysis in poro-elastic and poro-visco-elastic models with respect to boundary data. Quart. Appl. Math. **75**, 697–735 (2017)
4. H.T. Banks, K. Bekele-Maxwell, L. Bociu, M. Noorman, K. Tillman, The complex-step method for sensitivity analysis of non-smooth problems arising in biology. Eur. J. Math. Comput. Appl. **3**(3), 16–68 (2015)
5. H.T. Banks, K. Bekele-Maxwell, L. Bociu, C. Wang, Sensitivity via the complex-step method for delay differential equations with non-smooth initial data. Quart. Appl. Math. **75**, 231–248 (2017)
6. A. Bhole, B. Flynn, M. Liles, N.S.C. Dimarzio, J. Ruberti, Mechanical strain enhances survivability of collagen micronetworks in the presence of collagenase: implications for load-bearing matrix growth and stability. Philos. Trans. R. Soc. A **367**, 3339–3362 (2009)

7. M.A. Biot, General theory of three-dimensional consolidation. J. Appl. Phys. **12**(2), 155–164 (1941)
8. L. Bociu, S. Canic, B. Muha, J. Webster, Multilayered poroelasticity interacting with Stokes flow. SIAM J. Math. Anal. **53**(6), 6243–6279
9. L. Bociu, G. Guidoboni, R. Sacco, M. Verri, On the role of compressibility in poroviscoelastic models. Math. Biosci. Eng. **16**(5), 6167–6208 (2019)
10. L. Bociu, G. Guidoboni, R. Sacco, J. Webster, Analysis of nonlinear poro-elastic and poro-viscoelastic models. Arch. Rational Mech. Anal. **222**, 1445–1519 (2016)
11. L. Bociu, M. Noorman, Poro-visco-elastic models in biomechanics: sensitivity analysis. Commun. Appl. Anal. **23**(1), 61–77 (2019)
12. L. Bociu, J. Webster, Nonlinear quasi-static poroelasticity. J. Differ. Equa. **296**, 242–278 (2021)
13. R. Camp, M. Liles, J. Beale, N.S.B. Flynn, E. Moore, S. Murthy, J. Ruberti, Molecular mechanochemistry: low force switch slows enzymatic cleavage of human type I collagen monomer. J. Am. Chem. Soc. **133**, 4073–4078 (2011)
14. Y. Cao, S. Chen, A.J. Meir, Analysis and numerical approximations of equations of nonlinear poroelasticity. DCDS-B **18**, 1253–1273 (2013)
15. P. Causin, G. Guidoboni, A. Harris, D. Prada, R. Sacco, S. Terragni, A poroelastic model for the perfusion of the lamina cribrosa in the optic nerve head. Math. Biosci. **257**, 33–41 (2014)
16. E. Detournay, A.H.-D. Cheng, Comprehensive rock engineering: principles, practice and projects, in *Fundamentals of poroelasticity*, vol. II. Analysis and Design Method, ed. by C. Fairhurst (Pergamon Press, Oxford, 1993), pp. 113–171
17. M.R. DiSilvestro, J.-K.F. Suh, Biphasic poroviscoelastic characteristics of proteoglycan-depleted articular cartilage: simulation of degeneration. Ann. Biomed. Eng. **30**, 792–800 (2002)
18. B. Flynn, A.B.N. Saeidi, M. Liles, C. Dimarzio, J. Ruberti, Mechanical strain stabilizes reconstituted collagen fibrils against enzymatic degradation by mammalian collagenase matrix metalloproteinase 8 (MMP-8). PLoS One **5**, e12337 (2010)
19. R. Grytz, C. Girkin, V. Libertiaux, J. Downs, Perspectives on biomechanical growth and remodeling mechanisms in glaucoma. Mech. Res. Commun. **42**, 92–106 (2012)
20. R. Grytz, M. Fazio, M. Girard, V. Libertiaux, L. Bruno, S. Gardiner, C. Girkin, J. Downs, Material properties of the posterior human sclera. J. Mech. Behav. Biomed. Mater. **29**, 602–617 (2014)
21. G. Guidoboni, A. Harris, L. Carichino, Y. Arieli, B.A. Siesky, Effect of intraocular pressure on the hemodynamics of the central retinal artery: a mathematical model. Math. Biosci. Eng. **11**(3), 523–546 (2014)
22. R. Hollows, P. Graham, Intraocular pressure, glaucoma, and glaucoma suspects in a defined population. Br. J. Ophthalmol. **50**, 570–577 (1996)
23. W.M. Lai, J.S. Hou, V.C. Mow, A triphasic theory for the swelling and deformation behaviors of articular cartilage. ASME J. Biomech. Eng. **113**, 245–258 (1991)
24. A.F. Mak, The apparent viscoelastic behavior of articular cartilage - the contributions from the intrinsic matrix viscoelasticity and interstitial fluid flows. J. Biomech. Eng. **108**, 123–130 (1986)
25. J. Morgan-Davies, N. Taylor, A.R. Hill, P. Aspinall, C.J. O'Brien, A. Azuara-Blanco, Three dimensional analysis of the lamina cribrosa in glaucoma. Br. J. Ophthalmol. **88**(10), 1299–1304 (2004)
26. S. Nicaise, About the Lamé System in a Polygonal or a Polyhedral Domain and a Coupled Problem between the Lamé System and the Plate Equation I: Regularity of Solutions. Annali della Scuola Normale Superiore di Pisa. Classe di Scienze 4^e série **19**, 327–361 (1992)
27. S. Owczarek, A Galerkin method for Biot consolidation model. Math. Mech. Solids **15**, 42–56 (2010)
28. D. Prada, A. Harris, G. Guidoboni, B. Siesky, A.M. Huang, J. Arciero. Autoregulation in the optic nerve head. Major review. Surv. Ophthalmol. **61**(2), 164–186 (2016)
29. T. Roose, P.A. Netti, L. Munn, Y. Boucher, R. Jain, Solid stress generated by spheroid growth estimated using a linear poroelastic model. Microvascul. Res. **66**, 204–212 (2003)

30. R. Sacco, G. Guidoboni, A.G. Mauri, *A Comprehensive Physically Based Approach to Modeling in Bioengineering and Life Sciences* (Academic, London, 2019)
31. L.A. Setton, W. Zhu, V.C. Mow, The biphasic poroviscoelastic behavior of articular cartilage: role of the surface zone in governing the compressive behavior. J. Biomech. **26**, 581–592 (1993)
32. R. Shah, R. Wormald, Glaucoma. Clin. Evid. **9** (2009). Online
33. R.E. Showalter, Degenerate evolution equations and applications. Indiana Univ. Math. J. **23**(8), 655–677 (1974)
34. R.E. Showalter, Monotone operators in Banach space and nonlinear partial differential equations, in *AMS*. Mathematical Surveys and Monographs, vol. 49 (AMS, New York, 1996)
35. R.E. Showalter, Diffusion in poro-elastic media. JMAA **251**, 310–340 (2000)
36. A. Smillie, I. Sobey, Z. Molnar, A hydro-elastic model of hydrocephalus. J. Fluid Mech. **539**, 417–443 (2005)
37. M.A. Soltz, G.A. Ateshian, Experimental verification and theoretical prediction of cartilage interstitial fluid pressurization at an impermeable contact interface in confined compression. J. Biomech. **31**, 927–934 (1998)
38. D.E. Stewart, *Dynamics with Inequalities: Impacts and Hard Constraints* (SIAM, New York, 2011)
39. N. Su, R.E. Showalter, Partially saturated flow in a poroelastic medium. DCDS-B **1**, 403–420 (2001)
40. J.-K. Suh, S. Bai, Finite element formulation of biphasic poroviscoelastic model for articular cartilage. J. Biomech. Eng. **120**, 195–201 (1998)
41. C.C. Swan, R.S. Lakes, R.A. Brand, K.J. Stewart, Micromechanically based poroelastic modeling of fluid flow in haversian bone. J. Biomech. Eng. **125**, 25–37 (2003)
42. K. Terzaghi, *Principle of Soil Mechanics*. Eng. News Record, A Series of Articles (1925)
43. F. Tröltzsch, *Optimal Control of Partial Differential Equations. Theory, Methods, and Applications*, vol. 112 (AMS, Providence, 2010)
44. M. Verri, G. Guidoboni, L. Bociu, R. Sacco, The role of structural viscosity in deformable porous media with applications in biomechanics. Math. Biosci. Eng. **15**(4), 933–959 (2018)
45. A. Zenisek, The existence and uniqueness theorem in Biot's consolidation theory. Appl. Math. **29**, 194–211 (1984)

Global Gradient Estimate for a Divergence Problem and Its Application to the Homogenization of a Magnetic Suspension

Thuyen Dang, Yuliya Gorb, and Silvia Jiménez Bolaños

1 Introduction

The purpose of this paper is to generalize the results obtained by the authors in [10], where the rigorous analysis of the homogenization of a particulate flow consisting of a non-dilute suspension of a viscous Newtonian fluid with magnetizable particles was developed. Here, the fluid is assumed to be described by the Stokes flow and the particles are either paramagnetic or diamagnetic. The coefficients of the corresponding partial differential equations are locally periodic and a one-way coupling between the fluid domain and the particles is also assumed. Such one-way coupling has been observed in nature, see [11, Chapter 1]. For details and information about the applications and literature on the magnetic suspension, we turn to [10] and the references cited therein; however, the mathematical formulation of the considered problem is given in Sect. 2.2 below. References on the effective viscosity of a suspension without the coupling with magnetic field include [5, 6, 12–16, 18, 21, 23, 25, 31, 33].

In [10], a restrictive assumption about the magnetic permeability of the suspension, denoted by $\mathbf{a}$, was made. Here, the function $\mathbf{a}(\cdot)$ is locally periodic and elliptic, where the latter means that $\lambda \mathbf{I} \leq \mathbf{a}(x)$ and $\|\mathbf{a}\|_{L^\infty} \leq \Lambda$, for all $x \in \Omega$, with the

T. Dang
University of Houston, Houston, TX, USA
e-mail: ttdang9@central.uh.edu

Y. Gorb (✉)
National Science Foundation, Alexandria, VA, USA
e-mail: ygorb@nsf.gov

S. Jiménez Bolaños
Colgate University, Hamilton, NY, USA
e-mail: sjimenez@colgate.edu

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_6

suspension domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, including both the ambient fluid and the particles, and $\lambda, \Lambda > 0$ given in (A2)–(A3) below. The assumption on the function $\mathbf{a}$ made in [10] is as follows: for a given $s \in (4, 6]$, there exists a small number $\delta = \delta(\Lambda, d, \Omega) > 0$, for which the magnetic permeability $\mathbf{a}$ satisfies the following condition:

$$\operatorname{ess\,sup} \mathbf{a} - \operatorname{ess\,inf} \mathbf{a} \leq \delta. \tag{1}$$

As a consequence of (1), in [10] we obtained the following gradient estimate for the magnetic potential φ^ε:

$$\int_\Omega \left|\nabla \varphi^\varepsilon\right|^s \mathrm{d}x \leq C \int_\Omega |\mathbf{k}|^s \mathrm{d}x, \tag{2}$$

where the constant $C > 0$ is independent of ε, φ^ε and $\mathbf{k}$; with $0 < \varepsilon \ll 1$ the scale of the microstructure, $\mathbf{k} \in H^1(\Omega, \mathbb{R}^d)$ divergence-free, satisfying the compatibility condition $\displaystyle\int_{\partial\Omega} \mathbf{k} \cdot \mathbf{n}_{\partial\Omega}\, \mathrm{d}s = 0$, and appearing in the Neumann boundary condition on $\partial\Omega$, the boundary of the domain Ω, given by:

$$\left(\mathbf{a} \nabla \varphi^\varepsilon\right) \cdot \mathbf{n}_{\partial\Omega} = \mathbf{k} \cdot \mathbf{n}_{\partial\Omega}, \tag{3}$$

where $\mathbf{n}_{\partial\Omega}$ is the outward-pointing unit normal vector to $\partial\Omega$. The regularity result (2) was then used in the derivation of the effective (or homogenized) response of the given suspension that was rigorously justified in [10].

The main goal of this paper is to relax the assumption (1) on the magnetic permeability $\mathbf{a}$. To achieve this, we consider the Dirichlet boundary condition given in (4b) below, rather than one given in (3), and obtain the Lipschitz estimate (5), instead of (2), for the gradient of the magnetic potential φ^ε, see Theorem 1 below. In this paper, we are able to remove the condition (1) and have $\mathbf{a}$ only required to be *piecewise Hölder continuous*. Such relaxation will be based on the following observations:

- The De Giorgi-Nash-Moser estimate [20, Theorem 8.24] states that the solutions of scalar equations are Hölder continuous.
- If $1 > \varepsilon \geq \varepsilon_0$, for some $\varepsilon_0 > 0$, the uniform gradient bound (5) can be obtained by the result of Li and Vogelius [29]. The case when $\varepsilon_0 > \varepsilon > 0$ is resolved by the compactness method, which is discussed below.
- Provided $\mathbf{a}$ is also symmetric (this assumption is only necessary for the corrector results in Theorem 2), the gradients of the solutions of the cell problems are in $L^\infty(Y)$.

The main tools used in the proof of this theorem are (i) the regularity results of Li and Vogelius [29], and (ii) the celebrated *compactness method*, which was first used in homogenization in the seminal works of Avellaneda and Lin [3, 4]. Its machinery and applications in homogenization are carefully explained in [34]. In the context of homogenization, this method utilizes compactness in order to gain an improved regularity from a limiting equation via a proof by contradiction. This improvement of regularity is iterated and then used in a blow-up argument. Usually, the implementation of this method follows three steps, coined by Avellaneda and Lin [3, 4] as *(i)* “improvement,” *(ii)* “iteration,” and *(iii)* “blowup.”

The main contribution of this improved regularity result is that it will allow us to significantly widen the range of applicability of the results obtained in [10].

The outline of the paper is as follows. In Sect. 2, the main notations are introduced and the formulation of the fine-scale problem is discussed. Theorem 1, which provides an improved gradient estimate for the magnetic potential, is stated and discussed in Sect. 3. In Sect. 4, we obtain the interior Lipschitz and Hölder estimates, which provide the foundation for the boundary and corrector estimates discussed in Sect. 5. With all the results at hand, we then present the proof of our main theorem, also in Sect. 5. In Sect. 6, the homogenization results are obtained and summarized in Theorem 2. The conclusions are given in Sect. 7. The classical Schauder estimate is recalled in Appendix.

2 Formulation

2.1 Notation

For a measurable set A and a measurable function $f: A \to \mathbb{R}$, we define by $|A|$ the measure of A and $\fint_A f(x)\,\mathrm{d}x := \frac{1}{|A|}\int_A f(x)\,\mathrm{d}x$.

Throughout this paper, the scalar-valued functions, such as the pressure p, are written in usual typefaces, while vector-valued or tensor-valued functions, such as the velocity $\mathbf{u}$ and the Cauchy stress tensor $\boldsymbol{\sigma}$, are written in bold. Sequences are indexed by numeric superscripts (ϕ^i), while elements of vectors or tensors are indexed by numeric subscripts (x_i). Finally, the Einstein summation convention is used whenever applicable.

2.2 Setup of the Problem

Consider $\Omega \subset \mathbb{R}^d$, for $d \geq 2$, a simply connected and bounded domain of class $C^{1,\alpha}$, $0 < \alpha < 1$, and let $Y := (0, 1)^d$ be the unit cell in $\mathbb{R}^d$. The unit cell Y is decomposed into:

$$Y = Y_s \cup Y_f \cup \Gamma,$$

where Y_s, representing the magnetic inclusion, and Y_f, representing the fluid domain, are open sets in $\mathbb{R}^d$, and Γ is the closed $C^{1,\alpha}$ interface that separates them.

Let $i = (i_1, \ldots, i_d) \in \mathbb{Z}^d$ be a vector of indices and $\{\mathrm{e}^1, \ldots, \mathrm{e}^d\}$ be the canonical basis of $\mathbb{R}^d$. For a fixed small $\varepsilon > 0$, we define the dilated sets:

$$Y_i^\varepsilon := \varepsilon(Y + i), \quad Y_{i,s}^\varepsilon := \varepsilon(Y_s + i), \quad Y_{i,f}^\varepsilon := \varepsilon(Y_f + i), \quad \Gamma_i^\varepsilon := \partial Y_{i,s}^\varepsilon.$$

Typically, in homogenization theory, the positive number $\varepsilon \ll 1$ is referred to as the *size of the microstructure*. The effective (or *homogenized*) response of the given suspension corresponds to the case $\varepsilon = 0$.

We denote by $\mathbf{n}_i$, $\mathbf{n}_\Gamma$, and $\mathbf{n}_{\partial\Omega}$ the unit normal vectors to Γ_i^ε pointing outward $Y_{i,s}^\varepsilon$, to Γ pointing outward Y_s, and to $\partial\Omega$ pointing outward, respectively; and also, we denote by $\mathrm{d}\mathcal{H}^{d-1}$ the $(d-1)$-dimensional Hausdorff measure. In addition, we define the sets:

$$I^\varepsilon := \{i \in \mathbb{Z}^d : Y_i^\varepsilon \subset \Omega\}, \quad \Omega_s^\varepsilon := \bigcup_{i \in I^\varepsilon} Y_{i,s}^\varepsilon, \quad \Omega_f^\varepsilon := \Omega \setminus \Omega_s^\varepsilon, \quad \Gamma^\varepsilon := \bigcup_{i \in I^\varepsilon} \Gamma_i^\varepsilon,$$

see Fig. 1.

The magnetic permeability $\mathbf{a}$ is a $d \times d$ matrix satisfying the following conditions:

(A1) Y-periodicity: for all $z \in \mathbb{R}^d$, for all $m \in \mathbb{Z}$, and for all $k \in \{1, \ldots, d\}$ we have:

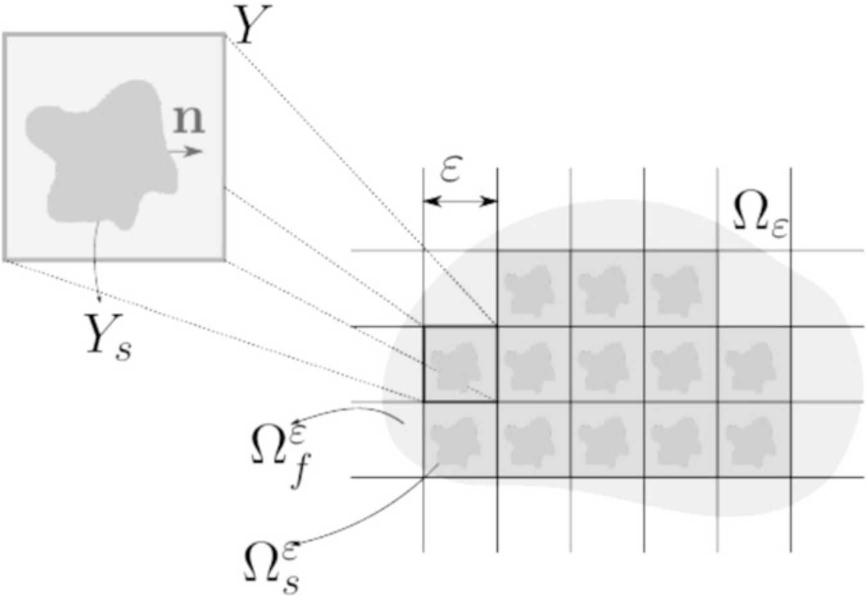


Fig. 1 Reference cell Y and domain Ω

$$\mathbf{a}(z + m\mathrm{e}^k) = \mathbf{a}(z).$$

(A2) Boundedness and measurability: there exists $\Lambda > 0$ such that:

$$\|\mathbf{a}\|_{L^\infty\left(\mathbb{R}^d\right)} \le \Lambda.$$

(A3) Ellipticity: there exists $\lambda > 0$ such that for all $\xi \in \mathbb{R}^d$, for all $x \in \mathbb{R}^d$, we have:

$$\mathbf{a}(x)\xi \cdot \xi \ge \lambda \left|\xi\right|^2.$$

Denote by $\mathfrak{M}(\lambda, \Lambda)$ the set of matrices that satisfy (A2)–(A3) and $\mathfrak{M}_{\mathrm{per}}(\lambda, \Lambda)$ the subset of matrices in $\mathfrak{M}(\lambda, \Lambda)$ that also satisfy (A1).

3 Statement and Discussion of the Main Result

The main result of this paper is summarized in the following theorem:

Theorem 1 (Global Lipschitz Estimate) *Let Ω be a bounded $C^{1,\alpha}$ domain, $g \in C^{1,\alpha'}(\partial\Omega)$, and $f \in L^\infty(\Omega)$, where $0 < \alpha' < \alpha < 1$. Suppose that $\mathbf{a} \in \mathfrak{M}_{\mathrm{per}}(\lambda, \Lambda)$ is piecewise C^α-continuous. There exists $C = C(\alpha, \alpha', \lambda, \Lambda, d, \Omega) > 0$ such that, for all $\varepsilon > 0$, the (unique) solution φ^ε of:*

$$-\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right] = f, \qquad \text{in } \Omega \tag{4a}$$

$$\varphi^\varepsilon = g, \qquad \text{on } \partial\Omega \tag{4b}$$

satisfies:

$$\left\|\nabla\varphi^\varepsilon\right\|_{L^\infty(\Omega)} \le C\left(\|g\|_{C^{1,\alpha'}(\partial\Omega)} + \|f\|_{L^\infty(\Omega)}\right). \tag{5}$$

Remark 1 For each $\varepsilon > 0$, let N_ε be the number of subdomains inside Ω such that in each of them the function $\mathbf{a}$ is C^α-continuous. Denote those subdomains by D_m, $1 \le m \le N_\varepsilon$. Then, for $0 < \alpha' < \min\left\{\alpha, \frac{\alpha}{(\alpha+1)d}\right\}$, by Li and Vogelius [29, Corollary 1.3], one has:

$$\left\|\nabla\varphi^\varepsilon\right\|_{L^\infty(\Omega)} \le C\left(\|g\|_{C^{1,\alpha'}(\partial\Omega)} + \|f\|_{L^\infty(\Omega)}\right), \tag{6}$$

where C depends on $\Omega, d, \alpha, \alpha', \lambda, \Lambda, \|\mathbf{a}\|_{C^{\alpha'}(\overline{D_m}, \mathbb{R}^{d\times d})}$, and the $C^{1,\alpha}$-modulus of $\cup_{m=1}^{N_\varepsilon}\partial D_m$ (defined in page 92 [29]). As $\varepsilon \to 0$, the number N_ε increases, while the sizes of the subdomains decrease, which leads to the blowup of the $C^{1,\alpha}$-modulus. Therefore, estimate (6) is not uniform in ε.

However, if $\varepsilon_0 \le \varepsilon \le 1$ for some constant $\varepsilon_0 > 0$, then one can control the number, size, distance, and $C^{1,\alpha}$-modulus of the subdomains in Ω, uniformly with respect to ε. Note that the upper and lower bounds of those quantities are positive and independent of ε. Thus, by Li and Vogelius [29, Corollary 1.3], there exists C independent of ε such that (5) holds when $\varepsilon_0 \le \varepsilon \le 1$. Therefore, Theorem 1 will be proven, if one can specify a constant $\varepsilon_0 > 0$ such that (5) holds for $0 < \varepsilon < \varepsilon_0$.

Our proof of Theorem 1 follows the classical steps in regularity theory: (i) derive an interior Lipschitz estimate, (ii) derive a boundary Lipschitz estimate, and finally (iii) combine the estimates in (i) and (ii) to obtain the global Lipschitz estimate. Step (i) is obtained via the *compactness method*. For step (ii), we additionally need to establish the following preliminary results:

- Interior and boundary Hölder estimates, see Propositions 1–3.
- Estimates for the Green's function, which are obtained using the Hölder bounds above, see Proposition 5. The existence of the Green's function for scalar uniformly elliptic equations is established in [22, 26, 30].
- Estimates for the Dirichlet boundary corrector, see Proposition 6.

If the coefficient $\mathbf{a}$ is not Hölder continuous, then the classical results in [3, 34, 37] cannot be applied directly. Nevertheless, some of their proofs can be adapted for the case at hand. In those situations, we will explicitly point out what needs to be modified in their proofs, in order to relax the continuity assumption on the coefficient matrix $\mathbf{a}$.

4 Interior Estimates

We start with an estimate for homogenized equations, i.e., the equations with constant coefficients, which are limits of some fine-scale problems.

Lemma 1 *Let $\lambda > 0$, $\Lambda > 0$, $\gamma > 0$ and $0 < \mu < \frac{1}{2}$ be fixed. For each constant matrix $\mathbf{b} \in \mathfrak{M}(\lambda, \Lambda)$, $h \in L^{d+\gamma}(B(x_0, 1))$, with $\|h\|_{L^{d+\gamma}(B(x_0,1))} \le 1$, there exists $\theta = \theta(\gamma, \mu, \lambda, \Lambda, d) > 0$ such that if $\phi \in H^1(B(x_0, 1))$ satisfies:*

$$-\operatorname{div}(\mathbf{b}\nabla\phi) = h \text{ in } B(x_0, 1),$$

then the following estimate holds:

$$\sup_{|x-x_0|<\theta} \left| \phi(x) - \phi(x_0) - (x - x_0) \cdot \fint_{B(x_0,\theta)} \nabla\phi(z)\,\mathrm{d}z \right| < \theta^{1+3\mu/4}. \tag{7}$$

Proof By the classical Schauder estimate for the scalar equation with constant coefficients (Theorem 3), we have $\phi \in C^{1,\mu}\left(B\left(x_0, 1/4\right)\right)$ and:

$$\begin{aligned}\|\phi\|_{C^{1,\mu}(B(x_0,1/4))} &\le C(\gamma,\mu,\lambda,\Lambda,d)\left(\|h\|_{L^{d+\gamma}(B(x_0,3/8))}+\|\phi\|_{H^1(B(x_0,3/8))}\right)\\ &\le C(\gamma,\mu,\lambda,\Lambda,d)\,\|h\|_{L^{d+\gamma}(B(x_0,1))}\\ &\le C(\gamma,\mu,\lambda,\Lambda,d).\end{aligned} \tag{8}$$

For $0<\theta<\frac{1}{4}$ and $|x-x_0|<\theta$, there exist z_x such that:

$$\begin{aligned}&\left|\phi(x)-\phi(x_0)-(x-x_0)\cdot\fint_{B(x_0,\theta)}\nabla\phi(z)\,\mathrm{d}z\right|\\ &=\left|\frac{x-x_0}{|B(x_0,\theta)|}\cdot\int_{B(x_0,\theta)}(\nabla\phi(z_x)-\nabla\phi(y))\,\mathrm{d}y\right|\\ &\le C(\gamma,\mu,\lambda,\Lambda,d)\,|x-x_0|^{1+\mu}\\ &\le C(\gamma,\mu,\lambda,\Lambda,d)\theta^{1+\mu}.\end{aligned}$$

Choosing θ small enough so that $C(\gamma,\mu,\lambda,\Lambda,d)\theta^{1+\mu}<\theta^{1+3\mu/4}$, we obtain (7).

□

The fact that θ does not depend on the matrix $\mathbf{b}$ or the source term h is crucial for the contradiction argument in Proposition 1 below. We now state the interior Lipschitz estimate. Note that, here, $\mathbf{a}$ is not necessarily Hölder continuous in the domain Ω.

Proposition 1 (Interior Lipschitz Estimate I) *Suppose that* $\mathbf{a}\in\mathfrak{M}_{\mathrm{per}}(\lambda,\Lambda)$ *and* $f\in L^\infty(\Omega)$. *Fix* $x_0\in\Omega$ *and* $R>0$*, so that* $B(x_0,R)\subset\Omega$. *There exist* $\varepsilon_0=\varepsilon_0(\lambda,\Lambda,R,d)>0$ *and* $C=C(\lambda,\Lambda,R,d)>0$ *such that, for all* $0<\varepsilon<\varepsilon_0$ *and for every weak solution* $\varphi^\varepsilon\in H^1(B(x_0,R))$ *of the equation* $-\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right]=f$ *in* $B(x_0,R)$*, the following estimate holds:*

$$\left\|\nabla\varphi^\varepsilon\right\|_{L^\infty(B(x_0,R/2))}\le C\left(\left\|\varphi^\varepsilon\right\|_{L^\infty(B(x_0,R))}+\|f\|_{L^\infty(B(x_0,R))}\right). \tag{9}$$

Proof By dilation, we may assume that $R=1$. Fix $0<\mu<\frac{1}{2}$. We prove, by the compactness method, that there exists $\varepsilon_0=\varepsilon_0(\lambda,\Lambda,d,\mu)$ so that (9) holds for all $0<\varepsilon<\varepsilon_0$. To do this, we only need to show that there exists $C>0$, independent of ε, such that:

$$\max\left\{\left\|\varphi^\varepsilon\right\|_{L^\infty(B(x_0,1))},\|f\|_{L^\infty(B(x_0,1))}\right\}\le 1 \text{ implies } \left\|\nabla\varphi^\varepsilon\right\|_{L^\infty(B(x_0,1/2))}\le C. \tag{10}$$

Let $\boldsymbol{\omega} := (\omega^1, \ldots, \omega^d)$, where $\omega^i \in H^1_{\text{per}}(Y)/\mathbb{R},\ 1 \le i \le d$, is the solution of the *cell problem*:

$$-\operatorname{div}_y \left[\mathbf{a}(y)\left(\mathrm{e}^i + \nabla_y \omega^i(y)\right)\right] = 0 \text{ in } Y. \tag{11}$$

1. *Improvement.* In this step, we prove by contradiction that:

   For fixed $0 < \mu < \frac{1}{2}$, there exist θ and ε^*, with $0 < \theta < \frac{1}{4},\ 0 < \varepsilon^* < 1$, depending on λ, Λ, d, and μ, such that if $\mathbf{a} \in \mathfrak{M}_{\text{per}}(\lambda, L),\ f \in L^\infty(B(x_0, 1))$, $\varphi^\varepsilon \in H^1(B(x_0, 1))$ satisfy:

$$-\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right] = f \text{ in } B(x_0, 1), \tag{12a}$$

$$\max\left\{\|\varphi^\varepsilon\|_{L^\infty(B(x_0,1))}, \|f\|_{L^\infty(B(x_0,1))}\right\} \le 1, \tag{12b}$$

   then, for all $0 < \varepsilon < \varepsilon^*$, we have:

$$\sup_{|x-x_0|<\theta}\left|\varphi^\varepsilon(x) - \varphi^\varepsilon(x_0) - \left[x - x_0 + \varepsilon\boldsymbol{\omega}\left(\frac{x}{\varepsilon}\right)\right]\cdot \fint_{B(x_0,\theta)} \nabla\varphi^\varepsilon(z)\,\mathrm{d}z\right| \le \theta^{1+\mu/2}, \tag{13}$$

   where $\boldsymbol{\omega}$ solves (11).

   Take θ as in (7) of Lemma 1. By contradiction, suppose there exist sequences:

$$\varepsilon_n \to 0, \quad \mathbf{a}_n \in \mathfrak{M}_{\text{per}}(\lambda, L), \quad f_n \in L^\infty\left(B(x_0, 1)\right), \text{ and } \varphi_n \in H^1(B(x_0, 1))$$

   satisfying:

$$-\operatorname{div}\left[\mathbf{a}_n\left(\frac{x}{\varepsilon_n}\right)\nabla\varphi_n\right] = f_n \text{ in } B(x_0, 1), \tag{14}$$

$$\max\left\{\|\varphi_n\|_{L^\infty(B(x_0,1))}, \|f_n\|_{L^\infty(B(x_0,1))}\right\} \le 1, \tag{15}$$

   such that:

$$\sup_{|x-x_0|<\theta}\left|\varphi_n(x) - \varphi_n(x_0) - \left[x - x_0 + \varepsilon_n\boldsymbol{\omega}\left(\frac{x}{\varepsilon_n}\right)\right]\cdot \fint_{B(x_0,\theta)} \nabla\varphi_n(z)\,\mathrm{d}z\right| > \theta^{1+\mu/2}. \tag{16}$$

   Let $\mathcal{A}_n \in \mathfrak{M}(\lambda, \Lambda)$ denote the effective matrix corresponding to $\mathbf{a}_n$. By the Banach–Alaoglu theorem, the Caccioppoli inequality, the Rellich–Kondrachov theorem, and the Schauder theorem (see, e.g., [19, Theorem 4.4] and [7, Theorem 3.16, 6.4 and 9.16]), there exist functions $\varphi_0 \in L^2(B(x_0, 1)),\ f_0 \in L^\infty(B(x_0, 1))$ and a constant matrix $\mathbf{a}_0 \in \mathfrak{M}(\lambda, \Lambda)$ such that, up to subsequences, we have:

$$\begin{aligned}
\varphi_n &\rightharpoonup \varphi_0 \text{ in } L^2(B(x_0,1))\\
f_n &\overset{*}{\rightharpoonup} f_0 \text{ in } L^\infty(B(x_0,1))\\
\varphi_n &\rightharpoonup \varphi_0 \text{ in } H^1(B(x_0,1/2))\\
f_n &\to f_0 \text{ in } H^{-1}(B(x_0,1))\\
\mathcal{A}_n &\to \mathbf{a}_0.
\end{aligned}$$

By [9, Theorem 13.4 (iii)] or [37, Theorem 2.3.2], we have φ_0 is the solution of:

$$-\operatorname{div}(\mathbf{a}_0\nabla\varphi_0) = f_0 \text{ in } B(x_0,1/2). \tag{17}$$

Fix $x \in B(x_0,1)$ and let $U \subset B(x_0,1)$ be an open neighborhood of x. By the De Giorgi-Nash-Moser Theorem [20, Theorem 8.24], there exists $0 < \beta = \beta(d,\lambda/\Lambda) < 1$ such that:

$$\|\varphi_n\|_{C^\beta(\overline{U})} \leq C\left(\|\varphi_n\|_{L^\infty(B(x_0,1))} + \|f_n\|_{L^\infty(B(x_0,1))}\right) \leq 2C.$$

By the Arzela–Ascoli Theorem, up to a subsequence, φ_n uniformly converges to φ^* in $C(U)$ for some φ^*. Since $\varphi_n \rightharpoonup \varphi_0$ in $L^2(B(x_0,1))$, we conclude that $\varphi^* = \varphi_0$ a.e. in U. Therefore, $\lim_{n\to\infty}\varphi_n(x) = \varphi_0(x)$ a.e. in $B(x_0,1)$. Letting $n \to \infty$ in (15), the argument above and [7, Theorem 3.13] yield:

$$\max\left\{\|\varphi_0\|_{L^\infty(B(x_0,1))}, \|f_0\|_{L^\infty(B(x_0,1))}\right\} \leq 1,$$

which, together with (17), implies that (7) still holds for $\phi = \varphi_0$ (observe that, from (8), shrinking the domain from $B(x_0,1)$ to $B(x_0,1/2)$ does not affect estimates), that is:

$$\sup_{|x-x_0|<\theta}\left|\varphi_0(x) - \varphi_0(x_0) - (x-x_0)\cdot \fint_{B(x_0,\theta)} \nabla\varphi_0(z)\,\mathrm{d}z\right| < \theta^{1+3\mu/4}. \tag{18}$$

On the other hand, letting $n \to \infty$ in (16) and since $\|\boldsymbol{\omega}\|_{L^\infty(Y)} < \infty$, we obtain:

$$\sup_{|x-x_0|<\theta}\left|\varphi_0(x) - \varphi_0(x_0) - (x-x_0)\cdot \fint_{B(x_0,\theta)} \nabla\varphi_0(z)\,\mathrm{d}z\right| \geq \theta^{1+\mu/2},$$

which contradicts (18), since $0 < \theta < \frac{1}{4}$.

2. *Iteration* Let $0 < \varepsilon < \varepsilon^*$. Direct evaluation yields that:

$$P^\varepsilon(x) := \frac{1}{\theta^{1+\mu/2}}\Big\{\varphi^\varepsilon(\theta x) - \varphi^\varepsilon(\theta x_0)$$

$$-\left[\theta(x-x_0)+\varepsilon\boldsymbol{\omega}\left(\frac{\theta x}{\varepsilon}\right)\right]\cdot\fint_{B(x_0,\theta)}\nabla\varphi^\varepsilon(z)\,\mathrm{d}z\Bigg\}$$

solves the following equation:

$$-\operatorname{div}\left[\mathbf{a}\left(\frac{\theta x}{\varepsilon}\right)\nabla P^\varepsilon(x)\right]=\tilde{f}\text{ in }B(x_0,1),\tag{19}$$

where $\tilde{f}:=\theta^{1-\mu/2}f(\theta x)$. Moreover, by (13) and (12b), we have:

$$\left\|P^\varepsilon\right\|_{L^\infty(B(x_0,1))}\le 1\ \text{ and }\ \left\|\tilde{f}\right\|_{L^\infty(B(x_0,1))}\le 1,$$

so by using (13) again, we obtain:

$$\sup_{|x-x_0|<\theta}\left|P^\varepsilon(x)-P^\varepsilon(x_0)-\left[x-x_0+\frac{\varepsilon}{\theta}\boldsymbol{\omega}\left(\frac{\theta x}{\varepsilon}\right)\right]\cdot\fint_{B(x_0,\theta)}\nabla P^\varepsilon\right|(z)\,\mathrm{d}z\le\theta^{1+\mu/2}.\tag{20}$$

From (20) and scaling down, we have:

$$\sup_{|x-x_0|<\theta^2}\left|\varphi^\varepsilon(x)-\varphi^\varepsilon(x_0)-(x-x_0)\cdot a_2^\varepsilon+\varepsilon b_2^\varepsilon\right|\le\theta^{2(1+\mu/2)},$$

where

$$\begin{aligned}a_2^\varepsilon&:=\fint_{B(x_0,\theta)}\nabla\varphi^\varepsilon(z)\,\mathrm{d}z+\theta^{\mu/2}\fint_{B(x_0,\theta)}\nabla P^\varepsilon(z)\,\mathrm{d}z,\\ b_2^\varepsilon(y)&:=\boldsymbol{\omega}(y)\cdot\left(\fint_{B(x_0,\theta)}\nabla\varphi^\varepsilon(z)\,\mathrm{d}z+\theta^{\mu/2}\fint_{B(x_0,\theta)}\nabla P^\varepsilon(z)\,\mathrm{d}z\right),\ \text{ for }y:=\frac{x}{\varepsilon}\in Y.\end{aligned}\tag{21}$$

By the De Giorgi-Nash-Moser estimate and Caccioppoli inequality [2, Lemma C.2], there exists a constant C, depending only on λ, Λ, and d, such that:

$$\|\boldsymbol{\omega}\|_{L^\infty(Y)}\le C,\qquad\left|\fint_{B(x_0,\theta)}\nabla\varphi^\varepsilon(z)\,\mathrm{d}z\right|\le C/\theta,\qquad\left|\fint_{B(x_0,\theta)}\nabla P^\varepsilon(z)\,\mathrm{d}z\right|\le C/\theta.$$

Therefore, (21) implies that:

$$\begin{aligned}\left|a_2^\varepsilon\right|&\le(C/\theta)\left(1+\theta^{\mu/2}\right),\\ \left\|b_2^\varepsilon\right\|_{L^\infty(Y)}&\le(C/\theta)\left(1+\theta^{\mu/2}\right).\end{aligned}$$

Reiterating this process, we obtain that there exists $C = C(\gamma, \lambda, \Lambda, d, \mu) > 0$ such that:

$$\begin{aligned} \left|a_k^\varepsilon\right| &\leq (C/\theta)\left(1+\theta^{\mu/2}+\cdots+\theta^{(k-1)\mu/2}\right), \\ \left\|b_k^\varepsilon\right\|_{L^\infty(Y)} &\leq (C/\theta)\left(1+\theta^{\mu/2}+\cdots+\theta^{(k-1)\mu/2}\right), \end{aligned} \tag{22}$$

and:

$$\sup_{|x-x_0|<\theta^k}\left|\varphi^\varepsilon(x)-\varphi^\varepsilon(x_0)-(x-x_0)\cdot a_k^\varepsilon+\varepsilon b_k^\varepsilon\right| \leq \theta^{k(1+\mu/2)}. \tag{23}$$

3. *Blowup* Let $\varepsilon_0 := \min\left\{\varepsilon^*, \frac{1}{5\sqrt{d}}\right\}$ and $0 < \varepsilon < \varepsilon_0$.

Choose k such that $\theta^{k+1} \leq 4\varepsilon\sqrt{d} < \theta^k$. Then from (23), there exists $C = C(\theta, d) > 0$ so that:

$$\sup_{|x-x_0|<4\varepsilon\sqrt{d}}\left|\varphi^\varepsilon(x)-\varphi^\varepsilon(x_0)-(x-x_0)\cdot a_k^\varepsilon+\varepsilon b_k^\varepsilon\right| \leq C\varepsilon^{1+\mu/2}, \tag{24}$$

which, together with (22), leads to:

$$\left\|\varphi^\varepsilon-\varphi^\varepsilon(x_0)\right\|_{L^\infty(B(x_0,4\varepsilon\sqrt{d}))} \leq C\varepsilon. \tag{25}$$

Denote by z_0^ε the center of the cell Y_i^ε containing x_0, and define:

$$v^\varepsilon(x) := \frac{1}{\varepsilon}\left[\varphi^\varepsilon(\varepsilon x+z_0^\varepsilon)-\varphi^\varepsilon(\varepsilon x_0+z_0^\varepsilon)\right], \quad x \in \Omega.$$

Then, $\nabla v^\varepsilon(x) = \nabla\varphi^\varepsilon\left(\varepsilon x+z_0^\varepsilon\right)$ and, moreover, v^ε solves:

$$-\operatorname{div}\left[\mathbf{a}\left(x+\frac{z_0^\varepsilon}{\varepsilon}\right)\nabla v^\varepsilon(x)\right] = \varepsilon f(\varepsilon x+z_0^\varepsilon) \quad \text{in } B\left(0, 3\sqrt{d}\right). \tag{26}$$

Observe that:

$$\begin{aligned} \frac{1}{\varepsilon}\left(B\left(x_0, \varepsilon\sqrt{d}\right)-z_0^\varepsilon\right) &\subset B\left(0, 2\sqrt{d}\right) \\ &\subset B\left(0, 3\sqrt{d}\right) \subset \frac{1}{\varepsilon}\left(B\left(x_0, 4\varepsilon\sqrt{d}\right)-z_0^\varepsilon\right). \end{aligned} \tag{27}$$

Applying [29, Theorem 1.1] to (26), we obtain that there exists a constant $C > 0$, independent of ε and x_0, such that:

$$\begin{aligned}
&\left\|\nabla v^{\varepsilon}\right\|_{L^{\infty}\left(\frac{1}{\varepsilon}\left(B\left(x_0,\varepsilon\sqrt{d}\right)-z_0^{\varepsilon}\right)\right)} \\
&\quad\le \left\|\nabla v^{\varepsilon}\right\|_{L^{\infty}\left(B\left(0,2\sqrt{d}\right)\right)} \\
&\quad\le C\left[\left\|v^{\varepsilon}\right\|_{L^{\infty}\left(B\left(0,3\sqrt{d}\right)\right)} + \sup_{x\in B\left(0,3\sqrt{d}\right)}\left|\varepsilon f(\varepsilon x+z_0^{\varepsilon})\right|\right] \qquad (28)\\
&\quad\le C\left[\left\|v^{\varepsilon}\right\|_{L^{\infty}\left(\frac{1}{\varepsilon}\left(B\left(x_0,4\varepsilon\sqrt{d}\right)-z_0^{\varepsilon}\right)\right)} + \|f\|_{L^{\infty}\left(\frac{1}{\varepsilon}\left(B\left(x_0,4\varepsilon\sqrt{d}\right)-z_0^{\varepsilon}\right)\right)}\right].
\end{aligned}$$

Scaling down (28) and using (25), we obtain:

$$\begin{aligned}
&\left\|\nabla \varphi^{\varepsilon}\right\|_{L^{\infty}\left(B\left(x_0,\varepsilon\sqrt{d}\right)\right)} \\
&\le C\left[\frac{1}{\varepsilon}\left\|\varphi^{\varepsilon}-\varphi^{\varepsilon}(x_0)\right\|_{L^{\infty}\left(\left(B\left(x_0,4\varepsilon\sqrt{d}\right)\right)\right)} + \|f\|_{L^{\infty}\left(B\left(x_0,4\varepsilon\sqrt{d}\right)\right)}\right] \le C,
\end{aligned}$$

where $C > 0$ is independent of x_0 and ε.

Remark 2 Under stronger smoothness assumptions on the coefficient **a**, similar estimates to (6) are proved in the literature. In particular, if **a** is in VMO($\mathbb{R}^d$), the real-variable method of L. Caffarelli and I. Peral [8] yields an uniform $W^{1,p}-$estimate; on the other hand, if **a** is Hölder continuous, then one has the uniform Lipschitz estimate. Those results hold also for elliptic systems and even for Neumann boundary condition. We refer the reader to [3, 4, 27, 28, 34–37] and the references cited therein.

However, in this paper, we focus on the case when **a** is only piecewise Hölder continuous. A similar argument as in the papers cited above together with the regularity theorem of Li and Vogelius [29, Theorem 1.1] yield the interior Lipschitz estimate as showed in Proposition 1. Moreover, some additional care is needed to ensure that the constant C in (28) is independent of *both* ε and x_0. In the Blow-up step of the proof above, one may try to let

$$s^{\varepsilon}(x) := \frac{1}{\varepsilon}\varphi^{\varepsilon}(\varepsilon x + x_0) \qquad (29)$$

so that

$$-\operatorname{div}\left[\mathbf{a}(x)\nabla s^{\varepsilon}(x)\right] = \varepsilon f(\varepsilon x + x_0), \qquad (30)$$

then by applying [29, Theorem 1.1], one obtains

$$\left\|\nabla s^{\varepsilon}\right\|_{L^{\infty}\left(B\left(0,\frac{1}{2\varepsilon_0}\right)\right)} \le C'\left[\left\|s^{\varepsilon}\right\|_{L^{\infty}\left(B\left(0,\frac{1}{\varepsilon_0}\right)\right)} + \|f\|_{L^{\infty}\left(B\left(0,\frac{1}{\varepsilon_0}\right)\right)}\right].$$

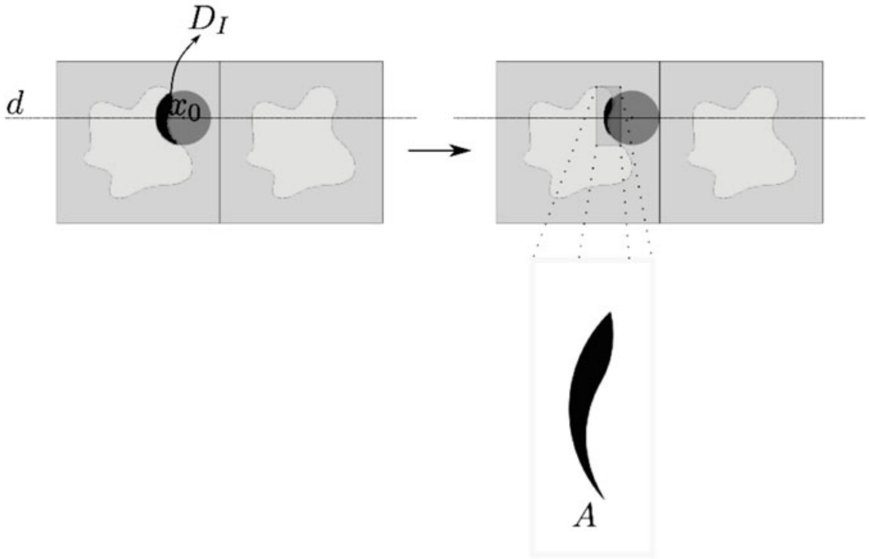


Fig. 2 As the center x_0 of the ball $B\left(x_0, \frac{1}{\varepsilon_0}\right)$ slides on the line d to the right, the subdomain D_I shrinks to 0, which makes the $C^{1,\alpha}$ modulus to become unbounded [29, page 93]. Moreover, in some cases, it is possible that a cusp also appears at some points (point A on the zoomed in figure above)

However, C' *indeed depends on* x_0. The reason is that, when one shifts x_0 in the scaling (29), one also changes the $C^{1,\alpha}$-modulus of the subdomains, where the latter, in the context of our problem, are generated by taking intersections of the ball centered at x_0 and the heterogenous domain. In short, we do not have the uniform control of the subdomains when using the scaling (29) for arbitrary x_0, see Fig. 2. In order to circumvent the dependence on x_0, we use a different scaling and combine with a geometric argument, as demonstrated in the proof of Proposition 1.

The following result follows from Proposition 1, the De Giorgi-Nash-Moser estimate and a change of variable.

Proposition 2 (Interior Lipschitz Estimate II) *Suppose that* $\mathbf{a} \in \mathfrak{M}_{\text{per}}(\lambda, \Lambda)$ *and* $f \in L^\infty(\Omega)$. *Fix* $x_0 \in \Omega$ *and* $R > 0$ *so that* $B(x_0, R) \subset \Omega$. *There exist* $\varepsilon_0 = \varepsilon_0(\lambda, \Lambda, d) > 0$ *and* $C = C(\lambda, \Lambda, d) > 0$ *such that, for all* $0 < \varepsilon < \varepsilon_0$, *the weak solution* $\varphi^\varepsilon \in H^1(B(x_0, R))$ *of the equation* $-\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right] = f$ *in* $B(x_0, R)$ *satisfies:*

$$\left|\nabla\varphi^\varepsilon(x_0)\right| \leq C'\left[\left(\fint_{B(x_0,R)} \left|\nabla\varphi^\varepsilon(z)\right|^2 \mathrm{d}z\right)^{\frac{1}{2}} + R \sup_{x\in B(x_0,R)} |f(x)|\right]. \tag{31}$$

Proof Without loss of generality, we assume $x_0 = 0$. By Proposition 1, with $R = 1$ and considering $\varphi^\varepsilon - \fint_{B(0,1)} \varphi^\varepsilon(z)\,\mathrm{d}z$, which solves $-\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right] = f$ in $B(0,1)$, we have that there exist $\varepsilon_0 > 0$ and $C' > 0$, depending only on λ, Λ, d, such that:

$$\begin{aligned}
\left|\nabla\varphi^\varepsilon(0)\right| &\leq \left\|\nabla\varphi^\varepsilon\right\|_{L^\infty(B(0,1/4))} \\
&\leq C'\left(\left\|\varphi^\varepsilon - \fint_{B(0,1)} \varphi^\varepsilon(z)\,\mathrm{d}z\right\|_{L^\infty(B(0,1/2))} + \|f\|_{L^\infty(B(0,1/2))}\right) \\
&\leq C'\left(\left\|\varphi^\varepsilon - \fint_{B(0,1)} \varphi^\varepsilon(z)\,\mathrm{d}z\right\|_{L^2(B(0,1))} + \|f\|_{L^\infty(B(0,1))}\right) \\
&\leq C'\left(\left\|\nabla\varphi^\varepsilon\right\|_{L^2(B(0,1))} + \|f\|_{L^\infty(B(0,1))}\right) \\
&\leq C'\left[\left(\fint_{B(0,1)} \left|\nabla\varphi^\varepsilon(z)\right|^2 \mathrm{d}z\right)^{\frac{1}{2}} + \sup_{x\in B(0,1)} |f(x)|\right],
\end{aligned} \tag{32}$$

where we have used the De Giorgi-Nash-Moser estimate and Poincaré's inequality.

For $R > 0$ and $x \in B(0,1)$, let $v^\varepsilon(x) := R^{-1}\varphi^\varepsilon(Rx)$, then $\nabla v^\varepsilon(x) = \nabla\varphi^\varepsilon(Rx)$ and:

$$-\operatorname{div}\left[\mathbf{b}\left(\frac{x}{\varepsilon}\right)\nabla v^\varepsilon(x)\right] = Rf(Rx),$$

where $\mathbf{b}(z) := \mathbf{a}(Rz)$. We have $\mathbf{b} \in \mathfrak{M}(\lambda, \Lambda)$ is $R^{-1}Y$−periodic. Note that the proof of Proposition 1 does not depend on the period, hence, (32) holds for v^ε in particular:

$$\begin{aligned}
\left|\nabla\varphi^\varepsilon(0)\right| &= \left|\nabla v^\varepsilon(0)\right| \\
&\leq C'\left[\left(\fint_{B(0,1)} \left|\nabla v^\varepsilon(x)\right|^2 \mathrm{d}x\right)^{\frac{1}{2}} + R\sup_{x\in B(0,1)} |f(Rx)|\right] \\
&= C'\left[\left(\fint_{B(0,1)} \left|\nabla\varphi^\varepsilon(Rx)\right|^2 \mathrm{d}x\right)^{\frac{1}{2}} + R\sup_{x\in B(0,1)} |f(Rx)|\right].
\end{aligned} \tag{33}$$

By a change of variable in (33), we obtain (31). □

We recall the interior Hölder estimate, adapted from [34, Proposition 1] (or [3, Lemma 9]) that will be used to obtain the boundary Hölder estimate in the next section.

Proposition 3 (Interior Hölder Estimate) *Suppose that* $\mathbf{a} \in \mathfrak{M}_{\mathrm{per}}(\lambda, \Lambda)$ *and* $f \in L^{d+\gamma}(\Omega)$*, for some* $\gamma > 0$*. Fix* $x_0 \in \Omega$ *and* $R > 0$ *such that* $B(x_0, R) \subset \Omega$*. Let*

$0 < \mu := \frac{\gamma}{d+\gamma} < 1$. *There exists* $C = C(\gamma, \lambda, \Lambda, p, R, d) > 0$ *such that, for all* $\varepsilon > 0$*, the weak solution* $\varphi^\varepsilon \in H^1(B(x_0, R))$ *of the equation* $-\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right] = f$ *in* $B(x_0, R)$ *satisfies:*

$$\left[\varphi^\varepsilon\right]_{C^{0,\mu}(B(x_0,R/2))} \leq C\left(\left\|\varphi^\varepsilon\right\|_{L^2(B(x_0,R))} + \|f\|_{L^{d+\gamma}(B(x_0,R))}\right), \tag{34}$$

where $[h]_{C^{0,\mu}(A)} := \sup_{x\neq y\in A}\frac{|h(x)-h(y)|}{|x-y|^\mu}$.

The proof of Proposition 3 is similar to [34, Proposition 1]. Indeed, a closer look at the proof of [34, Proposition 1] reveals that the Hölder continuity assumption on **a** is needed only for the classical Schauder estimate for elliptic systems to hold. However, this paper is devoted to the scalar case, and we use the De Giorgi-Nash-Moser Theorem [20, Theorem 8.24], for which the assumption **a** is bounded is sufficient.

Remark 3 For the case of elliptic systems, the De Giorgi-Nash-Moser Theorem does not hold, see counterexamples by De Giorgi, Giusti and Miranda, and others, cf. [19, Section 9.1], and the references cited therein. Because of that, this paper is concerned with the scalar case only.

5 Boundary Estimates, Green Functions, Dirichlet Correctors, and Proof of Main Theorem

The following result is adapted from [3, Section 2.3] and [37, Section 5.2].

Proposition 4 (Boundary Hölder Estimate) *Suppose that* $\mathbf{a} \in \mathfrak{M}_{\text{per}}(\lambda, \Lambda)$*, and* Ω *is a* C^1*-domain. Fix* $x_0 \in \partial\Omega$*,* $0 < r < \operatorname{diam}(\Omega)$*, and* $0 < \mu < 1$*. Let* $g \in C^{0,1}\left(B(x_0, r)\cap\partial\Omega\right)$*. There exist* $\varepsilon_0 = \varepsilon_0(\mu, \lambda, \Lambda, d) > 0$ *and* $C = c(\mu, \lambda, \Lambda, d) > 0$ *such that, for all* $0 < \varepsilon < \varepsilon_0$*, every weak solution* $\varphi^\varepsilon \in H^1(B(x_0, r))$ *of the equation:*

$$\begin{aligned}-\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right] &= 0 \text{ in } B(x_0, r)\cap\Omega,\\ \varphi^\varepsilon &= g \text{ on } B(x_0, r)\cap\partial\Omega\end{aligned}$$

satisfies:

$$\begin{aligned}&\left[\varphi^\varepsilon\right]_{C^{0,\mu}(B(x_0,r/2)\cap\Omega)}\\ &\quad\leq Cr^{-\mu}\left[\left(⨍_{B(x_0,r)\cap\Omega}\left|\varphi^\varepsilon(z)\right|^2\mathrm{d}z\right)^{\frac{1}{2}} + |g(x_0)| + r\,\|g\|_{C^{0,1}(B(x_0,r)\cap\partial\Omega)}\right].\end{aligned} \tag{35}$$

The proof of Proposition 4 follows the proof of [37, Theorem 5.2.1] with minor modifications. In [37, Theorem 5.2.1], the assumption that $\mathbf{a} \in \mathrm{VMO}(\mathbb{R}^d)$ is used only in two places: (1) when $\varepsilon \geq \varepsilon_0$, which is beyond the scope of this particular theorem, and (2) to obtain the interior Hölder estimate, which we already relaxed in Proposition 3.

Thanks to Propositions 3 and 4, we can drop the assumption that $\mathbf{a} \in \mathrm{VMO}(\mathbb{R}^d)$ of Theorem 5.4.1–2 and Lemma 5.4.5 in [37]. The results are summarized in the following proposition.

Proposition 5 (Green's Functions) *Suppose that* $\mathbf{a} \in \mathfrak{M}_{\mathrm{per}}(\lambda, \Lambda)$ *and* Ω *is a* C^1*-domain. Fix* $0 < \mu, \sigma, \sigma_1 < 1$ *and let* $\delta(x) := \operatorname{dist}(x, \partial\Omega)$. *Then, there exist* $\varepsilon_0 = \varepsilon_0(\mu, \lambda, \Lambda, d) > 0$ *and* $C = C(\lambda, \Lambda, \sigma, \sigma_1, \Omega) > 0$ *such that, for all* $0 < \varepsilon < \varepsilon_0$, *the Green's functions* $G^\varepsilon(x, y)$ *exist and satisfy the following:*

$$\left|G^\varepsilon(x, y)\right| \leq \begin{cases} C\frac{1}{|x-y|^{d-2}} & \text{if } d \geq 3, \\ C\left[1 + \ln\left(\frac{r_0}{|x-y|}\right)\right] & \text{if } d = 2. \end{cases} \tag{36a}$$

$$\left|G^\varepsilon(x, y)\right| \leq \begin{cases} \frac{C\delta(x)^\sigma}{|x-y|^{d-2+\sigma}} & \text{if } \delta(x) < \frac{1}{2}\left|x-y\right|, \\ \frac{C\delta(y)^{\sigma_1}}{|x-y|^{d-2+\sigma_1}} & \text{if } \delta(y) < \frac{1}{2}\left|x-y\right|, \\ \frac{C\delta(x)^\sigma\delta(y)^{\sigma_1}}{|x-y|^{d-2+\sigma+\sigma_1}} & \text{if } \delta(x) < \frac{1}{2}\left|x-y\right| \text{ or } \delta(y) < \frac{1}{2}\left|x-y\right|, \end{cases} \tag{36b}$$

$$\int_\Omega \left|\nabla_y G^\varepsilon(x, y)\right| \delta(y)^{\sigma-1}\, \mathrm{d}y \leq C\delta(x)^\sigma, \tag{36c}$$

where $x, y \in \Omega$, $x \neq y$ *and* $r_0 := \operatorname{diam}(\Omega)$.

As a consequence, for $0 < c < 1$ *and* $g \in C^{0,1}(\Omega)$, *there exists* $C = C(\lambda, \Lambda, d) > 0$ *such that, for any* $x_0 \in \partial\Omega$, *for any* ε *satisfying* $c\varepsilon \leq \min\{c\varepsilon_0, r\} \leq r < r_0 := \operatorname{diam}(\Omega)$, *and for any solution* φ^ε *of the Dirichlet problem* $-\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right] = 0$ *in* Ω, $\varphi^\varepsilon = g$ *on* $\partial\Omega$, *the following estimate holds:*

$$\left(\fint_{B(x_0,r)\cap\Omega} \left|\nabla\varphi^\varepsilon\right|^2\right)^{\frac{1}{2}} \leq C\left[\|\nabla g\|_{L^\infty(\Omega)} + \varepsilon^{-1}\|g\|_{L^\infty(\Omega)}\right]. \tag{37}$$

We now define the boundary Dirichlet corrector: For $1 \leq i \leq d$, let $\Phi^{i,\varepsilon} \in H^1(\Omega)$ be the solution of the problem:

$$\begin{aligned} -\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\Phi^{i,\varepsilon}(x)\right] &= 0 \text{ for } x \in \Omega, \\ \Phi^{i,\varepsilon}(z) &= z_i \text{ for } z \in \partial\Omega. \end{aligned} \tag{38}$$

The following proposition provides a bound on the boundary Dirichlet corrector.

Proposition 6 *Let Ω be a bounded $C^{1,\alpha}$-domain. Suppose that* $\mathbf{a} \in \mathfrak{M}_{\text{per}}(\lambda, L)$ *is piecewise C^α-continuous. Then, for all $\varepsilon > 0$, the solution $\Phi^{i,\varepsilon}$ of (38) satisfies:*

$$\left\| \nabla \Phi^{i,\varepsilon} \right\|_{L^\infty(\Omega)} \leq C, \tag{39}$$

where constant C depends only on λ, Λ and Ω.

The proof is similar to [37, Theorem 5.4.4]. One only needs to use three observations:

- The case $c\varepsilon \geq \min\{c\varepsilon_0, r\}$ follows from [29, Theorem 1.2], by the same argument used in Remark 1.
- Let $\boldsymbol{\omega} = (\omega^1, \omega^2, \dots, \omega^d)$ be the solutions of the cell problems (11). Given only that $\mathbf{a}$ is piecewise C^α-continuous, then $\nabla\boldsymbol{\omega}$ is bounded in $L^\infty(Y, \mathbb{R}^{d\times d})$, see the first paragraph in the proof of [17, Theorem 3.2] or [38, Corollary 3.5].
- The interior Lipschitz estimate in Proposition 2 only requires $\mathbf{a}$ is piecewise Hölder continuous.

We now combine Propositions 6, 2, and [29, Theorem 1.2] to obtain a discontinuous coefficient-version of [37, Theorem 5.5.1].

Proposition 7 (Boundary Lipschitz Estimate) *Suppose that* $\mathbf{a} \in \mathfrak{M}_{\text{per}}(\lambda, L)$ *is piecewise C^α-continuous and Ω is a $C^{1,\alpha}$-domain. Fix $x_0 \in \partial\Omega$, $0 < r <$* $\operatorname{diam}(\Omega)$ *and $0 < \mu < 1$. Let $g \in C^{1,\alpha}(B(x_0, r) \cap \partial\Omega)$. There exist $\varepsilon_0 = \varepsilon_0(\mu, \lambda, \Lambda, d, \Omega) > 0$ and $C = C(\mu, \lambda, \Lambda, d, \Omega) > 0$ such that, for all $0 < \varepsilon < \varepsilon_0$, the weak solution $\varphi^\varepsilon \in H^1(B(x_0, r))$ of the equation:*

$$\begin{aligned} -\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right] &= 0 \textit{ in } B(x_0, r) \cap \Omega, \\ \varphi^\varepsilon &= g \textit{ on } B(x_0, r) \cap \partial\Omega \end{aligned}$$

satisfies:

$$\begin{aligned} &\left\|\nabla\varphi^\varepsilon\right\|_{L^\infty(B(x_0,r/2)\cap\partial\Omega)} \\ &\leq C\left[r^{-1}\left(\fint_{B(x_0,r)\cap\Omega}\left|\varphi^\varepsilon\right|^2\right)^{\frac{1}{2}} + r^\alpha \left\|\nabla_{\tan}g\right\|_{C^{0,\alpha}(B(x_0,r)\cap\partial\Omega)}\right. \\ &\qquad\left. + \left\|\nabla_{\tan}g\right\|_{L^\infty(B(x_0,r)\cap\partial\Omega)} + r^{-1}\left\|g\right\|_{L^\infty(B(x_0,r)\cap\partial\Omega)}\right]. \end{aligned} \tag{40}$$

The estimate (5) of Theorem 1 is a consequence of Propositions 2 and 7, by an argument similar to [37, Theorem 5.6.2].

6 Application to Magnetic Suspensions

In this section, we apply the regularity results obtained above to the rigorous homogenization procedure discussed in [10]. For that, we first recap the formulation of the fine-scale problem and the homogenization result itself. We begin by introducing the definition of two-scale convergence, which will be used below.

Definition 1 A sequence $\{v^\varepsilon\}_{\varepsilon>0}$ in $L^2(\Omega)$ is said to *two-scale converge* to $v = v(x, y)$, with $v \in L^2(\Omega \times Y)$, if and only if:

$$\lim_{\varepsilon\to 0}\int_\Omega v^\varepsilon(x)\psi\left(x, \frac{x}{\varepsilon}\right)\mathrm{d}x = \frac{1}{|Y|}\int_\Omega\int_Y v(x, y)\psi(x, y)\,\mathrm{d}y\,\mathrm{d}x,$$

for any test function $\psi = \psi(x, y)$, with $\psi \in \mathcal{D}(\Omega, C^\infty_{\text{per}}(Y))$, see [1, 9, 32]. In this case, we write $v^\varepsilon \xrightarrow{2} v$.

Let the kinematic viscosity be denoted by $\nu = \frac{\eta}{\rho_f}$, where $\eta > 0$ and $\rho_f > 0$ are the fluid viscosity and the fluid density, respectively. The dimensionless quantities that appear in this problem are the (hydrodynamic) *Reynolds number* $R_e = UL/\nu$, the *Froude number* $F_r = U/\sqrt{FL}$, and the *coupling parameter* $S = \frac{B^2}{\rho_f \Lambda U^2}$, where L, U, B, and F are the characteristic scales corresponding to length, fluid velocity, magnetic field, and body density force, respectively. Moreover, $\Lambda > 0$ is defined in (A2).

From now on, we suppose Ω is $C^{3,\alpha}$, which is needed for the corrector result below. Suppose further that $\mathbf{g} \in H^1(\Omega, \mathbb{R}^d)$, $k \in C^{1,\alpha}(\partial\Omega)$, and $f \in L^\infty(\Omega)$. Let $\mathbf{u}^\varepsilon$ and p^ε be the fluid velocity and the fluid pressure, respectively. Also, in a space free of current, the magnetic field strength is given by $\mathbf{H}^\varepsilon = \nabla\varphi^\varepsilon$, for some magnetic potential $\varphi^\varepsilon(x)$. Let $\mathbf{u}^\varepsilon \in H^1_0(\Omega, \mathbb{R}^d)$, $p^\varepsilon \in L^2(\Omega)/\mathbb{R}$, and $\varphi^\varepsilon \in H^1(\Omega)$ be the solution of the following boundary value problem:

$$-\operatorname{div}\left[\boldsymbol{\sigma}(\mathbf{u}^\varepsilon, p^\varepsilon) + \boldsymbol{\tau}(\varphi^\varepsilon)\right] = \frac{1}{F_r^2}\mathbf{g}, \qquad \text{in } \Omega^\varepsilon_f \tag{41a}$$

$$\operatorname{div}\mathbf{u}^\varepsilon = 0, \qquad \text{in } \Omega^\varepsilon_f \tag{41b}$$

$$\mathbb{D}(\mathbf{u}^\varepsilon) = 0, \qquad \text{in } \Omega^\varepsilon_s \tag{41c}$$

$$-\operatorname{div}\left[\mathbf{a}\left(\frac{x}{\varepsilon}\right)\nabla\varphi^\varepsilon\right] = f \qquad \text{in } \Omega, \tag{41d}$$

together with the balance equations:

$$\int_{\Gamma^\varepsilon_i}\left[\boldsymbol{\sigma}(\mathbf{u}^\varepsilon, p^\varepsilon) + \boldsymbol{\tau}(\varphi^\varepsilon)\right]\mathbf{n}_i\,\mathrm{d}\mathcal{H}^{d-1} = 0, \tag{42a}$$

$$\int_{\Gamma_i^\varepsilon} \left(\left[\boldsymbol{\sigma}(\mathbf{u}^\varepsilon, p^\varepsilon) + \boldsymbol{\tau}(\varphi^\varepsilon) \right] \mathbf{n}_i \right) \times \mathbf{n}_i \, \mathrm{d}\mathcal{H}^{d-1} = 0, \tag{42b}$$

and boundary conditions:

$$\mathbf{u}^\varepsilon = 0, \text{ on } \partial\Omega, \tag{43a}$$

$$\varphi^\varepsilon = k, \text{ on } \partial\Omega, \tag{43b}$$

where

$$\boldsymbol{\sigma}(\mathbf{u}^\varepsilon, p^\varepsilon) := \frac{2}{R_e}\mathbb{D}(\mathbf{u}^\varepsilon) - p^\varepsilon \mathbf{I}, \tag{44a}$$

$$\mathbb{D}(\mathbf{u}^\varepsilon) := \frac{\nabla \mathbf{u}^\varepsilon + \nabla^\top \mathbf{u}^\varepsilon}{2}, \tag{44b}$$

$$\boldsymbol{\tau}(\varphi^\varepsilon) := S\mathbf{a}\left(\frac{x}{\varepsilon}\right)\left(\nabla\varphi^\varepsilon \otimes \nabla\varphi^\varepsilon - \frac{1}{2}\left|\nabla\varphi^\varepsilon\right|^2 \mathbf{I}\right) \tag{44c}$$

are the *rate of strain*, the *Cauchy stress*, and the *Maxwell stress* tensors, respectively. For the detailed derivation and the physical meaning of the equations above, we refer the readers to [10] and the references therein. Observe that, in the context of this paper, we consider the Dirichlet boundary condition (43b), instead of a Neumann boundary condition (3) in [10], to relax the regularity assumption on the magnetic permeability needed in [10]. Then, the weak formulation for (41d) and (43b) is given by:

$$\begin{aligned} &\int_\Omega \mathbf{a}\left(\frac{x}{\varepsilon}\right) \nabla\left(\varphi^\varepsilon - k\right) \cdot \nabla\xi \, \mathrm{d}x \\ &\quad = -\int_\Omega \mathbf{a}\left(\frac{x}{\varepsilon}\right) \nabla k \cdot \nabla \xi \, \mathrm{d}x + \int_\Omega f\xi \, \mathrm{d}x, \quad \forall \xi \in H_0^1(\Omega). \end{aligned} \tag{45}$$

One immediately has that $\|\varphi^\varepsilon\|_{H^1(\Omega)} \leq C\left(\|k\|_{H^{1/2}(\partial\Omega)} + \|f\|_{L^q(\Omega)}\right)$, which implies that φ^ε is two-scale convergent (up to a subsequence). Choosing a test function as in [10, Lemma 3.7], we obtain the cell problem (46) and the first two effective equations defined in (50) below.

Moreover, Theorem 1 ensures that $\nabla\varphi^\varepsilon$ is uniformly bounded in $L^\infty(\Omega, \mathbb{R}^d)$, with respect to $\varepsilon \in (0, \varepsilon_0)$. Therefore, we obtain the existence, uniqueness and a priori bounds for $\mathbf{u}^\varepsilon$ and p^ε as in [10, Corollary 3.11]. Here, we have relaxed the restrictive assumption (1) made in [10] and we can use our results in the case when the constant magnetic permeability is anisotropic, namely when $\mathbf{a}$ is a matrix.

To carry on with the homogenization formulation, for $1 \leq i, j \leq d$, denote by $\mathbf{U}^{ij}$ the vector defined by $\mathbf{U}_k^{ij} := y_j \delta_{ik}$. Consider $\omega^i \in H^1_{\text{per}}(Y)/\mathbb{R}$, the solution of:

$$-\operatorname{div}_y \left[\mathbf{a}(y)\left(\mathrm{e}^i + \nabla_y \omega^i(y)\right)\right] = 0 \text{ in } Y. \tag{46}$$

Also, consider $\boldsymbol{\chi}^{ij} \in H^1_{\text{per}}(Y, \mathbb{R}^d)/\mathbb{R}$ and $q^{ij} \in L^2(Y)/\mathbb{R}$, solving:

$$\begin{aligned}
\operatorname{div}_y \left[\mathbb{D}_y\left(\mathbf{U}^{ij} - \boldsymbol{\chi}^{ij}\right) + q^{ij}\mathbf{I}\right] &= 0 \text{ in } Y_f,\\
\operatorname{div}_y \boldsymbol{\chi}^{ij} &= 0 \text{ in } Y,\\
\mathbb{D}_y\left(\mathbf{U}^{ij} - \boldsymbol{\chi}^{ij}\right) &= 0 \text{ in } Y_s,\\
\int_\Gamma \left[\mathbb{D}_y\left(\mathbf{U}^{ij} - \boldsymbol{\chi}^{ij}\right) - q^{ij}\mathbf{I}\right]\mathbf{n}_\Gamma \,\mathrm{d}\mathcal{H}^{d-1} &= 0,\\
\int_\Gamma \left[\mathbb{D}_y\left(\mathbf{U}^{ij} - \boldsymbol{\chi}^{ij}\right) - q^{ij}\mathbf{I}\right]\mathbf{n}_\Gamma \times \mathbf{n}_\Gamma \,\mathrm{d}\mathcal{H}^{d-1} &= 0,
\end{aligned} \tag{47}$$

and consider $\boldsymbol{\xi}^{ij} \in H^1_{\text{per}}(Y, \mathbb{R}^d)/\mathbb{R}$ and $r^{ij} \in L^2(Y)/\mathbb{R}$, solving:

$$\begin{aligned}
\operatorname{div}_y \left[\mathbb{D}_y\left(\boldsymbol{\xi}^{ij}\right) + r^{ij}\mathbf{I} + \boldsymbol{\tau}^{ij}\right] &= 0 \text{ in } Y_f,\\
\operatorname{div}_y \boldsymbol{\xi}^{ij} &= 0 \text{ in } Y,\\
\mathbb{D}_y\left(\boldsymbol{\xi}^{ij}\right) &= 0 \text{ in } Y_s,\\
\int_\Gamma \left[\mathbb{D}_y\left(\boldsymbol{\xi}^{ij}\right) + r^{ij}\mathbf{I} + \boldsymbol{\tau}^{ij}\right]\mathbf{n}_\Gamma \,\mathrm{d}\mathcal{H}^{d-1} &= 0,\\
\int_\Gamma \left[\mathbb{D}_y\left(\boldsymbol{\xi}^{ij}\right) + r^{ij}\mathbf{I} + \boldsymbol{\tau}^{ij}\right]\mathbf{n}_\Gamma \times \mathbf{n}_\Gamma \,\mathrm{d}\mathcal{H}^{d-1} &= 0.
\end{aligned} \tag{48}$$

We also define:

$$\begin{aligned}
\mathcal{A}_{jk} &:= \frac{1}{|Y|}\int_Y \mathbf{a}(y)(\mathrm{e}^k + \nabla\omega^k(y)) \cdot (\mathrm{e}^j + \nabla\omega^j(y))\,\mathrm{d}y,\\
\mathcal{N}^{ij}_{mn} &:= \frac{1}{|Y|}\int_Y \mathbb{D}_y(\mathbf{U}^{ij} - \boldsymbol{\chi}^{ij}) : \mathbb{D}_y(\mathbf{U}^{mn} - \boldsymbol{\chi}^{mn})\,\mathrm{d}y,\\
\boldsymbol{\tau}^{ij}_{\text{ref}} &:= \mathbf{a}(y)\left[(\mathrm{e}^i + \nabla_y\omega^i) \otimes (\mathrm{e}^j + \nabla_y\omega^j) - \frac{1}{2}(\mathrm{e}^i + \nabla_y\omega^i) \cdot (\mathrm{e}^j + \nabla_y\omega^j)\mathbf{I}\right],\ y \in Y,\\
\mathcal{B}^{ij} &:= \frac{1}{|Y|}\int_Y \left(\mathbb{D}_y(\boldsymbol{\xi}^{ij}) + \boldsymbol{\tau}^{ij}\right)\mathrm{d}y,
\end{aligned} \tag{49}$$

where $\mathcal{A}$ is the *effective magnetic permeability*, which is symmetric and elliptic. The tensor $\mathcal{N} := \left\{\mathcal{N}^{ij}_{mn}\right\}_{1\leq i,j,m,n\leq d}$ is the *effective viscosity*, and it is a fourth rank tensor. Moreover, $\mathcal{N}$ is symmetric, i.e., $\mathcal{N}^{ij}_{mn} = \mathcal{N}^{mn}_{ij} = \mathcal{N}^{ji}_{mn} = \mathcal{N}^{ij}_{nm}$, and it satisfies the Legendre-Hadamard condition (or strong ellipticity condition), i.e., there exist $\beta > 0$ such that, for all $\zeta, \eta \in \mathbb{R}^d$, one has $\mathcal{N}^{ij}_{mn}\zeta_i\zeta_m\eta_j\eta_n \geq \beta\,|\zeta|^2\,|\eta|^2$. The matrix $\boldsymbol{\tau}_{\mathrm{ref}}$ is the *Maxwell stress tensor* on Y, and $\mathcal{B}$ is the *effective coupling matrix.*

By the same argument as in Theorem 3.5, Lemma 3.9, and Lemma 3.14 of [10], the following result holds:

Theorem 2 *Let* $(\varphi^\varepsilon, \mathbf{u}^\varepsilon, p^\varepsilon) \in H^1(\Omega) \times H^1_0(\Omega, \mathbb{R}^d) \times L^2_0(\Omega)$ *be the solution of* (41). *Then*

$$\begin{aligned} \varphi^\varepsilon &\rightharpoonup \varphi^0 \text{ in } H^1(\Omega),\\ \mathbf{u}^\varepsilon &\rightharpoonup \mathbf{u}^0 \text{ in } H^1_0(\Omega, \mathbb{R}^d),\\ p^\varepsilon &\rightharpoonup \pi^0 \text{ in } L^2_0(\Omega), \end{aligned}$$

where φ^0, $\mathbf{u}^0$, *and* π^0 *are solutions of:*

$$\begin{aligned} -\operatorname{div}\left(\mathcal{A}\nabla\varphi^0\right) &= f && \text{in } \Omega,\\ \varphi^0 &= k && \text{on } \partial\Omega,\\ \operatorname{div}\left[\frac{2}{R_e}\mathcal{N}^{ij}\mathbb{D}\left(\mathbf{u}^0\right)_{ij} - \pi^0 + S\mathcal{B}^{ij}\frac{\partial\varphi^0}{\partial x_i}\frac{\partial\varphi^0}{\partial x_j}\right] &= \frac{1}{F_r^2}\mathbf{g} && \text{in } \Omega,\\ \operatorname{div}\mathbf{u}^0 &= 0 && \text{in } \Omega\\ \mathbf{u}^0 &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{50}$$

with $\mathcal{A}, \mathcal{N}^{ij}, \mathcal{B}^{ij}$, $1 \leq i, j \leq d$, *defined in* (49). *Moreover, the first-order correctors satisfy:*

$$\begin{aligned} \lim_{\varepsilon\to 0}\left\|\nabla\varphi^\varepsilon(\cdot) - \nabla\varphi^0(\cdot) - \nabla_y\varphi^1\left(\cdot, \frac{\cdot}{\varepsilon}\right)\right\|_{L^2(\Omega,\mathbb{R}^d)} &= 0,\\ \lim_{\varepsilon\to 0}\left\|\mathbb{D}(\mathbf{u}^\varepsilon)(\cdot) - \mathbb{D}(\mathbf{u}^0)(\cdot) - \mathbb{D}_y(\mathbf{u}^1)\left(\cdot, \frac{\cdot}{\varepsilon}\right)\right\|_{L^2(\Omega,\mathbb{R}^{d\times d})} &= 0, \end{aligned}$$

where

$$\begin{aligned} \varphi^1(x, y) &:= \omega^i(y)\frac{\partial\varphi^0}{\partial x_i}(x),\\ \mathbf{u}^1(x, y) &:= -\mathbb{D}\left(\mathbf{u}^0(x)\right)_{ij}\boldsymbol{\chi}^{ij}(y) + S\frac{\partial\varphi^0}{\partial x_i}(x)\frac{\partial\varphi^0}{\partial x_j}(x)\boldsymbol{\xi}^{ij}(y). \end{aligned}$$

7 Conclusions

This paper concerns a homogenized description of a non-dilute suspension of magnetic particles in a viscous flow. The results demonstrated in this paper generalize the ones obtained by the authors in [10], where a more restrictive assumption on the magnetic permeability (1) was used and a Neumann boundary condition (3) was imposed instead of the Dirichlet condition (4b). Theorem 2 above demonstrates the *effective response* of a viscous fluid with a locally periodic array of paramagnetic/diamagnetic particles suspended in it, given by the system of equations (41). The effective equations are described by (50), with the effective coefficients given in (49). These effective quantities depend only on the instantaneous position of the particles, their geometry, and the magnetic and flow properties of the original suspension described by (41). Using the tools introduced in [29] and the compactness method, an improved regularity estimate for the gradient of the magnetic potential of the original fine-scale problem (41) was obtained, see Theorem 1. This theorem allows us to drop the restrictive assumption (1) mentioned above. Comparing to the classical results on regularity of this type, we do require the coefficient matrix belongs to a VMO-space, see, e.g., [3, 34, 37]. Recently, in [17, Proposition 3.1], the authors obtained an L^q-bound of the gradient of the solution of the scalar divergence equation, uniform with respect to ε, for $q < \infty$. Our result, in Theorem 1, shows that the gradient bound actually holds for the case $q = \infty$.

Acknowledgments The authors would like to thank Antoine Gloria for bringing up the results in [17] to their attention. The first author would also like to thank Christophe Prange for allowing him to audit his class, in which he first learned about the compactness method used in this paper. The work of the first author was partially supported by the NSF through grant DMS-1350248. The work of the third author was supported by NSF grant DMS-2110036. This material is based upon work supported by and while serving at the National Science Foundation for the second author Yuliya Gorb. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect views of the National Science Foundation.

Appendix

Theorem 3 (Interior Schauder Estimates [20, 24]) *Let* $\mathbf{b} \in \mathfrak{M}(\lambda, \Lambda)$ *be a constant matrix and* $w \in H^1(\Omega)$ *be a weak solution of:*

$$\mathbf{b}_{ij} D_i D_j w = f + \sum_{i=1}^{d} D_i f_i .$$

For every $\alpha \in (0, 1)$*, there exists a uniform constant* $C = C(\alpha, d, \lambda, \Lambda)$ *such that if* $\Omega' \subset\subset \Omega$*, with* $\delta = \operatorname{dist}(\Omega', \partial\Omega)$*, then the following estimates hold:*

(i) If $f \in L^p(\Omega)$, $f_i \in L^q(\Omega)$ *and* $\alpha = 1 - \frac{d}{q} = 2 - \frac{d}{p} \in (0, 1)$*, then* $w \in C^\alpha(\Omega')$ *and:*

$$\|w\|_{C^\alpha(\Omega')} \leq C\delta^{-\frac{d}{2}+1-\alpha}\left(\|f\|_{L^p(\Omega)} + \sum_{i=1}^{d}\left\|f_j\right\|_{L^q(\Omega)} + \|w\|_{H^1(\Omega)}\right).$$

(ii) If $f \in L^p(\Omega)$, $\alpha = 1 - \frac{d}{p} \in (0, 1)$ *and* $f_i \in C^\alpha(\Omega)$*, then* $\nabla w \in C^\alpha(\Omega')$ *and:*

$$\|\nabla w\|_{C^\alpha(\Omega')} \leq C\delta^{-\frac{d}{2}-\alpha}\left(\|f\|_{L^p(\Omega)} + \sum_{i=1}^{d}\|f_i\|_{C^\alpha(\Omega)} + \|w\|_{H^1(\Omega)}\right).$$

References

1. G. Allaire, Homogenization and two-scale convergence. SIAM J. Math. Anal. **23**(6), 1482–1518 (1992). https://doi.org/10.1137/0523084
2. S. Armstrong, T. Kuusi, J.C. Mourrat, *Quantitative Stochastic Homogenization and Large-Scale Regularity, Grundlehren Der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, vol. 352 (Springer, Cham, 2019). https://doi.org/10.1007/978-3-030-15545-2
3. M. Avellaneda, F.H. Lin, Compactness methods in the theory of homogenization. Commun. Pure Appl. Math. **40**(6), 803–847 (1987). https://doi.org/10.1002/cpa.3160400607
4. M. Avellaneda, F.H. Lin, Compactness methods in the theory of homogenization. II. Equations in nondivergence form. Commun. Pure Appl. Math. **42**(2), 139–172 (1989). https://doi.org/10.1002/cpa.3160420203
5. L. Berlyand, Y. Gorb, A. Novikov, Fictitious fluid approach and anomalous blow-up of the dissipation rate in a two-dimensional model of concentrated suspensions. Arch. Ration. Mech. Anal. **193**(3), 585–622 (2009). https://doi.org/10.1007/s00205-008-0152-2
6. L. Berlyand, E. Khruslov, Homogenized non-Newtonian viscoelastic rheology of a suspension of interacting particles in a viscous Newtonian fluid. SIAM J. Appl. Math. **64**(3), 1002–1034 (2004). https://doi.org/10.1137/S0036139902403913
7. H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations* Universitext (Springer, New York, 2011)
8. L.A. Caffarelli, I. Peral, On $\mathrm{W}^{1,\mathrm{p}}$-estimates for elliptic equations in divergence form. Commun. Pure Appl. Math. **51**(1), 1–21 (1998). https://doi.org/10.1002/(SICI)1097-0312(199801)51:1<1::AID-CPA1>3.3.CO;2-N
9. D. Cioranescu, P. Donato, *An Introduction to Homogenization, Oxford Lecture Series in Mathematics and Its Applications*, vol. 17 (The Clarendon Press, Oxford University Press, New York, 1999)
10. T. Dang, Y. Gorb, S. Jiménez Bolaños, Homogenization of nondilute suspension of viscous fluid with magnetic particles. SIAM J. Appl. Math. **81**(6), 2547–2568 (2021). https://doi.org/10.1137/21M1413833
11. P.A. Davidson, *An Introduction to Magnetohydrodynamics* (Cambridge University Press, Cambridge, 2001)
12. L. Desvillettes, F. Golse, V. Ricci, The mean-field limit for solid particles in a Navier-Stokes flow. J. Stat. Phys. **131**(5), 941–967 (2008). https://doi.org/10.1007/s10955-008-9521-3
13. M. Duerinckx, A. Gloria, On Einstein's effective viscosity formula (2020). arXiv:2008.03837 [math-ph]

14. M. Duerinckx, A. Gloria, Corrector equations in fluid mechanics: effective viscosity of colloidal suspensions. Arch. Ration. Mech. Anal. **239**(2), 1025–1060 (2021). https://doi.org/10.1007/s00205-020-01589-1
15. M. Duerinckx, A. Gloria, Quantitative homogenization theory for random suspensions in steady Stokes flow (2021). arXiv:2103.06414 [math]
16. M. Duerinckx, A. Gloria, Sedimentation of random suspensions and the effect of hyperuniformity (2021). arXiv:2004.03240 [math-ph]
17. G.A. Francfort, A. Gloria, O. Lopez-Pamies, Enhancement of elasto-dielectrics by homogenization of active charges. J. Math. Pures Appl. (9) **156**, 392–419 (2021). https://doi.org/10.1016/j.matpur.2021.10.002
18. D. Gérard-Varet, M. Hillairet, Analysis of the viscosity of dilute suspensions beyond Einstein's formula. Arch. Ration. Mech. Anal. **238**(3), 1349–1411 (2020). https://doi.org/10.1007/s00205-020-01567-7
19. M. Giaquinta, L. Martinazzi, *An Introduction to the Regularity Theory for Elliptic Systems, Harmonic Maps and Minimal Graphs, Appunti. Scuola Normale Superiore Di Pisa (Nuova Serie) [Lecture Notes. Scuola Normale Superiore Di Pisa (New Series)]*, vol. 11, 2nd edn. (Edizioni della Normale, Pisa, 2012). https://doi.org/10.1007/978-88-7642-443-4
20. D. Gilbarg, N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics (Springer, Berlin, 2001)
21. Y. Gorb, F. Maris, B. Vernescu, Homogenization for rigid suspensions with random velocity-dependent interfacial forces. J. Math. Anal. Appl. **420**(1), 632–668 (2014). https://doi.org/10.1016/j.jmaa.2014.05.015
22. M. Grüter, K.O. Widman, The Green function for uniformly elliptic equations. Manuscripta Math. **37**(3), 303–342 (1982). https://doi.org/10.1007/BF01166225
23. B.M. Haines, A.L. Mazzucato, A proof of Einstein's effective viscosity for a dilute suspension of spheres. SIAM J. Math. Anal. **44**(3), 2120–2145 (2012). https://doi.org/10.1137/100810319
24. Q. Han, F. Lin, *Elliptic Partial Differential Equations, Courant Lecture Notes in Mathematics*, vol. 1, 2nd edn. (Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2011)
25. R.M. Höfer, Sedimentation of inertialess particles in Stokes flows. Commun. Math. Phys. **360**(1), 55–101 (2018). https://doi.org/10.1007/s00220-018-3131-y
26. C.E. Kenig, W.M. Ni, On the elliptic equation Lu-k+K exp[2u]=0. Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **12**(2), 191–224 (1985)
27. C.E. Kenig, F. Lin, Z. Shen, Homogenization of elliptic systems with Neumann boundary conditions. J. Am. Math. Soc. **26**(4), 901–937 (2013). https://doi.org/10.1090/S0894-0347-2013-00769-9
28. Y.Y. Li, L. Nirenberg, Estimates for elliptic systems from composite material, in *Communications on Pure and Applied Mathematics*, vol. 56 (Wiley, 2003), pp. 892–925. https://doi.org/10.1002/cpa.10079
29. Y.Y. Li, M. Vogelius, Gradient estimates for solutions to divergence form elliptic equations with discontinuous coefficients. Arch. Ration. Mech. Anal. **153**(2), 91–151 (2000). https://doi.org/10.1007/s002050000082
30. W. Littman, G. Stampacchia, H.F. Weinberger, Regular points for elliptic equations with discontinuous coefficients. Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3) **17**, 43–77 (1963)
31. A. Mecherbet, Sedimentation of particles in Stokes flow. Kinet. Relat. Models **12**(5), 995–1044 (2019). https://doi.org/10.3934/krm.2019038
32. G. Nguetseng, A general convergence result for a functional related to the theory of homogenization. SIAM J. Math. Anal. **20**(3), 608–623 (1989). https://doi.org/10.1137/0520043
33. B. Niethammer, R. Schubert, A local version of Einstein's formula for the effective viscosity of suspensions. SIAM J. Math. Anal. **52**(3), 2561–2591 (2020). https://doi.org/10.1137/19M1251229
34. C. Prange, Uniform estimates in homogenization: compactness methods and applications. J. Equ. Dérivées Partielles 1–25 (2014). https://doi.org/10.5802/jedp.110

35. Z. Shen, $W^{1,p}$ estimates for elliptic homogenization problems in nonsmooth domains. Indiana Univ. Math. J. **57**(5), 2283–2298 (2008). https://doi.org/10.1512/iumj.2008.57.3344
36. Z. Shen, Boundary estimates in elliptic homogenization. Anal. PDE **10**(3), 653–694 (2017). https://doi.org/10.2140/apde.2017.10.653
37. Z. Shen, *Periodic Homogenization of Elliptic Systems, Operator Theory: Advances and Applications*, vol. 269 (Birkhäuser/Springer, Cham, 2018). https://doi.org/10.1007/978-3-319-91214-1
38. Q. Zhang, J. Cui, Interior Hölder and gradient estimates for the homogenization of the linear elliptic equations. Sci. China Math. **56**(8), 1575–1584 (2013). https://doi.org/10.1007/s11425-013-4645-6

On Static and Evolutionary Homogenization in Crystal Plasticity for Stratified Composites

Elisa Davoli and Carolin Kreisbeck

1 Introduction

Motivated by new trends in technology that require materials with nonstandard properties, the study of artificially engineered composites (metamaterials) has been the subject of an intense research activity at the triple point between mathematics, physics, and materials science.

Here, we investigate the effective deformation behavior of a special class of mechanical metamaterials exhibiting the following two features whose interplay generates a highly anisotropic material response: (1) the geometry of the heterogeneities is characterized by periodically alternating layers of two different components and (2) the material properties of the two components show strong differences. To be precise, we assume that one is rigid, while the other one is softer, allowing for large-strain elastoplastic deformations along prescribed slip directions.

The asymptotic analysis of variational models for such stratified materials with fully rigid components and complete adhesion between the phases was initiated by CHRISTOWIAK and KREISBECK in [5], which is the starting point for this work. More precisely, the subject of [5] is a two-dimensional homogenization problem in the context of finite elastoplasticity, with geometrically nonlinear but rigid elasticity, where the softer component can be deformed along a single active plastic slip system with linear self-hardening. At the core of the homogenization

E. Davoli (✉)
Institute of Analysis and Scientific ComputingTU Wien, Vienna, Austria
e-mail: elisa.davoli@tuwien.ac.at

C. Kreisbeck
Mathematisch-Geographische Fakultät, Katholische Universität Eichstätt-Ingolstadt, Eichstätt, Germany
e-mail: carolin.kreisbeck@ku.de

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_7

result via Γ-convergence [5, Theorem 1.1] lies the characterization of the weak closures of the set of admissible deformations via an asymptotic rigidity result. In [12], these techniques have been carried forward to a model for plastic composites without linear hardening in the spirit of [8]. This leads to a variational limit problem on the space of functions of bounded variation. Natural generalizations of these models to three (and higher) dimensions, where the material heterogeneities are either layers or fibers, are studied [6] and [15], respectively. Note that these two references, which are formulated in the context of nonlinear elasticity, use energy densities with p-growth for $1 < p < +\infty$, and consider also nontrivial elastic energies on the stronger components. This allows treating also very stiff (but not necessarily rigid) reinforcements. As for their mathematical structure, all the aforementioned papers feature energies of integral form, characterized by linear or superlinear growth, subject to non-convex differential constraints. For related work on alternative approaches to layered and fiber-reinforced high-contrast composites with different choices of scaling relations between the elastic constants, thickness, and adhesive parameters, see, e.g., [2, 4, 14, 28].

Our goal in this work is twofold. First, we provide an analysis of minimizers for the effective energy functional derived in [5] as a homogenized Γ-limit; in particular, we address the question of uniqueness and identify necessary conditions for minimizers. Second, for some specific case studies, we complement the static homogenization of [5] by an evolutionary Γ-convergence analysis. Generally speaking, evolutionary Γ-convergence aims at transferring the concept of limit passages in parameter-dependent stationary variational problems to time-dependent settings. For energetic rate-independent systems, such a theory was developed by Mielke, Roubíček, and Stefanelli in [24]. The main feature of Γ-convergence (see, e.g., [3, 10]) is to guarantee convergence of solutions. In other words, (almost) minimizers of the parameter-dependent functionals converge to minimizers of the limit functional. Analogously, evolutionary Γ-convergence for rate-independent systems implies that energetic solutions again converge to energetic solutions of the limit system. In cases where energetic solutions do not exist, which is the situation in this work, one can work instead with solutions to associated approximate incremental problems. For a comprehensive introduction to the topic, we refer to [23, Sections 2.3–2.5, Section 3.5.4]; for applications in linearized elastoplasticity, see, e.g., [17, 26] on homogenization or [25] on a rigorous justification through a rigorous linearization of finite-strain plasticity.

In order to describe our results in more detail, some notation needs to be introduced. Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain. Assume that Ω represents the reference configuration of a high-contrast material with bilayered microstructure encoded by the alternation of two horizontal layers, a soft and a rigid one, see Fig. 1. Without loss of generality, we can assume that

$$\int_{\Omega} x \, \mathrm{d}x = 0, \tag{1}$$

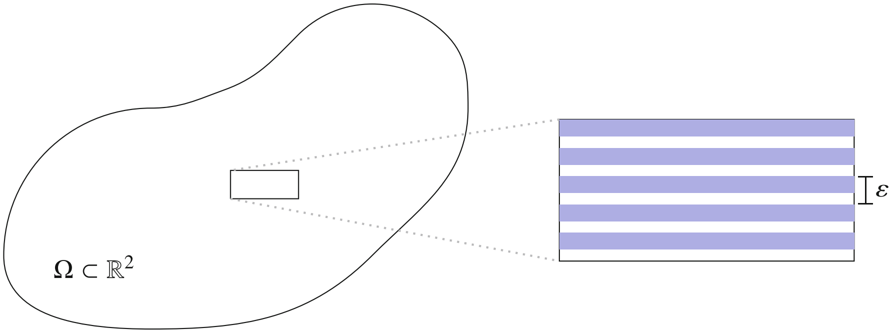


Fig. 1 Illustration of the reference configuration Ω and the stratified microstructure at length-scale ϵ

namely that the barycenter of Ω lies in the origin. To mathematically describe the geometry of the heterogeneities, consider the periodicity cell $Y := [0, 1)^2$, which we subdivide into $Y = Y_{\text{soft}} \cup Y_{\text{rig}}$ with $Y_{\text{soft}} := [0, 1) \times [0, \lambda)$ for $\lambda \in (0, 1)$ and $Y_{\text{rig}} := Y \setminus Y_{\text{soft}}$. All sets are extended by periodicity to $\mathbb{R}^2$. The (small) parameter $\varepsilon > 0$ describes the thickness of a pair (one rigid, one softer) of fine layers and can be viewed as the intrinsic periodicity scale of the microstructure. The collection of all rigid and soft layers in Ω corresponds to the sets $\varepsilon Y_{\text{rig}} \cap \Omega$ and $\varepsilon Y_{\text{soft}} \cap \Omega$, respectively.

Regarding the material properties, the body as a whole exhibits an elastoplastic behavior characterized by finite single-slip crystal plasticity with rigid elasticity throughout. The deformations on the individual rigid layers instead are restricted to global rotations or translations. We point out that several different models of finite elastoplasticity have been proposed and analyzed in the literature, see, e.g., [18] for a general introduction; recent contributions include an analysis of the incompatibility tensor [1], a formulation keeping track of frame invariance of intermediate configurations [16], as well as discussions of different (multiplicative) decompositions of deformation gradients [11, 13, 29]. Here, we adopt the classical approach introduced in [19, 21]: the gradient of every deformation $u : \Omega \to \mathbb{R}^2$ decomposes into the product of an elastic strain, F_{el}, and a plastic one, F_{pl}, satisfying

$$\nabla u = F_{\text{el}} F_{\text{pl}}. \tag{2}$$

Due to the assumption that the elastic behavior of the body is purely rigid, one has

$$F_{\text{el}} \in SO(2) \qquad \text{almost everywhere (a.e.) in } \Omega,$$

where $SO(2)$ denotes the set of rotations in $\mathbb{R}^2$. As for the plastic part, the presence of a single active slip, with slip direction $s \in \mathcal{S}^1 := \{s \in \mathbb{R}^2 : |s| = 1\}$ and associated slip plane normal $m = s^{\perp}$, translates into

$$F_{\mathrm{pl}} = \mathrm{Id} + \gamma s \otimes m, \tag{3}$$

where the shear coefficient γ measures the amount of slip. While the material is free to glide along the slip system in the softer phase, it is required that γ vanishes on the layers consisting of the rigid material, i.e., $\gamma = 0$ in $\varepsilon Y_{\mathrm{rig}} \cap \Omega$.

Collecting these modeling assumptions, we define, for $\varepsilon > 0$, the class $\mathcal{A}_\varepsilon^{(s)}$ of admissible deformations as

$$\begin{aligned}\mathcal{A}_\varepsilon^{(s)} := \{u \in \mathcal{Y} : \nabla u = R_u(\mathrm{Id} + \gamma_u s \otimes m) \text{ a.e. in } \Omega,&\\ R_u \in L^\infty(\Omega; SO(2)), \gamma_u \in L^2(\Omega), \gamma_u = 0 \text{ a.e. in } \varepsilon Y_{\mathrm{rig}} \cap \Omega\},&\end{aligned} \tag{4}$$

where $\mathcal{Y} = W^{1,2}(\Omega;\mathbb{R}^2) \cap L_0^2(\Omega;\mathbb{R}^2)$, and $L_0^2(\Omega;\mathbb{R}^2) = \{u \in L^2(\Omega;\mathbb{R}^2) : \int_\Omega u \, \mathrm{d}x = 0\}$ denotes the space of L^2-functions with zero average in Ω. In light of the completely rigid behavior of the body in $\varepsilon Y_{\mathrm{rig}} \cap \Omega$ and the plastic deformation behavior along a single-slip system, for which we assume linear hardening, in $\varepsilon Y_{\mathrm{soft}} \cap \Omega$, the stored elastoplastic energy of a deformation $u \in \mathcal{Y}$ reads as

$$E_\varepsilon^{(s)}(u) = \begin{cases} \displaystyle\int_{\varepsilon Y_{\mathrm{soft}} \cap \Omega} |\gamma_u|^2 \, \mathrm{d}x & \text{if } u \in \mathcal{A}_\varepsilon^{(s)}, \\ +\infty & \text{otherwise.} \end{cases} \tag{5}$$

Notice that a model for homogeneous materials with the properties of the softer component above has been studied in [7, 9]. We refer the reader to [8] for a corresponding analysis in the absence of hardening.

In order to state the homogenization result from [5], let us introduce, for any slip direction $s = (s_1, s_2) \in \mathcal{S}^1$, the following class of deformations:

$$\begin{aligned}\mathcal{A}^{(s)} := \{u \in \mathcal{Y} : \nabla u = R_u(\mathbb{I} + \gamma_u e_1 \otimes e_2),&\\ R_u \in SO(2), \gamma_u \in L^2(\Omega), \gamma_u \in K^{(s)} \text{ a.e. in } \Omega\},&\end{aligned} \tag{6}$$

with

$$K^{(e_2)} := \{0\}, \;\; K^{(e_1)} := \mathbb{R}, \;\text{ and } K^{(s)} := \begin{cases} [0, -2\lambda \frac{s_1}{s_2}] & \text{if } s_1 s_2 < 0, \\ [-2\lambda \frac{s_1}{s_2}, 0] & \text{if } s_1 s_2 > 0, \end{cases} \quad \text{for } s \notin \{e_1, e_2\}. \tag{7}$$

Throughout the chapter, we will often use—without further mention—one of the following alternative representations of the set in (6) (see [5]), that is,

$$\mathcal{A}^{(s)} = \{u \in \mathcal{Y} : u(x) = R_u(x + \Gamma_u(x)e_1) \text{ for } x \in \Omega,\ \Gamma_u \in W^{1,2}(\Omega) \cap L^2_0(\Omega),$$

$$R_u \in \mathrm{SO}(2),\ \partial_1\Gamma_u = 0,\ \partial_2\Gamma_u \in K^{(s)} \text{ a.e. in } \Omega\} \tag{8}$$

$$= \{u \in \mathcal{Y} : \nabla u \in \mathcal{M}^{(s)},\ \gamma_u \in K^{(s)} \text{ a.e. in } \Omega\}, \tag{9}$$

where γ_u is defined as in (6), and

$$\mathcal{M}^{(s)} := \{F \in \mathbb{R}^{2\times 2} : \det F = 1, |Fs| = 1\}. \tag{10}$$

We observe that if $u \in \mathcal{A}^{(s)}$, then $\operatorname{curl}\nabla u = 0$ implies $\partial_1\gamma_u = 0$. Hence, the set $\mathcal{A}^{(s)}$ is intrinsically one-dimensional.

With this notation in place, we can now formulate [5, Theorem 1.1], which characterizes the macroscopic material response of the considered stratified high-contrast composites in terms of Γ-convergence: the Γ-limit as $\epsilon \to 0$ of the energies $(E^{(s)}_{\varepsilon})_\epsilon$ in the weak $W^{1,2}$-topology is given for $u \in \mathcal{Y}$ by the functional

$$E^{(s)}(u) := \begin{cases} \dfrac{{s_1}^2}{\lambda}\displaystyle\int_\Omega \gamma_u^2 \,\mathrm{d}x - 2s_1s_2\int_\Omega \gamma_u \,\mathrm{d}x & \text{for } u \in \mathcal{A}^{(s)}, \\ +\infty & \text{otherwise.} \end{cases} \tag{11}$$

As Γ-convergence is invariant under continuous perturbations, the previous result is not affected by adding external loading to the stored energy functionals $E^{(s)}_\epsilon$ in (5). If we augment $E^{(s)}$ with a term describing work due to a body force with density $g \in L^2(\Omega;\mathbb{R}^2)$, the functional to study is

$$I^{(s)}_g(u) := E^{(s)}(u) - \int_\Omega g \cdot u \,\mathrm{d}x \tag{12}$$

for $u \in \mathcal{Y}$.

The essence of our first main result can then be summarized in simple terms as follows:

Conclusion (Uniqueness of Rotations and Shear Coefficients) For any slip direction $s \in \mathcal{S}^1$ and any applied load $g \in L^2(\Omega;\mathbb{R}^2)$, either the rotation or the shear coefficient associated with minimizers of $I^{(s)}_g$ is uniquely determined. □

We refer to Sect. 2 for the precise assumptions, as well as to Lemma 1 and Proposition 1 for the formulation of the statement and the proof of this result. Besides this general observation about uniqueness, we also discuss necessary and sufficient conditions for minimizers under specific conditions on the slip directions and the applied loads, including criteria for trivial deformation behavior in the form of rigid-body motions.

In the second part of this chapter (see Sect. 3), we expand the homogenization of the introduced static model to a quasistatic context by incorporating time-dependent loadings and dissipation acting on the shear variable. To be precise, we work with dissipations $\mathcal{D}$ that are given as the difference of the shear coefficients measured in the L^1-norm. Our second main contribution regards the asymptotic analysis of such extended models in the framework of evolutionary Γ-convergence for energetic rate-independent systems (see [23], as well as the beginning of Sect. 3, where we give a brief outline of this theory adapted to the setting of this chapter). An intuitive, rough version of our findings reads the following:

Conclusion (Evolutionary* $\boldsymbol{\Gamma}$*-Convergence) Suppose that the slip direction is aligned or orthogonal to the orientation of the material microstructure, meaning $s = e_1$ or $s = e_2$. Considering a family of rate-independent systems with (suitably regularized versions of) the stored energies $E_\varepsilon^{(s)}$ from (5) and dissipation distance $\mathcal{D}$, it follows that energetic and dissipative effects decouple in the limit $\epsilon \to 0$. Moreover, (approximate) solutions (to the associated time-discrete incremental problems) of the ε-dependent systems converge to energetic solutions of a system involving the stored energy $E^{(s)}$ from (11) and $\mathcal{D}$.

If $s = e_1$, the limiting system is very restrictive and corresponds to a purely energetic evolution without any dissipation, so that one can speak of a loss of dissipation through homogenization. □

The precise formulation of Conclusion 1 for $s = e_2$ and $s = e_1$ can be found in Theorems 3 and 4, respectively. In both cases, the proof strategy relies on the well-established scheme in [23, Section 2.3–2.5]. The only delicate point is the construction of a so-called mutual recovery sequence, for which we utilize tailored arguments that keep track of the special geometry of the problem.

This chapter is organized as follows: Sect. 2 is entirely devoted to the study of the static minimization problem, while Sect. 3 deals with evolutionary Γ-convergence.

1.1 Notation

Throughout the manuscript, $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^2$, and $\mathcal{S}^1$ is the unit sphere in $\mathbb{R}^2$, i.e., $\mathcal{S}^1 := \{s \in \mathbb{R}^2 : |s| = 1\}$. For $a \in \mathbb{R}^2$ with $a \neq 0$, we use the shorthand notation $\bar{a} = a/|a|$. We take e_1 and e_2 as the standard unit vectors in $\mathbb{R}^2$ and define $a^\perp := a_1e_2 - a_2e_1 \in \mathbb{R}^2$ for any $a = (a_1, a_2) \in \mathbb{R}^2$. Analogously, for functions $g : \Omega \to \mathbb{R}^2$, the map $g^\perp$ is given as $g^\perp(x) = g(x)^\perp$ for all $x \in \Omega$. Given $v_1, v_2 \in \mathbb{R}^2$, we denote by $(v_1|v_2)$ the 2×2 matrix having these vectors as first and second columns, respectively. For the trace of a matrix $A \in \mathbb{R}^{2\times 2}$, we write $\operatorname{Tr} A$, while $\mathrm{id} : \mathbb{R}^2 \to \mathbb{R}^2$ denotes the identity map and $\mathbb{I}$ its differential. Furthermore, $\mathbb{1}_U : \mathbb{R}^2 \to \{0, 1\}$ is the indicator function for a set $U \subset \mathbb{R}^2$, i.e., $\mathbb{1}_U(x) = 1$ if and only if $x \in U$. Our notation for the dual of a vector space $\mathcal{Y}$ is $\mathcal{Y}'$, and the corresponding duality pairing is denoted by $\langle\cdot,\cdot\rangle_{\mathcal{Y}',\mathcal{Y}}$ or simply by $\langle\cdot,\cdot\rangle$. As for L^p- and Sobolev spaces, we follow the classical notational conventions, indicating explicitly the target space other than $\mathbb{R}$, we write, e.g., $L^2(\Omega;\mathbb{R}^2)$ and $W^{1,2}(\Omega) = W^{1,2}(\Omega;\mathbb{R})$. Finally, when speaking of the convergence of a family $(a_\epsilon)_\epsilon$ with a continuous parameter $\epsilon > 0$, we actually mean that each of the sequences $(a_{\epsilon_j})_j$ with $\epsilon_j \searrow 0$ converges for $j \to +\infty$ to the same limit.

2 Minimizers of the Static Homogenized Limit Problem

In this section, we analyze, for every $g \in L^2(\Omega; \mathbb{R}^2)$ and $s \in \mathcal{S}^1$, the variational problem

$$\text{minimize} \quad I_g^{(s)}(u) = E^{(s)}(u) - \int_\Omega g \cdot u \,\mathrm{d}x \qquad \text{for } u \in \mathcal{Y}. \tag{13}$$

The existence of solutions to (13) follows from the direct method in the calculus of variations. Indeed, observing that $I_g^{(s)}$ is the Γ-limit resulting from the homogenization procedure studied in [5], along with the fact that the term $u \mapsto -\int_\Omega g \cdot u \,\mathrm{d}x$ constitutes a continuous perturbation, yields the lower semicontinuity of the functional $I_g^{(s)}$ regarding the weak topology of $W^{1,2}(\Omega; \mathbb{R}^2)$. The sequential weak compactness of the sublevel sets of $I_g^{(s)}$ follows from the quadratic growth of $I_g^{(s)}$ with respect to $\|\gamma_u\|_{L^2(\Omega)}$ in combination with the special structure of $\mathcal{A}^{(s)}$, which contains only globally rotated shear deformations, see (6).

The aim of this section is to address the issues of uniqueness of solutions to (13) and their explicit characterization. We start by introducing some auxiliary quantities for our analysis: given $\Gamma \in L^2(\Omega)$ and $g \in L^2(\Omega; \mathbb{R}^2)$, let

$$\Lambda(g, \Gamma) := \widehat{g} + \int_\Omega \Gamma g \,\mathrm{d}x$$

with

$$\widehat{g} := \int_\Omega x_1 g(x) - x_2 g^\perp(x) \,\mathrm{d}x = \int_\Omega \widehat{G}(x) x \,\mathrm{d}x \tag{14}$$

and $\widehat{G}(x) := (g(x)| - g^\perp(x))$ for $x \in \Omega$.

Given this terminology, one obtains for $u \in \mathcal{A}^{(s)}$ with $u = R_u(\mathrm{id} + \Gamma_u e_1)$, cf. (8), that

$$\int_\Omega g \cdot u \,\mathrm{d}x = \int_\Omega x_1 g \cdot R_u e_1 + x_2 g \cdot (R_u e_1)^\perp + \Gamma_u g \cdot R_u e_1 \,\mathrm{d}x = \Lambda(g, \Gamma_u) \cdot R_u e_1. \tag{15}$$

Motivated by this observation, we can set up a variational problem that is equivalent to solving (13) and involves only the shear variable Γ. This reformulation is made precise in the following lemma.

Lemma 1 *Let* $J_g^{(s)} : W^{1,2}(\Omega) \cap L^2_0(\Omega) \to \mathbb{R}_\infty := \mathbb{R} \cup \{+\infty\}$ *be given by*

$$J_g^{(s)}(\Gamma) := \begin{cases} \dfrac{{s_1}^2}{\lambda} \displaystyle\int_\Omega (\partial_2 \Gamma)^2 \, \mathrm{d}x - 2 s_1 s_2 \int_\Omega \partial_2 \Gamma \, \mathrm{d}x - |\Lambda(g, \Gamma)| & \textit{if } \partial_1 \Gamma = 0, \\ & \partial_2 \Gamma \in K^{(s)} \textit{ a.e. in } \Omega, \\ +\infty & \textit{otherwise.} \end{cases}$$

Then, $u \in \mathcal{A}^{(s)}$ *is a minimizer of* $I_g^{(s)}$ *if and only if* Γ_u *is a minimizer of* $J_g^{(s)}$ *and*

$$\begin{cases} R_u e_1 = \overline{\Lambda(g, \Gamma_u)} & \textit{if } \Lambda(g, \Gamma_u) \neq 0, \\ R_u \in SO(2) & \textit{if } \Lambda(g, \Gamma_u) = 0. \end{cases} \tag{16}$$

Moreover, $\min_{u \in \mathcal{Y}} I_g^{(s)}(u) = \min_{\Gamma \in W^{1,2}(\Omega) \cap L^2_0(\Omega)} J_g^{(s)}(\Gamma)$.

Proof First, we verify via a simple application of the direct method that minimizers of $J_g^{(s)}$ exist. To check for coercivity, we argue in the case $s \neq e_1$ that the domain of $J_g^{(s)}$ is a bounded subset of $W^{1,2}(\Omega)$ and thus precompact in the weak topology of $W^{1,2}(\Omega)$. If $s = e_1$, consider a minimizing sequence $(\Gamma_n)_n$ for $J_g^{(e_1)}$ such that for every $n \in \mathbb{N}$,

$$J_g^{(e_1)}(\Gamma_n) \leq \inf_{\Gamma \in W^{1,2}(\Omega) \cap L^2_0(\Omega)} J_g^{(e_1)}(\Gamma) + \frac{1}{n},$$

and hence, $\partial_1 \Gamma_n = 0$ and

$$\frac{1}{\lambda} \int_\Omega (\partial_2 \Gamma_n)^2 \, \mathrm{d}x \leq -|\Lambda(g, 0)| + \frac{1}{n} + |\widehat{g}| + \|\Gamma_n\|_{L^2(\Omega)} \|g\|_{L^2(\Omega;\mathbb{R}^2)},$$

so that $(\Gamma_n)_n$ is bounded in $W^{1,2}(\Omega)$ by the Poincaré–Wirtinger inequality and therefore has a weakly converging subsequence in $W^{1,2}(\Omega)$. Finally, the existence of a minimizer follows as the functionals $J_g^{(s)}$ are all lower semicontinuous with respect to the weak topology in $W^{1,2}(\Omega)$.

To prove the statement, we start with the preliminary observation that by (15),

$$I_g^{(s)}(u) = J_g^{(s)}(\Gamma_u) - \Lambda(g, \Gamma_u) \cdot R_u e_1 + |\Lambda(g, \Gamma_u)| \tag{17}$$

for every $u \in \mathcal{A}^{(s)}$ with $u = R_u(\mathrm{id} + \Gamma_u e_1)$, see (8).

Now, let u^* be a minimizer of $I_g^{(s)}$ in $\mathcal{A}^{(s)}$ with $u^* = R^*(\mathrm{id} + \Gamma^* e_1)$. If we consider $u = R(\mathrm{id} + \Gamma^* e_1)$ for different $R \in SO(2)$, then (16) is a direct consequence of the optimality of u^* and of (17). Choosing a competitor $u = R_\Gamma(\mathrm{id} + \Gamma e_1)$ with $R_\Gamma e_1 = \overline{\Lambda(g, \Gamma)}$ if $\Lambda(g, \Gamma) \neq 0$ and $R_\Gamma = \mathbb{I}$ otherwise, we infer from the minimality of u^* for $I_g^{(s)}$ in combination with (17) that Γ^* is minimal for $J_g^{(s)}$.

Conversely, let Γ^* be a minimizer of $J_g^{(s)}$, and let R^* be such that (16) holds true. Then, the minimality of the map $u^* = R^*(\mathrm{id} + \Gamma^* e_1)$ for $I_g^{(s)}$ follows once again from (17). □

Remark 1

(a) Notice that while the values of the energy functionals $I_g^{(s)}$ depend quadratically and linearly on γ_v, the dependence of $J_g^{(s)}$ on its domain involves a combination of convex and concave terms.

(b) In the case $s = e_2$, the situation is particularly simple. As the class of admissible functions does not include any shear deformations and is thus very restrictive, the problem comes down to analyzing rigid-body motion in response to external loading. In other words, it reduces to discussing the standard model of nonlinear rigid elasticity. To be precise, since $\gamma_u = 0$ and $\Gamma_u = 0$ for any $u \in \mathcal{A}^{(e_2)}$, and therefore, $\Lambda(g, \Gamma_u) = \widehat{g}$, we conclude from Lemma 1 that u is a minimizer of $I_g^{(e_2)}$ if and only if

$$\begin{cases} R_u e_1 = \widehat{g}/|\widehat{g}| & \text{if } \widehat{g} \neq 0, \\ R_u \in SO(2) & \text{otherwise.} \end{cases}$$

In the case of non-unique rotations for minimizers of $I_g^{(s)}$, we can prove a partial uniqueness result for the shears Γ. In combination with Lemma 1, it shows that at least one of the building blocks of a minimizer, that is, rotation or shear, is uniquely determined.

Proposition 1 *Let $u, w \in \mathcal{A}^{(s)}$ be two minimizers of $I_g^{(s)}$. If $\Lambda(g, \Gamma_w) = 0$, then $\gamma_u = \gamma_w$, or equivalently, $\Gamma_u = \Gamma_w$.*

Proof For $\delta \in (0, 1)$, consider the map $z_\delta := (1 - \delta)u + \delta R_u R_w^T w$, which satisfies $z_\delta \in \mathcal{A}^{(s)}$ with $R_{z_\delta} = R_u$ and $\gamma_{z_\delta} = (1 - \delta)\gamma_u + \delta\gamma_w$, cf. (6). Then,

$$\begin{aligned} 0 &\leq I_g^{(s)}(z_\delta) - (1 - \delta) I_g^{(s)}(u) - \delta I_g^{(s)}(w) \\ &= (\delta^2 - \delta)\frac{s_1^2}{\lambda} \int_\Omega (\gamma_u - \gamma_w)^2 \, \mathrm{d}x - \delta \int_\Omega g \cdot (R_u R_w^T w - w) \, \mathrm{d}x. \end{aligned}$$

We observe that $R_u R_w^T w - w = (R_u - R_w)(\mathrm{id} + \Gamma_w e_1)$, so that

$$\begin{aligned} \int_\Omega g \cdot (R_u R_w^T - w) \, \mathrm{d}x &= \Big(\int_\Omega x_1 g(x) - x_2 g^\perp(x) + \Gamma_w(x) g(x) \, \mathrm{d}x \Big) \cdot (R_u - R_w) e_1 \\ &= \Lambda(g, \Gamma_w) \cdot (R_u e_1 - R_w e_1) = 0, \end{aligned}$$

in view of the assumption $\Lambda(g, \Lambda_w) = 0$. Thus,

$$0 \geq \frac{s_1^2}{\lambda} \int_\Omega (\gamma_u - \gamma_w)^2 \, \mathrm{d}x \geq 0, \tag{18}$$

which shows that $\gamma_u = \gamma_w$. □

In what follows, we specify our discussion to the case of a square-shaped reference configuration $\Omega = (-1, 1)^2$, which allows for a refined analysis. Indeed, in this setting, one can decouple the two spatial variables and view Γ_u and γ_u for $u \in \mathcal{A}^{(s)}$ as functions of one variable, i.e., as elements of $W^{1,2}(-1, 1)$ and $L^2(-1, 1)$, respectively. Moreover,

$$\Lambda(g, \Gamma) = \widehat{g} + \int_{-1}^{1} \Gamma(x_2)[g]_{x_1}(x_2) \, \mathrm{d}x_2 \quad \text{and}$$

$$\widehat{g} = \int_{-1}^{1} x_1 [g]_{x_2}(x_1) \, \mathrm{d}x_1 - \int_{-1}^{1} x_2 [g]_{x_1}^{\perp}(x_2) \, \mathrm{d}x_2$$

with $[g]_{x_1} := \int_{-1}^{1} g(x_1, \cdot) \, \mathrm{d}x_1$ and $[g]_{x_2} := \int_{-1}^{1} g(\cdot, x_2) \, \mathrm{d}x_2$.

Remark 2 (Examples of $\widehat{g}$ *for* $\Omega = (-1, 1)^2$*)*

(a) If g describes linear loading, i.e., $g(x) := Ax + b$ for $x \in \Omega$ with given (nontrivial) $A \in \mathbb{R}^{2\times 2}$ and $b \in \mathbb{R}^2$, a direct calculation shows that

$$\widehat{g} = \frac{4}{3} \begin{pmatrix} \operatorname{Tr} A \\ A_{12} - A_{21} \end{pmatrix}.$$

If A is symmetric with nonzero trace, then $\widehat{g}/|\widehat{g}| = e_1$, whereas $\widehat{g}/|\widehat{g}| = e_2$ for a skew-symmetric A.

(b) When g is constant, then trivially, $\widehat{g} = 0$. In that case, also $\Lambda(g, \Gamma) = 0$ for any admissible Γ due to the fact that $\Gamma \in L_0^2(-1, 1)$ has vanishing mean value. Besides, in light of the property that Ω has its barycenter in the origin, $\widehat{g} = 0$ if g is of the form

$$g(x) = \sum_{i=1}^{N} A_i \begin{pmatrix} x_1^{\alpha_i} x_2^{\beta_i} \\ x_1^{\gamma_i} x_2^{\delta_i} \end{pmatrix}, \quad x \in \Omega,$$

with $A_i \in \mathbb{R}^{2\times 2}$ and $\alpha_i, \beta_i, \gamma_i, \delta_i$ odd integers for $i = 1, \ldots, N \in \mathbb{N}$.

Under special assumptions on the loads or the slip systems, we can identify further conditions on minimizers of $J_g^{(s)}$, and hence also of $I_g^{(s)}$, as the next results illustrate.

Proposition 2 *Let* $\Omega = (-1,1)^2$*, and suppose that* $[g]_{x_1}$ *is constant. If* $\Gamma \in W^{1,2}(-1,1) \cap L^2_0(-1,1)$ *is a minimizer of* $J^{(s)}_g$*, then* $\Gamma = 0$*. Every minimizer of* $I^{(s)}_g$ *is a rigid-body motion.*

Proof The case $s = e_2$ is covered in Remark 1. Assume now that $s \neq e_2$. If $[g]_{x_1}$ is constant, then $\Lambda(g,\Gamma) = \widehat{g}$, and thus, $\Lambda(g,\Gamma)$ is independent of Γ. The functional $J^{(s)}_g$ is then strictly convex with a unique minimizer Γ satisfying

$$\Gamma' \in \operatorname{argmin}_{\gamma \in L^2(-1,1;K^{(s)})} \int_{-1}^{1} \frac{s_1^2}{\lambda}\gamma^2 - 2s_1s_2\gamma \, \mathrm{d}x_2,$$

where Γ' denotes the weak derivative of Γ. By Jensen's inequality, this implies that Γ' is constant with

$$\Gamma' \in \operatorname{argmin}_{\eta \in K^{(s)}} \frac{s_1^2}{\lambda}\eta^2 - 2s_1s_2\eta = \{0\}; \tag{19}$$

for the last identity, we make use of the fact that the vertex of the parabola in (19), that is, $\eta = \frac{s_2}{s_1}\lambda$, does not lie in $K^{(s)}$ for $s \neq e_1$.

Summing up, we have shown that $\Gamma' = 0$, and therefore $\Gamma = 0$. The second statement is then an immediate consequence of Lemma 1. □

For $s = e_1$, we have the following necessary condition for minimizers of $J^{(e_1)}_g$.

Proposition 3 *Let* $\Omega = (-1,1)^2$*, and let* $\Gamma \in W^{1,2}(-1,1) \cap L^2_0(-1,1)$ *be a critical point of* $J^{(e_1)}_g$*.*

(a) If $\Lambda(g,\Gamma) = 0$*, then* $\Gamma = 0$*.*
(b) If $\Lambda(g,\Gamma) \neq 0$*, the differential equation*

$$\Gamma'' = -\frac{\lambda}{4}\overline{\Lambda(g,\Gamma)} \cdot \Big([g]_{x_1} - \frac{1}{2}\int_{-1}^{1} [g]_{x_1} \, \mathrm{d}x_2\Big) \tag{20}$$

holds in the sense of distributions. In particular, $\Gamma \in W^{2,2}(-1,1)$*.*

Proof We first prove (a). Let $\psi \in C^\infty_c(-1,1)$. For any $\delta \in \mathbb{R}$, we define the variation

$$\Gamma_\delta = \Gamma + \delta\Psi,$$

where Ψ is the primitive of ψ with vanishing mean value. To obtain optimality conditions for the minimizer Γ, we calculate that

$$\begin{aligned}0 = \frac{d}{d\delta}_{|\delta=0} J_g^{(e_1)}(\Gamma_\delta) &= \frac{d}{d\delta}_{|\delta=0} \frac{2}{\lambda}\int_{-1}^{1} (\Gamma' + \delta\psi)^2 \,\mathrm{d}x_2 - |\Lambda(g, \Gamma + \delta\Psi)| \\ &= \int_{-1}^{1} \frac{4}{\lambda}\Gamma'\psi \,\mathrm{d}x_2 - \Big(\int_{-1}^{1} [g]_{x_1}\Psi \,\mathrm{d}x_2\Big) \cdot \overline{\Lambda(g,\Gamma)} = \int_{-1}^{1} \frac{4}{\lambda}\Gamma'\psi \,\mathrm{d}x_2.\end{aligned}$$

This implies that $\Gamma' = 0$ and concludes the proof in view of the fact that $\Gamma \in L_0^2(-1, 1)$.

Next, we prove (b). For any $\Psi \in C_c^\infty(-1, 1)$, consider the variation

$$\Gamma_\delta = \Gamma + \delta\Big(\Psi - \frac{1}{2}\int_{-1}^{1} \Psi \,\mathrm{d}x_2\Big).$$

The claim then follows from the computation

$$\begin{aligned}0 &= \frac{d}{d\delta}_{|\delta=0} J_g^{(e_1)}(\Gamma_\delta) \\ &= \frac{d}{d\delta}_{|\delta=0} \frac{2}{\lambda}\int_{-1}^{1} (\Gamma' + \delta\Psi')^2 \,\mathrm{d}x_2 - \Big|\Lambda\Big(g, \Gamma + \delta\Big(\Psi - \frac{1}{2}\int_{-1}^{1}\Psi \,\mathrm{d}x_2\Big)\Big)\Big| \\ &= \int_{-1}^{1} \frac{4}{\lambda}\Gamma'\Psi' \,\mathrm{d}x_2 - \Big(\int_{-1}^{1} [g]_{x_1} \cdot \Big(\Psi - \frac{1}{2}\int_{-1}^{1}\Psi \,\mathrm{d}x_2\Big) \,\mathrm{d}x_2\Big) \cdot \overline{\Lambda(g,\Gamma)} \\ &= \int_{-1}^{1} \frac{4}{\lambda}\Gamma'\Psi - \overline{\Lambda(g,\Gamma)} \cdot \Big([g]_{x_1} - \tfrac{1}{2}\int_{-1}^{1} [g]_{x_1} \,\mathrm{d}x_2\Big)\Psi \,\mathrm{d}x_2.\end{aligned}$$

Remark 3 Note that case (a) can only occur in a scenario where $\widehat{g} = 0$. If $[g]_{x_1}$ is constant in addition, one concludes from Proposition 3 the necessary condition $\Gamma'' = 0$ for critical points of $J_g^{(e_1)}$, which is in agreement with the characterization of minimizers in Proposition 2.

Remark 4 (Affine Loads) We discuss solutions to (20) in the special case of affine loads $g(x) = Ax + b$ for $x \in \Omega = (-1, 1)^2$ with $A \in \mathbb{R}^{2\times 2}$ and $b \in \mathbb{R}^2$. Then,

$$\begin{aligned}[g]_{x_1}(x_2) &= 2x_2Ae_2 + 2b \quad \text{for } x_2 \in (-1, 1), \\ [g]_{x_2}(x_1) &= 2x_1Ae_1 + 2b \quad \text{for } x_1 \in (-1, 1),\end{aligned}$$

so that $[g]_{x_1} - \frac{1}{2}\int_{-1}^{1}[g]_{x_1} \,\mathrm{d}x_2 = 2x_2Ae_2$, $\widehat{g} = \frac{4}{3}(Ae_1 - (Ae_2)^\perp)$, and

$$\Lambda(g, \Gamma) = \frac{4}{3}(Ae_1 - (Ae_2)^\perp) + 2Ae_2\int_{-1}^{1} x_2\Gamma(x_2) \,\mathrm{d}x_2.$$

In view of (20), any critical point for $J_g^{(e_1)}$ is a polynomial of third order, i.e.,

$$\Gamma(x_2) = \alpha x_2^3 + \beta x_2^2 + \gamma x_2 + \delta \text{ with coefficients } \alpha, \beta, \gamma, \delta \in \mathbb{R}.$$

From $\int_{-1}^{1} \Gamma(x_2)\, \mathrm{d}x_2 = 0$, we infer that $\delta = -\frac{\beta}{3}$. Since the right-hand side of (20) has null average in $(-1, 1)$, we conclude that $\beta = 0$ and thus

$$\Gamma(x_2) = \alpha x_2^3 + \gamma x_2.$$

By plugging this structure into the expression of $J_g^{(e_1)}$, the problem of finding minimizers for $J_g^{(e_1)}$ reduces to the following two-dimensional optimization problem:

$$\min_{(\alpha,\gamma)\in\mathbb{R}^2} \frac{4}{\lambda}\Big(\frac{9}{5}\alpha^2 + \gamma^2 + 2\alpha\gamma\Big) - \frac{4}{3}\Big|Ae_1 - (Ae_2)^{\perp} + Ae_2\Big(\frac{3}{5}\alpha + \gamma\Big)\Big|.$$

3 Homogenization via Evolutionary Γ-Convergence

Before stating our findings on evolutionary Γ-convergence, we introduce the necessary terminology and recall a few definitions and abstract results from [23, 24], adjusted to our setting, for a self-contained presentation and the readers' convenience. Let $\mathcal{Q} = \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{Y}, \mathcal{Z}$ are reflexive, separable Banach spaces endowed with the weak topology and $T > 0$. We write $q = (y, z) \in \mathcal{Q} = \mathcal{Y} \times \mathcal{Z}$. Furthermore, take an energy functional $\mathcal{E} : [0, T] \times \mathcal{Q} \to \mathbb{R}_\infty := \mathbb{R} \cup \{\infty\}$ of the form

$$\mathcal{E}(t, q) := E(q) - \langle l(t), y\rangle_{\mathcal{Y}',\mathcal{Y}}, \tag{21}$$

where $E : \mathcal{Q} \to \mathbb{R}_\infty$ and $l \in W^{1,1}(0, T; \mathcal{Y}')$, and let $\mathcal{D} : \mathcal{Z} \times \mathcal{Z} \to [0, +\infty]$ be a dissipation distance, i.e., $\mathcal{D}$ is definite and satisfies the triangle inequality. A triple $(\mathcal{Q}, \mathcal{E}, \mathcal{D})$ of such a state space, energy, and dissipation functional is referred to as an (energetic) rate-independent system.

A process $q : [0, T] \to \mathcal{Q}$ is called an energetic solution of the rate-independent system $(\mathcal{Q}, \mathcal{E}, \mathcal{D})$ if the global stability condition, i.e., $q(t) \in \mathcal{S}(t)$ with

$$\mathcal{S}(t) := \{q \in \mathcal{Q} : \mathcal{E}(t, q) < +\infty \text{ and } \mathcal{E}(t, q) \le \mathcal{E}(t, \tilde{q}) + \mathcal{D}(z, \tilde{z}) \text{ for all } \tilde{q} = (\tilde{y}, \tilde{z}) \in \mathcal{Q}\}, \tag{22}$$

and the energy balance

$$\mathcal{E}(t, q(t)) + \mathrm{Diss}_{\mathcal{D}}(z; [0, t]) = \mathcal{E}(0, q(0)) - \int_0^t \langle \dot{l}(\tau), y(\tau)\rangle\ \mathrm{d}\tau \tag{23}$$

hold for all $t \in [0, T]$, see [23, Definition 2.1.2]; here, $\dot{l}$ denotes the weak derivative of l with respect to time, and for $q = (y, z) \in \mathcal{Q}$,

$$\mathrm{Diss}_{\mathcal{D}}(z; [0, t]) := \sup\Big\{\sum_{i=1}^{N} \mathcal{D}(z(t_{j-1}), z(t_j)) : N \in \mathbb{N}, 0 = t_0 < t_1 < \ldots < t_N = t\Big\}.$$

Now, consider energy and dissipation functionals $\mathcal{E}_\epsilon$, $\mathcal{D}_\epsilon$ with parameter $\epsilon > 0$ and $\mathcal{E}_0$, $\mathcal{D}_0$ as introduced above (the same notations carry over as well, indicated by subscript ϵ and 0, respectively). We say that the family $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon)_\epsilon$ evolutionary Γ-converges (with well-prepared initial conditions) to $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_0)$ as $\epsilon \to 0$, formally

$$(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon) \xrightarrow{ev\text{-}\Gamma} (\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_0) \quad \text{as } \epsilon \to 0, \tag{24}$$

if energetic solutions to $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon)$ exist and if the limits of such solutions for $\epsilon \to 0$ are energetic solutions to $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_0)$, along with suitable convergences of the energetic and the dissipated contributions. To be precise, if $q_\epsilon : [0, T] \to \mathcal{Q}$ are energetic solutions to $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon)$ such that $q_\epsilon(t) \to q(t)$ in $\mathcal{Q}$ for all $t \in [0, T]$ and $\mathcal{E}_\epsilon(0, q_\epsilon(0)) \to \mathcal{E}_0(0, q(0))$ as $\epsilon \to 0$, then $q : [0, T] \to \mathcal{Q}$ is an energetic solution for $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_0)$, and it holds that

$$\begin{aligned} \mathcal{E}_\epsilon(t, q_\epsilon(t)) \to \mathcal{E}_0(t, q(t)) \quad & \text{for all } t \in [0, T], \\ \mathrm{Diss}_{\mathcal{D}_\epsilon}(z_\epsilon; [0, t]) \to \mathrm{Diss}_{\mathcal{D}_0}(z; [0, t]) \quad & \text{for all } t \in [0, T], \\ \langle \dot{l}_\epsilon(t), y_\epsilon(t)\rangle \to \langle \dot{l}_0(t), y(t)\rangle \quad & \text{for a.e. } t \in [0, T]. \end{aligned}$$

If energetic solutions to $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon)$ do not exist, one needs to employ a refined concept. In fact, we can fall back on the following related approach, based on approximate incremental problems, which automatically admit solutions. Based on [24, Section 4] on the relaxation of evolutionary problems, we state a generalized version suitable for parameter-dependent families of energies and dissipations, as mentioned also in [23, Section 2.5.1]. In the following, we say that $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon)_\epsilon$ approximately evolutionary Γ-converges to $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_0)$ and write

$$(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon) \xrightarrow{ev\text{-}\Gamma_{\mathrm{app}}} (\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_0) \quad \text{as } \epsilon \to 0, \tag{25}$$

if sequences of piecewise constant interpolants of approximate solutions to time-incremental problems for $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon)$ have subsequences that converge to energetic solutions of $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_0)$ in a suitable sense. To be precise, for $\epsilon > 0$, let

$$\mathcal{T}_\epsilon = \{0 = \tau_\epsilon^{(0)} < \tau_\epsilon^{(1)} < \ldots < \tau_\epsilon^{(N_\epsilon - 1)} < \tau_\epsilon^{(N_\epsilon)} = T\},$$

with $N_\epsilon \in \mathbb{N}$, be a family of partitions of $[0, T]$ with fineness

$$\nu(\mathcal{T}_\epsilon) = \max_{k=1,\dots,N_\epsilon} \tau_\epsilon^{(k)} - \tau_\epsilon^{(k-1)} \to 0 \text{ as } \epsilon \to 0,$$

and initial conditions $q_\epsilon^{(0)} \in \mathcal{Q}$ with $q_\epsilon^{(0)} \to q^{(0)}$ in $\mathcal{Q}$ and $\mathcal{E}_\epsilon(0, q_\epsilon^{(0)}) \to \mathcal{E}_0(0, q^{(0)}) \in \mathbb{R}$ as $\epsilon \to 0$. Consider the piecewise constant functions $q_\epsilon : [0, T] \to \mathcal{Q}$ on the partition $\mathcal{T}_\epsilon$ with

$$q_\epsilon(t) = \sum_{k=1}^{N_\epsilon} q_\epsilon^{(k)} \mathbb{1}_{[\tau_\epsilon^{(k-1)}, \tau_\epsilon^k)}, \tag{26}$$

where $q_\epsilon^{(k)} \in \mathcal{Q}$ for $k = 1, \dots, N_\epsilon$ are iteratively determined solutions to the approximate incremental problem

$$\mathcal{E}_\epsilon(\tau_\epsilon^{(k)}, q_\epsilon^{(k)}) + \mathcal{D}_\epsilon(q_\epsilon^{(k-1)}, q_\epsilon^{(k)}) \leq \inf_{\tilde{q} \in \mathcal{Q}} \left[\mathcal{E}_\epsilon(\tau_\epsilon^{(k)}, \tilde{q}) + \mathcal{D}_\epsilon(q_\epsilon^{(k-1)}, \tilde{q})\right] + \nu(\mathcal{T}_\epsilon)\delta_\epsilon^{(k)};$$

here, $\delta_\epsilon^{(k)} > 0$ satisfies $\sup_{\epsilon>0} \sum_{k=1}^{N_\epsilon} \delta_\epsilon^{(k)} < +\infty$. Then, (25) means that there exist a subsequence of $(q_\epsilon)_\epsilon$ (not relabeled) and an energetic solution $q : [0, T] \to \mathcal{Q}$ for $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_0)$ such that

$$\begin{aligned} z_\epsilon(t) &\to z(t) \quad \text{in } \mathcal{Z} && \text{for all } t \in [0, T], \\ \mathcal{E}_\epsilon(t, q_\epsilon(t)) &\to \mathcal{E}_0(t, q(t)) && \text{for all } t \in [0, T], \\ \mathrm{Diss}_{\mathcal{D}_\epsilon}(z_\epsilon; [0, t]) &\to \mathrm{Diss}_{\mathcal{D}_0}(z; [0, t]) && \text{for all } t \in [0, T], \\ \langle \dot{l}_\epsilon(t), y_\epsilon(t) \rangle &\to \langle \dot{l}_0(t), y(t) \rangle && \text{for a.e. } t \in [0, T]. \end{aligned}$$

Next, we collect a list of conditions on $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon)_\epsilon$ and $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_0)$, which have been shown to provide sufficient criteria for (24) and (25):

(H1) $\mathcal{D}_\epsilon = \mathcal{D}$ for all $\epsilon > 0$, where $\mathcal{D} : \mathcal{Z} \times \mathcal{Z} \to [0, +\infty]$ is a lower-semicontinuous quasi-distance, i.e., definite and satisfying the triangle inequality.
(H2) $E_\epsilon : \mathcal{Q} \to \mathbb{R}_\infty$ are lower semicontinuous for all $\epsilon > 0$.
(H3) $E_\epsilon(q) \geq C\|q\|_{\mathcal{Q}}^\alpha - c$ for all $q \in \mathcal{Q}$, $\epsilon > 0$ with $C, c > 0$ and some $\alpha > 1$, and $l_\epsilon \to l_0$ in $W^{1,1}(0, T; \mathcal{Y}')$.
(H4) $(E_\epsilon)_\epsilon$ Γ-converges to E_0 with respect to the topology of $\mathcal{Q}$.
(H5) If $(t_\epsilon, q_\epsilon)_\epsilon$ is a stable sequence, that is, $q_\epsilon \in \mathcal{S}_\epsilon(t_\epsilon)$ for all $\epsilon > 0$ and $\sup_{\epsilon>0} \mathcal{E}_\epsilon(t_\epsilon, q_\epsilon) < +\infty$, such that $(t_\epsilon, q_\epsilon) \to (t, q)$ in $[0, T] \times \mathcal{Q}$, then $q \in \mathcal{S}_0(t)$.

The above-listed hypotheses are specializations of the more general assumptions in [23, Sections 2.4.2, 2.5.1], taylored to the setting relevant for this work. In fact, as a corollary of [23, Theorem 2.4.10], the evolutionary Γ-convergence (24) follows if (H1)–(H5) hold and $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}_\epsilon)$ admit energetic solutions.

The proof is based on a, by now classical, time-discretization strategy. If (H1), (H3), (H4), and (H5) hold, then (25) can be considered a consequence of [23, Theorem 2.5.1] or [24, Theorem 4.1]. Strictly speaking, the latter references write the statement explicitly only for relaxation, so assuming that $\mathcal{E}_\epsilon$ are all identical for all ϵ, but as mentioned there already, the results carry over to general families of energy functionals after a straightforward modification. Both for (24) and (25), one obtains that the limit solution $q : [0, T] \to \mathcal{Q}$ is measurable.

When it comes to verifying the hypotheses, the critical condition to check is (H5). A sufficient condition is the existence of a mutual recovery sequence, cf. [23, (2.4.13), Proposition 2.4.8 (ii), Lemma 2.1.14]:

(H6) For any sequence $(t_\epsilon, q_\epsilon)_\epsilon \subset [0, T] \times \mathcal{Q}$ with $\sup_{\epsilon>0} \mathcal{E}_\epsilon(t_\epsilon, q_\epsilon) < +\infty$ that converges to $(t, q) \in [0, T] \times \mathcal{Q}$ and any $\tilde{q} \in \mathcal{Q}$, there exists a sequence $(\tilde{q}_\epsilon)_\epsilon \subset \mathcal{Q}$ such that $\tilde{q}_\epsilon \to q$ in $\mathcal{Q}$ and

$$\limsup_{\epsilon\to 0} \mathcal{E}_\epsilon(t_\epsilon, \tilde{q}_\epsilon) + \mathcal{D}_\epsilon(q_\epsilon, \tilde{q}_\epsilon) - \mathcal{E}_\epsilon(t_\epsilon, q_\epsilon) \le \mathcal{E}_0(t, \tilde{q}) + \mathcal{D}_0(q, \tilde{q}) - \mathcal{E}_0(t, q). \tag{27}$$

Now, after this brief excursion into the theory of evolutionary convergence for rate-independent systems, we discuss two different scenarios of such results associated with the plasticity model introduced previously in Sect. 2. The key part of the proofs is the construction of mutual recovery sequences, where the main difficulty lies in accommodating the non-convex constraints in order to obtain admissible states. A full picture of evolutionary Γ-convergence in our homogenization setting of crystal plasticity seems currently out of reach. We present in the following a few first steps by studying specific cases where energetic and dissipative effects decouple, see Theorems 3 and 4 below.

Let us first introduce the general setting, fix notation and provide some preliminaries. It is assumed in the remainder of this section that $\Omega = (-1, 1)^2$ and $s \in \{e_1, e_2\}$. Let $\mathcal{Q} = \mathcal{Y} \times \mathcal{Z}$ with $\mathcal{Y} = W^{1,2}(\Omega; \mathbb{R}^2) \cap L^2_0(\Omega; \mathbb{R}^2)$ and $\mathcal{Z} = L^2(\Omega)$, where both spaces are equipped with the corresponding weak topologies. We write $q = (u, \gamma) \in \mathcal{Q}$.

In what follows, we consider dissipation distances $\mathcal{D} : \mathcal{Q} \times \mathcal{Q} \to [0, +\infty]$ given for every $q = (u, \gamma)$, $\tilde{q} = (\tilde{u}, \tilde{\gamma}) \in \mathcal{Q}$ by either

$$\mathcal{D}(q, \tilde{q}) = \mathcal{D}^1(q, \tilde{q}) := \delta \int_\Omega |\gamma - \tilde{\gamma}| \,\mathrm{d}x = \delta \|\gamma - \tilde{\gamma}\|_{L^1(\Omega)} \tag{28}$$

with a dissipation coefficient $\delta > 0$ or by

$$\mathcal{D}(q, \tilde{q}) = \mathcal{D}_\ge(q, \tilde{q}) := \begin{cases} \mathcal{D}^1(q, \tilde{q}) & \text{if } \gamma \ge \tilde{\gamma} \text{ a.e. in } \Omega, \\ +\infty & \text{otherwise;} \end{cases} \tag{29}$$

note that the second choice of dissipation incorporates a monotonicity assumption on the direction of plastic glide, making for a unidirectional process. The underlying energy functionals $\mathcal{E}_\epsilon : [0, T] \times \mathcal{Q} \to \mathbb{R}_\infty$ for $\epsilon > 0$ associated with our system are

$$\mathcal{E}_\epsilon(t, q) := E_\epsilon(q) - \int_\Omega g_\epsilon(t) \cdot u \, \mathrm{d}x, \tag{30}$$

where $g_\epsilon \in W^{1,1}(0, T; L^2(\Omega; \mathbb{R}^2))$ are given body forces and $E_\epsilon : \mathcal{Q} \to [0, +\infty]$ time-independent energy contributions; the latter will be specified below in each subsection but share the common property that their admissible states are contained in the sets

$$\begin{aligned}\mathcal{B}_\varepsilon^{(s)} &:= \{q = (u, \gamma) \in \mathcal{Q} : \nabla u = R(\mathbb{I} + \gamma s \otimes m),\, R \in L^\infty(\Omega; SO(2)),\, \gamma = 0 \text{ on } \epsilon Y_{\text{rig}}\} \\ &= \{q = (u, \gamma_u) \in \mathcal{Q} : u \in \mathcal{A}_\varepsilon^{(s)}\},\end{aligned} \tag{31}$$

recalling $\mathcal{A}_\epsilon^{(s)}$ from (4), i.e.,

$$\mathcal{A}_\varepsilon^{(s)} = \{u \in \mathcal{Y} : \nabla u = R_u(\mathbb{I} + \gamma_u s \otimes m),\, R_u \in L^\infty(\Omega; SO(2)),\, \gamma_u = 0 \text{ on } \epsilon Y_{\text{rig}}\}.$$

As it was shown in [5], the limits of sequences of $\mathcal{B}_\epsilon^{(s)}$ can be characterized via

$$\begin{aligned}\mathcal{B}^{(s)} &:= \{q \in \mathcal{Q} : q_\epsilon \to q \text{ in } \mathcal{Q},\ q_\epsilon \in \mathcal{B}_\epsilon^{(s)} \text{ for all } \epsilon > 0\} \\ &= \{(u, \gamma_u) \in \mathcal{Q} : u \in \mathcal{A}^{(s)}\},\end{aligned} \tag{32}$$

with $\mathcal{A}^{(s)}$ as defined in (6).

3.1 The Case $s = e_2$

For $\epsilon > 0$, consider the energy functional $\mathcal{E}_\epsilon : [0, T] \times \mathcal{Q} \to \mathbb{R}_\infty$ as in (30) with

$$E_\epsilon(q) := \begin{cases} \displaystyle\int_\Omega \gamma^2 \, \mathrm{d}x & \text{if } q = (u, \gamma) \in \mathcal{B}_\epsilon^{(e_2)}, \\ +\infty & \text{otherwise,} \end{cases} \qquad \text{for } q \in \mathcal{Q}, \tag{33}$$

where $\mathcal{B}_\epsilon^{(e_2)}$ is given in (31). The next theorem shows that passing to the limit $\epsilon \to 0$ in the dissipative system $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D})$ yields a purely energetic evolution without any dissipation, see Remark 5. We point out that the energy functionals E_ϵ are not lower semicontinuous in light of oscillating rotations and observe that the oscillations in the shear strains prevent the continuous convergence of the dissipation.

As a consequence, the existence of energetic solutions to $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D})$ is not guaranteed, which suggests to formulate the result within the framework of approximate Γ-convergence. Regarding the proof strategy, note that classical constructions of mutual recovery sequences such as the "quadratic trick" (see e.g. [23, Section 3.4.5]) cannot be applied here due to the nonlinearity of the problem.

Theorem 3 (Evolutionary Γ-Convergence for $s = e_2$) *Let $\mathcal{E}_\epsilon$ for $\epsilon > 0$ be as in (30) and (33), and let $\mathcal{D}$ be as in (28) or (29). Furthermore, suppose that $g_\epsilon \to g_0$ in $W^{1,1}(0, T; L^2(\Omega; \mathbb{R}^2))$ with $g_0 \in W^{1,1}(0, T; L^2(\Omega; \mathbb{R}^2))$. Then,*

$$(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D}) \xrightarrow{ev\text{-}\Gamma_{\mathrm{app}}} (\mathcal{Q}, \mathcal{E}_0, \mathcal{D}) \quad \text{as } \epsilon \to 0,$$

where

$$\mathcal{E}_0(t, q) := \begin{cases} -\displaystyle\int_\Omega g_0(t) \cdot Rx \,\mathrm{d}x & \text{for } q = (u, 0) \in \mathcal{Q} \text{ with } \nabla u = R \text{ for } R \in SO(2), \\ +\infty & \text{otherwise.} \end{cases} \tag{34}$$

Proof The statement follows, once (H1), (H3), (H4), and (H5) are verified. The first two hypotheses are straightforward to check. Since E_0 given by $E_0(q) = 0$ for $q = (0, Rx)$ with $R \in \mathrm{SO}(2)$ and $E_0(q) = +\infty$ otherwise was characterized as the Γ-limit of $(E_\epsilon)_\epsilon$ in [5] (see also (5)), (H4) is verified. For the remaining condition (H5), we construct a suitable mutual recovery sequence according to (H6).

To this end, let $\tilde{q} \in \mathcal{Q}$, and let $(t_\epsilon, q_\epsilon)_\epsilon \subset [0, T] \times \mathcal{Q}$ be a sequence of uniformly bounded energy for $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D})_\epsilon$ converging to $(t, q) \in [0, T] \times \mathcal{Q}$. Since $\sup_{\epsilon>0} \mathcal{E}_\epsilon(t_\epsilon, q_\epsilon) < +\infty$, we have $q_\epsilon = (u_\epsilon, \gamma_\epsilon) \in \mathcal{B}_\epsilon^{(e_2)}$ for all ϵ, and therefore, according to (32),

$$q = (u, \gamma) \in \mathcal{B}^{(e_2)}, \text{ or equivalently, } \quad u \in \mathcal{A}^{(e_2)} \text{ and } \gamma = \gamma_u.$$

This shows that u is affine with $\nabla u = R$ for some $R \in SO(2)$ and that $\gamma_u = 0$. Hence, $\mathcal{E}_0(t, q) < +\infty$ in view of (34). As we aim to prove the estimate (27) for $\mathcal{D}_\epsilon = \mathcal{D}_0 = \mathcal{D}$, there is no loss of generality in assuming $\mathcal{E}_0(t, \tilde{q}) < +\infty$, i.e., $\tilde{q} = (\tilde{u}, 0) \in \mathcal{Q}$ with $\tilde{u}$ affine such that $\nabla\tilde{u} = \tilde{R}$ for $\tilde{R} \in SO(2)$.

We set

$$\tilde{u}_\epsilon = \tilde{R} R^T u_\epsilon \in \mathcal{A}_\epsilon^{(e_2)}$$

and $\tilde{q}_\epsilon = (\tilde{u}_\epsilon, \gamma_\epsilon) \in \mathcal{B}_\epsilon^{(e_2)}$ for any ϵ. Then, $\tilde{u}_\epsilon \rightharpoonup \tilde{u}$ in $W^{1,2}(\Omega; \mathbb{R}^2)$, and also $\tilde{q}_\epsilon \to \tilde{q}$ in $\mathcal{Q}$. Via compact Sobolev embedding, it holds (up to the selection of subsequences) that $\tilde{u}_\epsilon \to \tilde{u}$ in $L^2(\Omega; \mathbb{R}^2)$ and $u_\epsilon \to u$ in $L^2(\Omega; \mathbb{R}^2)$, which implies

$$\limsup_{\epsilon\to 0} -\int_\Omega g_\epsilon(t_\epsilon)\cdot(\tilde{u}_\epsilon - u_\epsilon)\,\mathrm{d}x = -\int_\Omega g_0(t)\cdot(\tilde{u}-u)\,\mathrm{d}x.$$

In light of $E_\epsilon(q_\epsilon) = E_\epsilon(\tilde{q}_\epsilon)$, $\mathcal{D}(q_\epsilon, \tilde{q}_\epsilon) = 0$ due to $\tilde{\gamma}_\epsilon = \gamma_\epsilon$, and $\mathcal{D}_0(q, \tilde{q}) = 0$, we therefore obtain

$$\limsup_{\epsilon\to 0} \mathcal{E}_\epsilon(t_\epsilon, \tilde{q}_\epsilon) - \mathcal{E}_\epsilon(t_\epsilon, q_\epsilon) + \mathcal{D}(q_\epsilon, \tilde{q}_\epsilon) = \mathcal{E}_0(t, \tilde{q}) - \mathcal{E}_0(t, q) + \mathcal{D}_0(q, \tilde{q}).$$

This yields (27) and shows the existence of a mutual recovery sequence, as desired. □

Remark 5 Given the system $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D})$, we observe that passing from one admissible state with finite energy to the next does not cause any dissipation. This is because $\mathcal{D}$ is zero when evaluated in finite-energy states, and thus, the set of stable states for $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D})$ is the same as that of $(\mathcal{Q}, \mathcal{E}_0, 0)$. Hence, one may think of the limit system $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D})$ as dissipation-free.

We conclude the case $s = e_2$ with a brief discussion of stable states and energetic solutions for the limit system $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D})$. Using Remark 1 (b), its set of stable states, denoted for time $t \in [0, T]$ by $\mathcal{S}_0(t)$, can be determined via minimization of $\mathcal{E}_0(t, \cdot)$, that is,

$$\begin{aligned}\mathcal{S}_0(t) &= \{q \in \mathcal{Q} : \mathcal{E}_0(t, q) \leq \mathcal{E}_0(t, \tilde{q}) \text{ for all } \tilde{q} \in \mathcal{Q}\}\\ &= \Big\{(u, 0) : u(x) = Rx \text{ for } x \in \Omega,\ R \in \operatorname{argmin}_{S\in SO(2)} -\int_\Omega g_0(t)\cdot Sx\,\mathrm{d}x\Big\}\\ &= \begin{cases}\{(Rx, 0) : Re_1 = \widehat{g_0(t)}/|\widehat{g_0(t)}|\} & \text{if } \widehat{g_0(t)} \neq 0,\\ \{(Rx, 0) : R \in SO(2)\} & \text{if } \widehat{g_0(t)} = 0;\end{cases}\end{aligned}$$

recall (14) for the definition of $\widehat{g_0(t)}$. In particular, $\mathcal{S}_0(t)$ contains only rigid-body motions, and the energy balance for $q(t, \cdot) = (u(t, \cdot), 0) \in \mathcal{S}_0(t)$ with $u(t)(x) = u(t, x) = R(t)x$ becomes

$$-R(t)e_1 \cdot \widehat{g_0(t)} = -R(0)e_1 \cdot \widehat{g_0(0)} - \int_0^t R(\tau)e_1 \cdot \widehat{\dot{g}_0(\tau)}\,\mathrm{d}\tau$$

for $t \in [0, T]$, cf. (23). Under the assumption that $\widehat{g_0(t)} \neq 0$ for all $t \in [0, T]$, we conclude that $(\mathcal{Q}, \mathcal{E}_0, \mathcal{D})$ has a unique energetic solution, which is given by $q(t, \cdot) = (u(t, \cdot), 0)$ with

$$u(t, x) = \big(x_1\widehat{g_0(t)} + x_2\widehat{g_0(t)}^{\perp}\big)/|\widehat{g_0(t)}|.$$

3.2 The Case $s = e_1$

As energy functional $\mathcal{E}_\epsilon : [0, T] \times Q \to \mathbb{R}_\infty$ for $\epsilon > 0$, we choose either

$$\mathcal{E}_\epsilon(t, q) = \mathcal{E}_\epsilon^{\mathrm{rig}}(t, q) := \begin{cases} \int_\Omega \gamma^2 \,\mathrm{d}x - \int_\Omega g_\epsilon(t) \cdot u \,\mathrm{d}x & \text{if } u \in \mathcal{A}_\epsilon^{(e_1),\mathrm{rig}}, \\ +\infty & \text{otherwise} \end{cases} \tag{35}$$

or

$$\mathcal{E}_\epsilon(t, q) = \mathcal{E}_{\epsilon,\tau_\epsilon}^{\mathrm{reg}}(t, q) := \begin{cases} \int_\Omega \gamma^2 \,\mathrm{d}x - \int_\Omega g_\varepsilon(t) \cdot u \,\mathrm{d}x + \tau_\epsilon \|\partial_1 \gamma\|^2_{L^2(\Omega)} & \text{if } u \in \mathcal{A}_\epsilon^{(e_1)}, \\ +\infty, & \text{otherwise;} \end{cases} \tag{36}$$

here,

$$\mathcal{A}_\epsilon^{(e_1),\mathrm{rig}} := \{u \in \mathcal{A}_\epsilon^{(e_1)} : R_u = \text{const.}\}$$

with $\mathcal{A}_\epsilon^{(e_1)}$ as in (4), $\tau_\epsilon > 0$ are given real constants such that $\tau_\epsilon \to +\infty$ as $\epsilon \to 0$, $g_\epsilon \in W^{1,1}(0, T; L^2(\Omega; \mathbb{R}^2))$ represent loading terms with $g_\epsilon \to g_0$ in $W^{1,1}(0, T; L^2(\Omega; \mathbb{R}^2))$ as $\epsilon \to 0$.

We remark that, in the first situation with $\mathcal{E}_\epsilon^{\mathrm{rig}}$, the inclusion

$$\mathcal{A}_\epsilon^{(e_1),\mathrm{rig}} \subset \mathcal{A}^{(e_1)} = \{u \in \mathcal{Y} : \nabla u = R(\mathbb{I} + \gamma e_1 \otimes e_2), R \in SO(2), \gamma \in L^2(\Omega), \partial_1 \gamma = 0\}$$

for any $\epsilon > 0$ leads to an essentially one-dimensional model already at the level of the ϵ-dependent functionals.

The additional term in $\mathcal{E}_{\epsilon,\tau_\epsilon}^{\mathrm{reg}}$ corresponds to a unidirectional regularization of the shears on the softer layers. As such, it penalizes oscillations in the x_1-direction, but does not affect those in x_2, which are relevant to respect the layered structure. A common modeling choice for regularizing via higher order derivatives in models of finite plasticity involves the dislocation tensor $\mathcal{G}$, see, e.g., [22]. Considering the multiplicative splitting of the deformation gradient into an elastic and a plastic part in (2), $\mathcal{G}$ in the planar case is defined by

$$\mathcal{G}(F_{\mathrm{pl}}) := \frac{1}{\det F_{\mathrm{pl}}} \operatorname{curl} F_{\mathrm{pl}}.$$

In the present setting, with slip direction $s = e_1$ and slip plane normal $m = e_2$, it holds that $F_{\mathrm{pl}} = \mathbb{I} + \gamma e_1 \otimes e_2$ (see (3)), and therefore,

$$\mathcal{G}(F_{\mathrm{pl}}) = \partial_1(F_{\mathrm{pl}} e_2) - \partial_2(F_{\mathrm{pl}} e_1) = (\partial_1 \gamma) e_1.$$

Hence, the penalization in $\mathcal{E}^{\mathrm{reg}}_{\epsilon,\tau_\epsilon}$ can be seen as forcing the L^2-norm of the dislocation tensor to asymptotically become infinitesimal and leads to a limit model in the context of dislocation-free finite crystal plasticity, as recently studied in [20, 30].

We present below two (approximate) evolutionary Γ-convergence results for the rate-independent systems with the previously introduced energies and dissipations. It is important to notice that both limit systems show no interaction between energy and dissipation terms. This effect is enforced by the rigid energy in $\mathcal{E}^{\mathrm{rig}}_{\varepsilon}$ and the regularization in $\mathcal{E}^{\mathrm{reg}}_{\varepsilon,\tau_\varepsilon}$, which suppress oscillations in the rotations and shears, respectively. On a technical level, these assumptions are designed to make a limit analysis based on weak–strong convergence arguments feasible. The case of fully oscillating rotations and shears, which requires characterizing limits of products of weakly convergent sequences, may become accessible with new arguments from the theory of compensated compactness (see, e.g., [27, 31]) but is beyond the scope of this work and currently still open.

Theorem 4 (Evolutionary $\boldsymbol{\Gamma}$-Convergence for $\boldsymbol{s = e_1}$) *With the definitions in* (35)*,* (36)*,* (28)*, and* (29) *and*

$$\mathcal{E}_0(t,q) := \begin{cases} \dfrac{1}{\lambda}\displaystyle\int_\Omega \gamma^2 \,\mathrm{d}x - \int_\Omega g_0(t)\cdot u \,\mathrm{d}x & \textit{if } u \in \mathcal{A}^{(e_1)}, \\ +\infty & \textit{otherwise,} \end{cases}$$

the following evolutionary Γ*-convergence results hold:*

$$(a)\ (\mathcal{Q}, \mathcal{E}^{\mathrm{rig}}_{\epsilon}, \mathcal{D}_{\geq}) \xrightarrow{ev\text{-}\Gamma} (\mathcal{Q}, \mathcal{E}_0, \mathcal{D}_{\geq}) \textit{ as } \epsilon \to 0;$$

$$(b)\ (\mathcal{Q}, \mathcal{E}^{\mathrm{reg}}_{\epsilon,\tau_\epsilon}, \mathcal{D}^1) \xrightarrow{ev\text{-}\Gamma_{\mathrm{app}}} (\mathcal{Q}, \mathcal{E}_0, \mathcal{D}^1) \textit{ as } \epsilon \to 0.$$

Proof All the basic assumptions of the theory summarized above are satisfied. While (H1), (H3), as well as (H2) for (a) are immediate to check, (H4) follows from [5]. It remains to prove the existence of mutual recovery sequences, that is, (H6).

Consider $q, \tilde{q} \in \mathcal{Q}$ with $u, \tilde{u} \in \mathcal{A}^{(e_1)}$, i.e.,

$$\nabla u = R(\mathbb{I} + \gamma e_1 \otimes e_2) \quad \textit{and} \quad \nabla \tilde{u} = \tilde{R}(\mathbb{I} + \tilde{\gamma} e_1 \otimes e_2)$$

with $R, \tilde{R} \in SO(2)$ and $\gamma, \tilde{\gamma} \in L^2(\Omega)$ such that $\partial_1 \gamma = \partial_1 \tilde{\gamma} = 0$. Moreover, let $(t_\epsilon, q_\epsilon)_\epsilon \subset [0,T] \times \mathcal{Q}$ be a bounded energy sequence for $(\mathcal{Q}, \mathcal{E}_\epsilon, \mathcal{D})_\epsilon$ with

$$\sup_{\epsilon>0} \mathcal{E}_\epsilon(t_\epsilon, q_\epsilon) < +\infty, \tag{37}$$

such that $(t_\epsilon, q_\epsilon) \to (t, q)$ in $[0,T] \times \mathcal{Q}$.

As a consequence of this uniform energy bound, one obtains that $u_\epsilon \in \mathcal{A}_\epsilon^{(e_1)}$ for all $\epsilon > 0$; in the case of $\mathcal{E}_\varepsilon = \mathcal{E}_\varepsilon^{\mathrm{rig}}$, one even has $u_\epsilon \in \mathcal{A}_\epsilon^{(e_1),\mathrm{rig}}$. Moreover, we infer from the Radon–Riesz theorem that

$$R_\epsilon \to R \quad \text{in } L^2(\Omega;\mathbb{R}^{2\times 2}) \quad \text{and} \quad \gamma_\epsilon \rightharpoonup \gamma \quad \text{in } L^2(\Omega),$$

where, for the sake of a simpler notation, we write R_ϵ for R_{u_ϵ} and γ_ϵ for γ_{u_ϵ}. The task is to find a sequence $(\tilde{q}_\epsilon)_\epsilon \subset \mathcal{Q}$ with $\tilde{q}_\epsilon \rightharpoonup \tilde{q}$ in $\mathcal{Q}$ that satisfies

$$\limsup_{\epsilon\to 0} \mathcal{E}_\epsilon(t_\epsilon, \tilde{q}_\epsilon) - \mathcal{E}_\epsilon(t_\epsilon, q_\epsilon) + \mathcal{D}(q_\epsilon, \tilde{q}_\epsilon) \le \mathcal{E}_0(t, \tilde{q}) - \mathcal{E}_0(t, q) + \mathcal{D}_0(q, \tilde{q}), \tag{38}$$

with $\mathcal{D} = \mathcal{D}_0 = \mathcal{D}_\ge$ in (a) and $\mathcal{D} = \mathcal{D}_0 = \mathcal{D}^1$ in (b). Observe that necessarily, $\tilde{u}_\epsilon \in \mathcal{A}_\epsilon^{(e_1)}$ for all $\epsilon > 0$, and additionally, for $\mathcal{E}_\varepsilon = \mathcal{E}_\varepsilon^{\mathrm{rig}}$, one needs $\tilde{u}_\epsilon \in \mathcal{A}_\epsilon^{(e_1),\mathrm{rig}}$. We now detail the construction of mutual recovery sequences for the two scenarios described in (a) and (b).

(a) In this case, the problem is essentially (up to global rotations) one-dimensional with quadratic energy, which allows us to adapt what is frequently referred to as the "quadratic trick," cf. [23, Section 3.5.4]. In particular, $(\mathcal{Q}, \mathcal{E}_\epsilon^{\mathrm{rig}}, \mathcal{D}_\ge)$ has an energetic solution for each $\epsilon > 0$. To meet the required constraints for admissible sequences of $\mathcal{E}_\epsilon^{\mathrm{rig}}$, we define $\tilde{u}_\epsilon \in \mathcal{A}_\epsilon^{(e_1)}$ by setting

$$\tilde{R}_\epsilon := \tilde{R}_{u_\epsilon} = \tilde{R} \qquad \text{and} \qquad \tilde{\gamma}_\epsilon := \tilde{\gamma}_{u_\epsilon} = \gamma_\epsilon + \frac{1}{\lambda}(\tilde{\gamma} - \gamma)\mathbb{1}_{\epsilon Y_{\mathrm{soft}}}$$

for all $\epsilon > 0$, recalling that $\lambda \in (0, 1)$ denotes the relative thickness of the soft layers. Indeed, since $\partial_1 \gamma_\epsilon = 0$ due to $u_\epsilon \in \mathcal{A}^{(e_1)}$, also $\partial_1 \tilde{\gamma}_\epsilon = 0$, and hence, the vector field $\tilde{R}_\epsilon(\mathbb{I} + \tilde{\gamma}_\epsilon e_1 \otimes e_2)$ is indeed a gradient field, namely for the potential $\tilde{u}_\epsilon$.

In view of $\tilde{\gamma}_\epsilon \rightharpoonup \tilde{\gamma}$ in $L^2(\Omega)$ and $\tilde{u}_\epsilon \rightharpoonup \tilde{u}$ in $W^{1,2}(\Omega;\mathbb{R}^2)$, it follows that

$$\begin{aligned}
&\limsup_{\epsilon\to 0} \mathcal{E}_\epsilon^{\mathrm{rig}}(t_\epsilon, \tilde{q}_\epsilon) - \mathcal{E}_\epsilon^{\mathrm{rig}}(t_\epsilon, q_\epsilon)\\
&\quad = \limsup_{\epsilon\to 0} \int_{\Omega\cap\epsilon Y_{\mathrm{soft}}} \frac{1}{\lambda^2}(\tilde{\gamma} - \gamma)^2 + \frac{2}{\lambda}(\tilde{\gamma} - \gamma)\gamma_\epsilon \,\mathrm{d}x - \int_\Omega g_\epsilon(t_\epsilon) \cdot (\tilde{u}_\epsilon - u_\epsilon)\,\mathrm{d}x\\
&\quad = \int_\Omega \frac{1}{\lambda}(\tilde{\gamma} - \gamma)^2 + \frac{2}{\lambda}\tilde{\gamma}\gamma - \frac{2}{\lambda}\gamma^2 \,\mathrm{d}x - \int_\Omega g_0(t) \cdot (\tilde{u} - u)\,\mathrm{d}x\\
&\quad = \int_\Omega \frac{1}{\lambda}\tilde{\gamma}^2 - \frac{1}{\lambda}\gamma^2 \,\mathrm{d}x - \int_\Omega g_0(t) \cdot (\tilde{u} - u)\,\mathrm{d}x\\
&\quad = \mathcal{E}_0(t, \tilde{q}) - \mathcal{E}_0(t, q),
\end{aligned} \tag{39}$$

using that $\mathbb{1}_{\epsilon Y_{\text{soft}}} \overset{*}{\rightharpoonup} \lambda$ id in $L^\infty(\Omega)$ by the Riemann-Lebesgue lemma and that $\mathbb{1}_{\epsilon Y_{\text{soft}}}\gamma_\epsilon = \gamma_\epsilon$ for all ϵ.

Due to the monotonicity constraint in $\mathcal{D}_\geq$, we may assume that $\gamma \geq \tilde{\gamma}$, which implies that $\gamma_\epsilon \geq \tilde{\gamma}_\epsilon$. Then,

$$\begin{aligned}\limsup_{\epsilon\to 0} \mathcal{D}_\geq(\gamma_\epsilon, \tilde{\gamma}_\epsilon) &= \limsup_{\epsilon\to 0} \delta \int_\Omega \gamma_\epsilon - \tilde{\gamma}_\epsilon \,\mathrm{d}x = \limsup_{\epsilon\to 0} \frac{\delta}{\lambda} \int_\Omega \mathbb{1}_{\epsilon Y_{\text{soft}}}(\tilde{\gamma} - \gamma)\,\mathrm{d}x \\ &= \delta \int_\Omega \gamma - \tilde{\gamma}\,\mathrm{d}x = \mathcal{D}_\geq(\gamma, \tilde{\gamma}).\end{aligned} \tag{40}$$

Combining (39) and (40) gives (38), as desired.

(b) We start by observing that the uniform energy bound (37) implies

$$\lim_{\epsilon\to 0} \int_\Omega |\partial_1 \gamma_\epsilon|^2 \,\mathrm{d}x = 0; \tag{41}$$

indeed, with $\gamma_\epsilon = \gamma_{u_\epsilon}$, we obtain via the Poincaré–Wirtinger inequality that

$$\begin{aligned}\int_\Omega \gamma_\epsilon^2 \,\mathrm{d}x + \tau_\epsilon \|\partial_1\gamma_\epsilon\|^2_{L^2(\Omega)} &\leq \sup_{\epsilon>0} \mathcal{E}_\epsilon(t_\epsilon, q_\epsilon) + \|g_\epsilon\|_{W^{1,1}(0,T;L^2(\Omega))}\|u_\epsilon\|_{L^2(\Omega)} \\ &\leq C(1 + \|\nabla u_\epsilon\|_{L^2(\Omega;\mathbb{R}^{2\times 2})}) \leq 4C(1 + \|\gamma_\epsilon\|_{L^2(\Omega)}),\end{aligned}$$

with a constant $C > 0$ independent of ϵ. This shows that $(\gamma_\epsilon)_\epsilon$ is uniformly bounded in $L^2(\Omega)$, as well as $\|\partial_1\gamma_\epsilon\|_{L^2(\Omega)} \to 0$, considering that $\tau_\epsilon \to +\infty$ as ϵ tends to zero.

Let us define $\tilde{R}_\epsilon = \tilde{R}$ and

$$\tilde{\gamma}_\epsilon = \frac{1}{2}\int_{-1}^{1} \gamma_\epsilon \,\mathrm{d}x_1 + \frac{1}{\lambda}(\tilde{\gamma} - \gamma)\mathbb{1}_{\epsilon Y_{\text{soft}}} \tag{42}$$

for $\epsilon > 0$. By construction, we have again that $\partial_1\tilde{\gamma}_\epsilon = 0$. The ansatz in (42) can be viewed as yet a refined version of the modified "quadratic trick" in (a).

In proving (38), the convergence of the energy terms follows in analogy to (39), if we account for the fact that $\frac{1}{2}\int_{-1}^{1} \gamma_\epsilon \,\mathrm{d}x_1 \rightharpoonup \gamma$ in $L^2(\Omega)$ and if we use the estimate

$$\int_\Omega \Big(\frac{1}{2}\int_{-1}^{1} \gamma_\epsilon \,\mathrm{d}x_1\Big)^2 - \gamma_\epsilon^2 \,\mathrm{d}x \leq 0$$

by Jensen's inequality.

Regarding the dissipative terms, we use the one-dimensional Poincaré inequality with Poincaré constant $c > 0$ to argue that

$$
\begin{aligned}
\limsup_{\epsilon\to 0} \mathcal{D}^1(\gamma_\epsilon, \tilde{\gamma}_\epsilon) &\leq \limsup_{\epsilon\to 0} \delta \int_\Omega \left| \gamma_\epsilon - \frac{1}{2}\int_{-1}^{1} \gamma_\epsilon \, \mathrm{d}x_1 \right| \mathrm{d}x + \delta \int_\Omega \frac{1}{\lambda} |\tilde{\gamma} - \gamma| \mathbb{1}_{\epsilon Y_{\mathrm{soft}}} \, \mathrm{d}x \\
&\leq \delta \limsup_{\epsilon\to 0} c \int_\Omega |\partial_1 \gamma_\epsilon| \, \mathrm{d}x + \delta \int_\Omega |\tilde{\gamma} - \gamma| \, \mathrm{d}x \\
&\leq c\delta \lim_{\epsilon\to 0} \|\partial_1 \gamma_\epsilon\|_{L^2(\Omega)} + \mathcal{D}^1(\gamma, \tilde{\gamma}).
\end{aligned}
$$

The proof of (38) follows then by (41). □

Remark 6 Following the proof of Theorem 4(a), it is immediate to see that we get the analogous evolutionary Γ-convergence result if the monotonicity assumption in the dissipation is dropped, i.e., $(Q, \mathcal{E}_\epsilon^{\mathrm{rig}}, \mathcal{D}^1) \xrightarrow{ev\text{-}\Gamma} (Q, \mathcal{E}_0, \mathcal{D}^1)$ as $\epsilon \to 0$.

Acknowledgments E.D. acknowledges support from the Austrian Science Fund (FWF) through the projects F 65, I 4052, V 662, and Y1292, as well as from BMBWF through the OeAD-WTZ project CZ04/2019. C.K. was supported by the Dutch Research Council (NWO) through the project TOP2.17.01. Most of this work was done, while C.K. was affiliated with Utrecht University and partially supported by the Westerdijk Fellowship program. Both the authors are grateful for the opportunity to collaborate during a stay at MFO Oberwolfach in the scope of the Research in Pairs program.

References

1. S. Amstutz, N. Van Goethem, Incompatibility-governed elasto-plasticity for continua with dislocations. Proc. R. Soc. A. **473**, 20160734 (2017)
2. G. Bouchitté, M. Bellieud, Homogenization of a soft elastic material reinforced by fibers. Asymptot. Anal. **31**(2), 153–183 (2020)
3. A. Braides, *Γ-Convergence for Beginners*. Oxford Lecture Series in Mathematics and its Applications, vol. 22 (Oxford University Press, Oxford, 2002)
4. A. Brillard, M. El Jarroudi, Homogenization of a nonlinear elastic structure periodically reinforced along identical fibres of high rigidity. Nonlinear Anal. Real World Appl. **8**(1), 295–311 (2007)
5. F. Christowiak, C. Kreisbeck, Homogenization of layered materials with rigid components in single-slip finite crystal plasticity. Calc. Var. Partial Differ. Equ. **56**, 75 (2018)
6. F. Christowiak, C. Kreisbeck, Asymptotic rigidity of layered structures and its application in homogenization theory. Arch. Ration. Mech. Anal. **235**, 51–98 (2020)
7. S. Conti, Relaxation of single-slip single-crystal plasticity with linear hardening, in *Multiscale Materials Modeling*. Fraunhofer IRB, Freiburg (2006), pp. 30–35
8. S. Conti, F. Theil, Single-slip elastoplastic microstructures. Arch. Ration. Mech. Anal. **178**, 125–148 (2005)
9. S. Conti, G. Dolzmann, C. Kreisbeck, Asymptotic behavior of crystal plasticity with one slip system in the limit of rigid elasticity. SIAM J. Math. Anal. **43**, 2337–2353 (2011)
10. G. Dal Maso, *An Introduction to Γ-Convergence*. Progress in Nonlinear Differential Equations and their Applications, vol. 8 (Birkhäuser Boston, Boston, 1993)
11. E. Davoli, G.A. Francfort, A critical revisiting of finite elastoplasticity. SIAM J. Math. Analy. **47**, 526–565 (2015)

12. E. Davoli, R. Ferreira, C. Kreisbeck, Homogenization in BV of a model for layered composites in finite crystal plasticity. Adv. Calc. Var. **14**, 441–473 (2021)
13. G. Del Piero, On the decomposition of the deformation gradient in plasticity. J. Elasticity **131**(1), 111–124 (2018)
14. M. El Jarroudi, Homogenization of a nonlinear elastic fibre-reinforced composite: a second gradient nonlinear elastic material. J. Math. Anal. Appl. **403**(2), 487–505 (2013)
15. D. Engl, C. Kreisbeck, A. Ritorto, Asymptotic analysis of deformation behavior in high-contrast fiber-reinforced materials: Rigidity and anisotropy (2021). Preprint arXiv:2105.03971
16. D. Grandi, U. Stefanelli, Finite plasticity in $P^T P$. Part II: quasi-static evolution and linearization. SIAM J. Math. Anal. **49**, 1356–1384 (2017)
17. H. Hauke, Homogenization in gradient plasticity. Math. Models Methods Appl. Sci. **21**(8), 1651–1684 (2011)
18. R. Hill, *The Mathematical Theory of Plasticity* (Clarendon Press, Oxford, 1950)
19. E. Kröner, Allgemeine Kontinuumstheorie der Versetzungen und Eigenspannungen. Arch. Rational Mech. Anal. **4**, 273–334 (1960)
20. M. Kružík, D. Melching, U. Stefanelli, Quasistatic evolution for dislocation-free finite plasticity. ESAIM Control Optim. Calc. Var. **23**, 123 (2020)
21. E.H. Lee, Elastic-plastic deformation at finite strains. J. Appl. Mech. **36**, 1–6 (1969)
22. A. Mielke, S. Müller, Lower semicontinuity and existence of minimizers in incremental finite-strain elastoplasticity. ZAMM Z. Angew. Math. Mech. **86**(3), 233–250 (2006)
23. A. Mielke, T. Roubíček, *Rate-independent Systems*. Applied Mathematical Sciences, vol. 193 (Springer, New York, 2015). Theory and Application
24. A. Mielke, T. Roubíček, U. Stefanelli, Γ-limits and relaxations for rate-independent evolutionary problems. Calc. Var. Partial Differ. Equ. **31**(3), 387–416 (2008)
25. A. Mielke, U. Stefanelli, Linearized plasticity is the evolutionary Γ-limit of finite plasticity. J. Eur. Math. Soc. (JEMS) **15**(3), 923–948 (2013)
26. A. Mielke, A.M. Timofte, Two-scale homogenization for evolutionary variational inequalities via the energetic formulation. SIAM J. Math. Anal. **39**(2), 642–668 (2007)
27. F. Murat, Compacité par compensation: condition nécessaire et suffisante de continuité faible sous une hypothèse de rang constant Ann. Scuola Norm. Sup. Pisa Cl. Sci. **8**(1), 69–102 (1981)
28. R. Paroni, A. Sili, Non-local effects by homogenization or 3D-1D dimension reduction in elastic materials reinforced by stiff fibers. J. Differ. Equ. **260**(3), 2026–2059 (2016)
29. C. Reina, L.F. Djodom, M. Ortiz, S. Conti, Kinematics of elasto-plasticity: validity and limits of applicability of $\boldsymbol{F} = \boldsymbol{F}^{\mathrm{e}}\boldsymbol{F}^{\mathrm{p}}$ for general three-dimensional deformations. J. Mech. Phys. Solids **121**, 99–113 (2018)
30. U. Stefanelli, Existence for dislocation-free finite plasticity. ESAIM Control Optim. Calc. Var. **25**, 21 (2019)
31. L. Tartar, The compensated compactness method applied to systems of conservation laws, in *Systems of Nonlinear Partial Differential Equations*. NATO Science Series C: Mathematical and Physical Sciences, vol. 111 Springer, Dordrecht

On the Prescription of Boundary Conditions for Nonlocal Poisson's and Peridynamics Models

Marta D'Elia and Yue Yu

1 Introduction and Motivation

Nonlocal, integral models are valid alternatives to classical partial differential equations (PDEs) to describe systems where small-scale effects or interactions affect the global behavior. In particular, nonlocal models are characterized by integral operators that embed length scales in their definitions, allowing to capture long-range space interactions. Furthermore, the integral nature of such operators reduces the regularity requirements on the solutions that are allowed to feature discontinuous or singular behavior. Applications of interest span a large spectrum of scientific and engineering fields, including fracture mechanics [33, 47], anomalous subsurface transport [3, 15, 44, 45], phase transitions [8, 14, 30], image processing [6, 25, 32], magnetohydrodynamics [43], stochastic processes [7, 18, 35, 39], and turbulence [12, 40, 41].

Despite their improved accuracy, the usability of nonlocal equations is hindered by several modeling and computational challenges that are the subject of very active research. Modeling challenges include the lack of a unified and complete nonlocal theory [13, 20, 23], the nontrivial treatment of nonlocal interfaces [1, 10, 29, 46, 55, 58], and the non-intuitive prescription of nonlocal boundary conditions [22, 31, 50, 54, 59]. Computational challenges are due to the integral nature of nonlocal operators that yields discretization matrices that feature a much larger bandwidth compared to the sparse matrices associated with PDEs. For both variational methods

M. D'Elia (✉)
Sandia National Laboratories, Livermore, CA, USA
e-mail: mdelia@sandia.gov

Y. Yu
Lehigh University, Bethlehem, PA, USA
e-mail: yuy214@lehigh.edu

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_8

[2, 11, 19, 21] and meshfree methods [29, 42, 48, 50, 51, 53–55, 59], a lot of progress has been made during the last decade, resulting in improved numerical techniques that facilitate wider adoption, even at the engineering level.

In its simplest form, the action of a nonlocal (spatial) operator on a scalar function $u : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathcal{L}u(\boldsymbol{x}) = \int_{\mathcal{H}_\delta(\boldsymbol{x})} I(\boldsymbol{x}, \boldsymbol{y}, u)\, d\boldsymbol{y},$$

where $\mathcal{H}_\delta(\boldsymbol{x})$ defines a nonlocal neighborhood of size δ surrounding a point $\boldsymbol{x} \in \mathbb{R}^d$, d being the spatial dimension, and δ the so-called horizon or interaction radius. The latter defines the extent of the nonlocal interactions and embeds the nonlocal operator with a characteristic length scale. The integrand function I is application dependent and plays the role of a constitutive law. Its definition is not straightforward and represents one of the most investigated problems in nonlocal research [9, 16, 52, 56, 57].

In this work, we focus on the prescription of nonlocal boundary conditions, or volume constraints, when solving nonlocal equations in bounded domains. The challenge stems from the presence of nonlocal interactions, for which a point $\boldsymbol{x}$ in a domain interacts with points outside of the domain that are contained in the point's neighborhood $\mathcal{H}_\delta(\boldsymbol{x})$. This fact generates an interaction region of nonzero measure where volume constraints need to be prescribed to guarantee the uniqueness of a nonlocal solution [27]. However, often times, input data to a problem are not available (due to measurement cost or physical impediments) in volumetric regions, whereas they are only available on the surfaces surrounding the domain. In other words, the only available data are *local*. Thus, the question arises of *how to convert local boundary information into a nonlocal volume constraint.*

In the nonlocal literature, this issue has been addressed in several works, most of which propose conversion approaches that are either too restrictive (in terms of geometry or dimensionality constraints), too computationally expensive (requiring the solution of an optimization problem), or are not prone to wide usability (requiring a modification of available codes). Among these works, we mention [17, 24, 31, 54, 59].

The method we propose is inspired by the recent work [22] where the authors propose to first approximate the nonlocal solution with its local counterpart and then *correct* it by solving the nonlocal problem using the local solution to generate volume constraints. In [22], Neumann local boundary conditions are converted into Dirichlet or Neumann volume constraints in the context of nonlocal Poisson's problems and numerical tests are performed in one dimension. Based on this work, we propose to convert Dirichlet local boundary conditions into Dirichlet or Neumann volume constraints in the context of both nonlocal Poisson's and peridynamics equations. Furthermore, we show applicability of our strategy in a two-dimensional setting using nontrivial geometries.

The main idea of the proposed method can be summarized in three simple steps.

1. Using available local data, we solve the local counterpart of the nonlocal problem. This step assumes that the local limit (the limit as $\delta \to 0$) of the nonlocal operator is known,[1] that the local data and the domain are smooth enough to guarantee well-posedness, and that a solver for the corresponding local equation is available.
2. We use the local solution either to define the nonlocal Dirichlet data in the nonlocal interaction domain or to obtain the nonlocal Neumann data by computing the corresponding nonlocal flux. This step numerically corresponds to a matrix-vector multiplication and does not require the implementation of a new nonlocal (flux) operator; in fact, as we will explain later, the nonlocal Neumann operator is the nonlocal operator itself evaluated at points in the nonlocal interaction domain.
3. Use either the Dirichlet or Neumann data obtained in Step 2 to solve the nonlocal problem, for which volume constraints are now available.

The choice between converting into a Dirichlet or Neumann condition depends on the expected behavior close to the boundary. When nonlocal effects are more likely to happen far from the boundary (because of, e.g., a pre-crack in the middle of the domain), the Dirichlet approach, characterized by a smooth behavior of the solution, can be considered appropriate. On the other hand, when nonlocal effects are expected close to the boundary, the Neumann approach might be preferable, as it returns a solution that matches the nonlocal flux associated with the local solution, rather than matching the local solution itself. We summarize the main properties of the proposed approach below.

- This strategy delivers a nonlocal solution that is physically consistent with PDEs in the limit of vanishing nonlocality. Numerically, when employing proper numerical discretization methods, e.g., the optimization-based meshfree quadrature rule [50, 59], this property guarantees *asymptotic compatibility* [49], i.e., the nonlocal numerical solution converges to its local limit as δ and the discretization size h approach 0.
- This technique has no geometry or dimensionality constraints. It can be utilized with any domain shape and in all dimensions $d = 1, 2, 3$.
- The conversion of local data into nonlocal volume constraints is inexpensive. In fact, it corresponds to a matrix-vector product where the matrix is either a selection matrix (in the Dirichlet case) or a nonlocal flux matrix (in the Neumann case).
- This strategy does not require the implementation of new software. In fact, available local and nonlocal solvers can be used as black boxes.

[1] Local limits of nonlocal operators can be obtained by using Taylor's expansion; both the nonlocal Poisson's problem and the peridynamic model considered in this work have well-known local limits, namely, the (local) Poisson's equation and the Navier equation of linear elasticity, respectively.

Consequently, this strategy has the potential of dramatically increasing the usability of nonlocal models at the engineering and industry level thanks to its flexibility, intuitiveness, and ease of implementation.

Paper Outline This chapter is organized as follows. In the following section, we describe the nonlocal Poisson and linear peridynamic solid (LPS) models. For each of them, we introduce the strong and weak formulations and discuss conditions for their well-posedness. In Sect. 3, we illustrate the proposed strategies for the conversion of a local, Dirichlet boundary condition into a nonlocal Dirichlet (DtD strategy) or Neumann (DtN strategy) volume constraint. In Sect. 4, we prove that both approaches deliver nonlocal solutions that are asymptotically compatible with the corresponding local solution of both Poisson's and LPS problems. Specifically, we prove that the nonlocal solution converges to the local one with quadratic rate. In Sect. 5, we illustrate the properties of our methods with several two-dimensional numerical tests. In particular, we show that when the solutions are such that local and nonlocal operators are equivalent, our procedure satisfies the consistency property (the nonlocal solution coincides with the local one). Furthermore, for both models and both approaches, we confirm the quadratic convergence rate of the L^2-norm difference between local and nonlocal solutions. Finally, in Sect. 6, we summarize our achievements.

2 Preliminaries

In this section, we introduce the mathematical models used in this chapter and recall relevant results. In what follows, scalar fields are indicated by italic symbols and vector fields by bold symbols. Let Ω be a bounded open domain in $\mathbb{R}^d$, $d = 1, 2, 3$, with Lipschitz-continuous boundary $\partial\Omega$.

2.1 The Nonlocal Poisson's Problem

For the function $u(\boldsymbol{x})\colon \mathbb{R}^d \to \mathbb{R}$, we define the nonlocal Laplacian $\mathcal{L}^{NL}\colon \mathbb{R}^d \to \mathbb{R}$ of $u(\boldsymbol{x})$ as

$$\mathcal{L}^{NL}u(\boldsymbol{x}) := 2\int_{\mathbb{R}^d} \big(u(\boldsymbol{y}) - u(\boldsymbol{x})\big)\,\gamma(\boldsymbol{x}, \boldsymbol{y})\,d\boldsymbol{y} \qquad \boldsymbol{x} \in \mathbb{R}^d, \tag{1}$$

where $\gamma(\boldsymbol{x}, \boldsymbol{y})$ is a nonnegative symmetric kernel[2] such that, for $\boldsymbol{x} \in \Omega$,

[2] For more general, sign-changing, and nonsymmetric kernels, we refer the reader to [36] and [18], respectively.

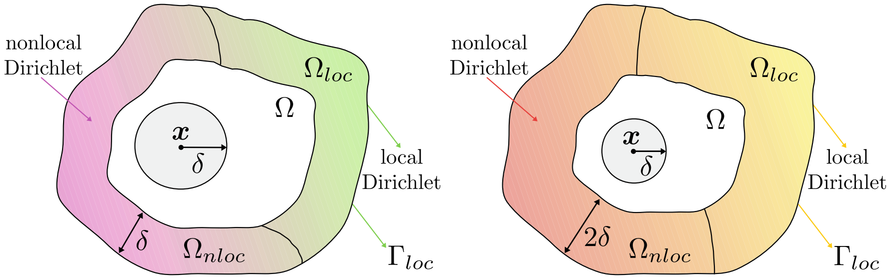


Fig. 1 The domain Ω, the support of γ at a point $\boldsymbol{x} \in \Omega$, $B_\delta(\boldsymbol{x})$, and the induced interaction domain Ω_I for the nonlocal Poisson's problem (left) and the LPS model (right)

$$\begin{cases} \gamma(\boldsymbol{x}, \boldsymbol{y}) > 0 \; \forall \, \boldsymbol{y} \in B_\delta(\boldsymbol{x}) \\ \gamma(\boldsymbol{x}, \boldsymbol{y}) = 0 \; \forall \, \boldsymbol{y} \in \mathbb{R}^d \setminus B_\delta(\boldsymbol{x}), \end{cases} \tag{2}$$

where $B_\delta(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}^d : \; \|\boldsymbol{x} - \boldsymbol{y}\| < \delta, \; \boldsymbol{x} \in \Omega\}$ and δ is the interaction radius or horizon. For the Laplacian operator $\mathcal{L}^{NL}$, we define the interaction domain of Ω associated with kernels like in (2) as follows:

$$\Omega_I = \{\boldsymbol{y} \in \mathbb{R}^d \setminus \Omega : \; \|\boldsymbol{y} - \boldsymbol{x}\| < \delta, \quad \text{for some } \boldsymbol{x} \in \Omega\}, \tag{3}$$

and set $\overline{\overline{\Omega}} = \Omega \cup \Omega_I$. The domain Ω_I contains all points outside of Ω that interact with points inside of Ω; as such, Ω_I is the volume where nonlocal boundary conditions, or volume constraints, must be prescribed to guarantee the well-posedness of the nonlocal equation associated with $\mathcal{L}^{NL}$ [27]. We refer to Fig. 1 (left) for an illustration of a two-dimensional domain, the support of γ, and the induced interaction domain. Here, the interaction domain is divided into the nonoverlapping partition $\Omega_I = \Omega_{nloc} \cup \Omega_{loc}$. In what follows, we assume that nonlocal data is available on Ω_{nloc}, whereas only local information is available on the physical boundary of Ω_{loc}, i.e., on $\Gamma_{loc} = \partial\Omega_{loc} \cap \partial\Omega$.

An important property of the Laplacian operator in (1) is its δ-convergence, i.e., as $\delta \to 0$ to the classical, local Laplacian Δ. In fact, when the kernel γ is properly scaled and when the fourth-order derivatives of u are bounded, we have the following pointwise relationship:

$$\mathcal{L}u(\boldsymbol{x}) = \Delta u(\boldsymbol{x}) + \mathcal{O}(\delta^2). \tag{4}$$

With the purpose of prescribing Neumann volume constraints, we introduce the nonlocal *flux* operator:

$$\mathcal{N}^{ND}u(\boldsymbol{x}) = -\int_{\overline{\overline{\Omega}}} (u(\boldsymbol{y}) - u(\boldsymbol{x}))\gamma(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y} \qquad \boldsymbol{x} \in \Omega_I.$$

To provide an interpretation of the interaction operator, we note that the integral $\int_{\Omega_I} \mathcal{N}^{ND}(\boldsymbol{v})\,d\boldsymbol{x}$ generalizes the concept of a local flux $\int_{\partial\Omega} \mathbf{q}\cdot\mathbf{n}\,dA$ through the boundary of a domain, with $\mathcal{N}(\boldsymbol{v})$ being the nonlocal counterpart of the local flux density $\mathbf{q}\cdot\mathbf{n}$. We refer to [27] for additional details regarding the nonlocal vector calculus and results such as integration by parts and nonlocal Green's identities.

We introduce the nonlocal energy semi-norm, nonlocal energy space, and nonlocal volume-constrained energy space

$$
\begin{aligned}
|||v|||^2 &:= \int_{\overline{\overline{\Omega}}}\int_{\overline{\overline{\Omega}}} (u(\boldsymbol{y})-u(\boldsymbol{x}))^2\gamma(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}\,d\boldsymbol{x}\\
V(\overline{\overline{\Omega}}) &:= \left\{ v\in L^2(\overline{\overline{\Omega}})\ :\ |||v|||_{\overline{\overline{\Omega}}}<\infty \right\}\\
V_\Lambda(\overline{\overline{\Omega}}) &:= \left\{ v\in V(\overline{\overline{\Omega}})\ :\ v=0 \text{ on } \Lambda\subset\Omega_I \right\}.
\end{aligned}
\tag{5}
$$

We also define the volume-trace space $\widetilde{V}_\Lambda(\overline{\overline{\Omega}}) := \{v|_\Lambda\ :\ v\in V(\overline{\overline{\Omega}})\}$, for $\Lambda\subset\Omega_I$, and the dual spaces $V'(\overline{\overline{\Omega}})$ and $V'_\Lambda(\overline{\overline{\Omega}})$ with respect to L^2-duality pairings.

We consider kernels such that the corresponding energy norm satisfies a Poincaré-like inequality, i.e., $\|v\|_{0,\overline{\overline{\Omega}}} \le C_{pn}|||v|||$ for all $v\in V_\Lambda(\overline{\overline{\Omega}})$, where C_{pn} is the nonlocal Poincaré constant. For such kernels, the paper [37] shows that C_{pn} is independent of δ if $\delta\in(0,\delta_0]$ for a given δ_0. In this paper, we consider a specific class of kernels, namely, integrable kernels such that there exist positive constants γ_1 and γ_2 for which $\gamma_1 \le \int_{\overline{\overline{\Omega}}}\gamma(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}$ and $\int_{\overline{\overline{\Omega}}}\gamma^2(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}\le\gamma_2^2$ for all $\boldsymbol{x}\in\Omega$. In this setting, $V(\overline{\overline{\Omega}})$ and $V_\Lambda(\overline{\overline{\Omega}})$ are equivalent to $L^2(\overline{\overline{\Omega}})$ and $L^2_c(\overline{\overline{\Omega}})$, and the operator $\mathcal{L}$ is such that $\mathcal{L}: L^2(\overline{\overline{\Omega}})\to L^2(\overline{\overline{\Omega}})$ [26].

Strong Form We introduce the strong form of a nonlocal Poisson's problem with Dirichlet or mixed volume constraints. We refer, again, to the configuration in Fig. 1 (left) and recall that $\Omega_I=\Omega_{nloc}\cup\Omega_{loc}$ such that $\Omega_{nloc}\cap\Omega_{loc}=\emptyset$. For $s\in V'(\overline{\overline{\Omega}})$, $v_n\in\widetilde{V}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, and $w_n\in\widetilde{V}_{\Omega_{loc}}(\overline{\overline{\Omega}})$, we define the *Dirichlet Poisson's problem* as follows: find $u_n\in V(\overline{\overline{\Omega}})$ such that

$$
\begin{cases}
-\mathcal{L}^{NL}u_n = s & \boldsymbol{x}\in\Omega\\
\qquad u_n = w_n & \boldsymbol{x}\in\Omega_{loc}\\
\qquad u_n = v_n & \boldsymbol{x}\in\Omega_{nloc},
\end{cases}
\tag{6}
$$

where $(6)_2$ and $(6)_3$ are two distinct Dirichlet volume constraints. Similarly, given $s\in V'(\overline{\overline{\Omega}})$, $v_n\in\widetilde{V}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, and $g_n\in V'(\Omega_{loc})$, we define the *mixed Poisson's problem* as follows: find $u_n\in V(\overline{\overline{\Omega}})$ such that

$$\begin{cases} -\mathcal{L}^{NL} u_n = s & \boldsymbol{x} \in \Omega \\ -\mathcal{N} u_n = g_n & \boldsymbol{x} \in \Omega_{loc} \\ u_n = v_n & \boldsymbol{x} \in \Omega_{nloc}, \end{cases} \tag{7}$$

where (7)$_2$ is the nonlocal counterpart of a flux condition, i.e., a Neumann boundary condition. As such, we refer to it as Neumann volume constraint.

Weak Form With the purpose of analyzing the δ-convergence of our strategies, we also introduce the weak form of problems (6) and (7). By multiplying both equations by a test function and using nonlocal integration by parts [26], we obtain the following weak formulations.

For $s \in V'(\overline{\overline{\Omega}})$, $v_n \in \widetilde{V}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, and $w_n \in \widetilde{V}_{\Omega_{loc}}(\overline{\overline{\Omega}})$, we define the *Dirichlet Poisson's problem* as follows: find $u_n \in V_c(\overline{\overline{\Omega}})$ such that $u_n = w_n$ in Ω_{loc}, $u_n = v_n$ in Ω_{nloc} and, for all $z \in V(\overline{\overline{\Omega}})$,

$$\int_{\overline{\overline{\Omega}}} \int_{\overline{\overline{\Omega}}} (u_n(\boldsymbol{x}) - u_n(\boldsymbol{y}))(z(\boldsymbol{x}) - z(\boldsymbol{y}))\gamma(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y}\, d\boldsymbol{x} = \int_{\Omega} s z\, d\boldsymbol{x}, \tag{8}$$

or, equivalently, $a(u, z) = F(z)$, where the bilinear form is given by $a(u, z) = \langle u, z\rangle_{V_{\Omega_I}}$. It can be shown [26] that for every $\gamma(\cdot, \cdot)$ satisfying the Poincaré inequality, $a(\cdot, \cdot)$ is coercive and continuous in $V_{\Omega_I}(\overline{\overline{\Omega}}) \times V_{\Omega_I}(\overline{\overline{\Omega}})$ and that $F(\cdot)$ is continuous in $V_{\Omega_I}(\overline{\overline{\Omega}})$. Thus, by the Lax–Milgram theorem, problem (8) is well-posed.

Similarly, given $s \in V'(\overline{\overline{\Omega}})$, $v_n \in \widetilde{V}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, and $g_n \in V'(\Omega_{loc})$, one can define the *mixed Poisson's problem* as follows: find $u_n \in V(\overline{\overline{\Omega}})$ such that $u_n = v_n$ in Ω_{nloc} and for all $z \in V_{\Omega_{nloc}}(\overline{\overline{\Omega}})$,

$$\int_{\overline{\overline{\Omega}}} \int_{\overline{\overline{\Omega}}} (u_n(\boldsymbol{x}) - u_n(\boldsymbol{y}))(z(\boldsymbol{x}) - z(\boldsymbol{y}))\gamma(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y}\, d\boldsymbol{x} = \int_{\Omega_{loc}} g_n z\, d\boldsymbol{x} + \int_{\Omega} s z\, d\boldsymbol{x}, \tag{9}$$

or, equivalently, $a(u, z) = F_{g_n}(z)$. Also in this case, it can be shown that $a(\cdot, \cdot)$ is coercive and continuous in $V_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, provided the kernel induces a Poincaré inequality. Furthermore, the functional F_{g_n} is continuous on $V_{\Omega_{nloc}}(\overline{\overline{\Omega}})$. Thus, by the Lax–Milgram theorem, problem (9) is also well-posed.

2.2 The Linear Peridynamic Solid Model

For the displacement function $\mathbf{u}(\boldsymbol{x})\colon \mathbb{R}^d \to \mathbb{R}^d$, we define the linear peridynamic solid (LPS) [28] operator[3] $\mathcal{L}^{LPS}\colon \mathbb{R}^d \to \mathbb{R}^d$ as

$$\begin{aligned}\mathcal{L}^{LPS}\mathbf{u}(\boldsymbol{x}) :=& \frac{C_1}{m(\delta)} \int_{\overline{\overline{\Omega}}} (\lambda - \mu)\, \gamma(|\mathbf{y}-\mathbf{x}|)\,(\mathbf{y}-\mathbf{x})\,(\theta(\mathbf{x})+\theta(\mathbf{y}))\,d\mathbf{y} \\ &+ \frac{C_2}{m(\delta)} \int_{\overline{\overline{\Omega}}} \mu\gamma(|\mathbf{y}-\mathbf{x}|)\frac{(\mathbf{y}-\mathbf{x})\otimes(\mathbf{y}-\mathbf{x})}{|\mathbf{y}-\mathbf{x}|^2}\,(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x}))\,d\mathbf{y},\end{aligned} \tag{10}$$

where the dilatation $\theta : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\theta(\mathbf{x}) := \frac{2}{m(\delta)} \int_{\overline{\overline{\Omega}}} \gamma(|\mathbf{y}-\mathbf{x}|)(\mathbf{y}-\mathbf{x})\cdot(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x}))\,d\mathbf{y}.$$

Here, for $d = 2$, $C_1 = 2$ and $C_2 = 16$. The kernel function γ is nonnegative and radial and satisfies the same assumptions as in (2). Furthermore, we consider kernels γ such that m, defined as

$$m(\delta) := \int_{B_\delta(\mathbf{x})} \gamma(|\mathbf{y}-\mathbf{x}|)\,|\mathbf{y}-\mathbf{x}|^2\,d\mathbf{y},$$

is bounded. This guarantees well-posedness of the volume-constrained problem associated with $\mathcal{L}^{LPS}$ [38]. The constants μ and λ are the shear and Lamé modulus, which, under the plane strain assumption [5], are related to the Young's modulus E and the Poisson ratio ν of a material, i.e., $\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}$, $\mu = \frac{E}{2(1+\nu)}$. It can be shown [38] that the LPS operator $\mathcal{L}^{LPS}$ converges to the Navier operator below:

$$\mathcal{L}_l\mathbf{u} := -\nabla\cdot(\lambda tr(\mathbf{E})\mathbf{I}+2\mu\mathbf{E}) = -(\lambda-\mu)\nabla[\mathrm{tr}(\mathbf{E})] - \mu\nabla\cdot(2\mathbf{E}+\mathrm{tr}(\mathbf{E})\mathbf{I}), \tag{11}$$

where $\mathbf{E} := \frac{1}{2}(\nabla\mathbf{u}+(\nabla\mathbf{u})^T)$ and $\mathrm{tr}(\mathbf{E}) = \nabla\cdot\mathbf{u}$. In particular, when the fourth-order derivatives of $\mathbf{u}$ are bounded, we have the following pointwise relationship:

$$\mathcal{L}^{LPS}\mathbf{u}(\boldsymbol{x}) = \mathcal{L}_l\mathbf{u}(\boldsymbol{x}) + \mathcal{O}(\delta^2). \tag{12}$$

For the LPS operator $\mathcal{L}^{LPS}$, we define the interaction domain of Ω as

$$\Omega_I = \{\boldsymbol{y} \in \mathbb{R}^d \setminus \Omega : \ \|\boldsymbol{y}-\boldsymbol{x}\| < 2\delta, \ \boldsymbol{x}\in\Omega\} \tag{13}$$

[3] Note that this model holds in the assumption of small displacements [28].

and set $\overline{\overline{\Omega}} = \Omega \cup \Omega_I$. Note that in this case, Ω_I is a layer of thickness 2δ surrounding Ω; this is due to the presence of a double integral in the definition of the operator. As before, Ω_I is the volume where nonlocal boundary conditions must be prescribed to guarantee the well-posedness of the nonlocal equation associated with $\mathcal{L}^{LPS}$. We refer to Fig. 1 (right) for an illustration of a two-dimensional domain, the support of γ, and the induced interaction domain. The same division as in Sect. 2.1 into a nonoverlapping partition is performed.

For the prescription of nonlocal flux conditions, we consider the following nonlocal flux operator for the LPS model. Let $\boldsymbol{x} \in \Lambda \subset \Omega_I$, and we have

$$\begin{aligned}\mathcal{N}^{LPS}\mathbf{u}(\boldsymbol{x}) :=& \frac{C_1}{m(\delta)} \int_{\overline{\overline{\Omega}}} (\lambda - \mu)\, \gamma(|\mathbf{y}-\mathbf{x}|)\, (\mathbf{y}-\mathbf{x})\, (\theta(\mathbf{x}) + \theta(\mathbf{y}))\, d\mathbf{y} \\ &+ \frac{C_2}{m(\delta)} \int_{\overline{\overline{\Omega}}} \mu\gamma(|\mathbf{y}-\mathbf{x}|) \frac{(\mathbf{y}-\mathbf{x}) \otimes (\mathbf{y}-\mathbf{x})}{|\mathbf{y}-\mathbf{x}|^2}\, (\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{x}))\, d\mathbf{y},\end{aligned} \tag{14}$$

where θ and m are defined as above. For more details on nonlocal flux conditions for nonlocal mechanics problems, we refer the interested reader to [34].

As for the Laplacian operator, we introduce the energy norm and the corresponding spaces [38].

$$\begin{aligned}|||\mathbf{u}|||^2_{LPS} =& \frac{1}{m(\delta)} \int_{\overline{\overline{\Omega}}} \int_{\overline{\overline{\Omega}} \cap B_\delta(\mathbf{x})} \frac{\gamma(|\mathbf{y}-\mathbf{x}|)}{|\mathbf{y}-\mathbf{x}|^2} \left[(\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{x})) \cdot (\mathbf{y}-\mathbf{x})\right]^2 d\mathbf{y}\, d\mathbf{x}, \\ V^{LPS}(\overline{\overline{\Omega}}) :=& \left\{\mathbf{u} \in [L^2(\overline{\overline{\Omega}})]^d : |||\mathbf{u}|||_{LPS} < \infty\right\} \\ V^{LPS}_{\Lambda}(\overline{\overline{\Omega}}) :=& \left\{\mathbf{u} \in V^{LPS}(\overline{\overline{\Omega}}) : \mathbf{u} = \mathbf{0} \text{ on } \Lambda \subset \Omega_I\right\}.\end{aligned} \tag{15}$$

Note that $|||\mathbf{u}|||_{LPS} = 0$ if and only if $\mathbf{u}$ represents an infinitesimally rigid displacement, i.e.,

$$\mathbf{u}(\boldsymbol{x}) \in \{\mathbb{Q}\boldsymbol{x} + \mathbf{b}, \mathbb{Q} \in \mathbb{R}^{d\times d}, \mathbb{Q}^T = -\mathbb{Q}, \mathbf{b} \in \mathbb{R}^d\}.$$

We also define the volume-trace space $\widetilde{V}^{LPS}_{\Lambda}(\overline{\overline{\Omega}}) := \{v|_\Lambda : v \in V^{LPS}(\overline{\overline{\Omega}})\}$, for $\Lambda \subset \Omega_I$, and the dual spaces $(V^{LPS})'(\overline{\overline{\Omega}})$ and $(V^{LPS})'_{\Lambda}(\overline{\overline{\Omega}})$ with respect to L^2-duality pairings. Note that when γ is an integrable function, similarly to the nonlocal Laplacian operator, the LPS operator acts as a map from $[L^2(\overline{\overline{\Omega}})]^d$ to $[L^2(\overline{\overline{\Omega}})]^d$.

Strong Form We introduce the strong form of the LPS problem with Dirichlet or mixed volume constraints. We refer, again, to the configuration in Fig. 1 (right). For $\mathbf{s} \in (V^{LPS})'(\overline{\overline{\Omega}})$, $\mathbf{v}_n \in \widetilde{V}^{LPS}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, and $\mathbf{w}_n \in \widetilde{V}^{LPS}_{\Omega_{loc}}(\overline{\overline{\Omega}})$, we define the *Dirichlet LPS problem* as follows: find $\mathbf{u}_n \in V^{LPS}(\overline{\overline{\Omega}})$ such that

$$\begin{cases} -\mathcal{L}^{LPS}\mathbf{u}_n = \mathbf{s} & \boldsymbol{x} \in \Omega \\ \quad\quad\ \mathbf{u}_n = \mathbf{w}_n & \boldsymbol{x} \in \Omega_{loc} \\ \quad\quad\ \mathbf{u}_n = \mathbf{v}_n & \boldsymbol{x} \in \Omega_{nloc}, \end{cases} \tag{16}$$

where $(16)_2$ and $(16)_3$ are distinct Dirichlet volume constraints. Similarly, given $\mathbf{s} \in (V^{LPS})'(\overline{\overline{\Omega}})$, $\mathbf{v}_n \in \widetilde{V}^{LPS}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, and $\mathbf{g}_n \in (V^{LPS})'(\Omega_{loc})$, we define the *mixed LPS problem* as follows: find $\mathbf{u}_n \in V^{LPS}(\overline{\overline{\Omega}})$ such that

$$\begin{cases} -\mathcal{L}^{LPS}\mathbf{u}_n = \mathbf{s} & \boldsymbol{x} \in \Omega \\ -\mathcal{N}^{LPS}\mathbf{u}_n = \mathbf{g}_n & \boldsymbol{x} \in \Omega_{loc} \\ \quad\quad\ \mathbf{u}_n = \mathbf{v}_n & \boldsymbol{x} \in \Omega_{nloc}. \end{cases} \tag{17}$$

Weak Form With the purpose of analyzing the δ-convergence of our strategies, we also introduce the weak form of problems (16) and (17). For clarity, and to avoid heavy notation, we present the formulations in the scalar setting. We first introduce the following integration by parts result [23, 26]: for every $u \in V^{LPS}(\overline{\overline{\Omega}})$ and $z \in V^{LPS}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, we have

$$\begin{aligned} &\int_{\Omega} -\mathcal{L}^{LPS}u(\boldsymbol{x})z(\boldsymbol{x})\,d\boldsymbol{x} \\ &\quad = \frac{C_1 d\,(\lambda-\mu)}{(m(\delta))^2}\int_{\overline{\overline{\Omega}}}\left[\int_{\overline{\overline{\Omega}}}\gamma(|\mathbf{y}-\mathbf{x}|)(\mathbf{y}-\mathbf{x})\cdot(u(\mathbf{y})-u(\mathbf{x}))\,d\mathbf{y}\right]\times \\ &\qquad\qquad\qquad\qquad \left[\int_{\overline{\overline{\Omega}}}\gamma(|\mathbf{y}-\mathbf{x}|)(\mathbf{y}-\mathbf{x})\cdot(z(\mathbf{y})-z(\mathbf{x}))\,d\mathbf{y}\right]d\mathbf{x} \\ &\quad + \frac{C_2\mu}{2m(\delta)}\int_{\overline{\overline{\Omega}}}\int_{\overline{\overline{\Omega}}}\gamma(|\mathbf{y}-\mathbf{x}|)(u(\mathbf{y})-u(\mathbf{x}))(z(\boldsymbol{y})-z(\boldsymbol{x}))d\mathbf{y}d\mathbf{x} \\ &\quad + \int_{\Omega_I}\mathcal{N}^{LPS}u(\boldsymbol{x})z(\boldsymbol{x})\,d\boldsymbol{x} \\ &\quad := a^{LPS}(u,z) + \int_{\Omega_{loc}}\mathcal{N}^{LPS}u(\boldsymbol{x})z(\boldsymbol{x})\,d\boldsymbol{x}. \end{aligned} \tag{18}$$

It is important to note that the bilinear form $a(\cdot,\cdot)$ induces a norm in the space $V^{LPS}_{\Lambda}(\overline{\overline{\Omega}})$, for all $\Lambda \subset \Omega_I$, or, in other words, $a(u,u)$ is equivalent to $|||u|||^2_{LPS}$ for all $u \in V^{LPS}_{\Lambda}(\overline{\overline{\Omega}})$. Thus, $a(\cdot,\cdot)$ is continuous and coercive.

By multiplying both Eqs. (16) and (17) by a test function and using nonlocal integration by parts, we obtain the following weak formulations. For $s \in (V^{LPS})'(\overline{\overline{\Omega}})$,

$v_n \in \widetilde{V}^{LPS}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, and $w_n \in \widetilde{V}^{LPS}_{\Omega_{loc}}(\overline{\overline{\Omega}})$, $u_n \in V^{LPS}(\overline{\overline{\Omega}})$ is a weak solution of the *Dirichlet LPS problem* if $u_n = w_n$ in Ω_{loc}, $u_n = v_n$ in Ω_{nloc}, and

$$a^{LPS}(u, z) = \int_{\Omega} sz\, d\boldsymbol{x}, \qquad \forall\, z \in V^{LPS}_{\Omega_I}(\overline{\overline{\Omega}}). \tag{19}$$

Similarly, given $s \in (V^{LPS})'(\overline{\overline{\Omega}})$, $v_n \in \widetilde{V}^{LPS}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$, and $g_n \in (V^{LPS})'(\Omega_{loc})$, $u_n \in V^{LPS}(\overline{\overline{\Omega}})$ is a weak solution of the *mixed LPS problem* if $u_n = v_n$ in Ω_{nloc} and

$$a^{LPS}(u, z) = \int_{\Omega_{loc}} g_n z\, d\boldsymbol{x} + \int_{\Omega} sz\, d\boldsymbol{x}, \qquad \forall\, z \in V^{LPS}_{\Omega_{nloc}}(\overline{\overline{\Omega}}). \tag{20}$$

The well-posedness of (19) and (20) follows from the fact that $a^{LPS}(\cdot, \cdot)$ is continuous and coercive in $V^{LPS}(\overline{\overline{\Omega}})$ and from the continuity of the right-hand sides. In fact, these properties allow us to apply the Lax–Milgram theorem that guarantees existence and uniqueness of solutions.

3 Proposed Strategies

In practice, data may only be available on the boundary $\partial\overline{\overline{\Omega}}$ and not in Ω_I; in particular, the values of the diffusive quantity, for the nonlocal Poisson's equation, and of the displacement, for the LPS model, may be available on parts of $\partial\overline{\overline{\Omega}}$, while nonlocal volume constraints may be available on the remaining part of Ω_I. Thus, as indicated in Fig. 1, we split the interaction domain into two parts: a "nonlocal part," Ω_{nloc}, where nonlocal volume constraints are available, and a "local part," Ω_{loc}, where only local, boundary data are available. As this is not enough for the well-posedness of the problem, we now introduce a strategy that, starting from this incomplete dataset, delivers volume constraints on Ω_{loc}, hence allowing for the solution of the nonlocal problems. We present our strategies for the nonlocal Poisson equation, as the approach is identical for the LPS model (the properties of the method are analyzed for both models).

Assumption 1 Only the following data are available:

1. $w_l \in H^{\frac{1}{2}}(\Gamma_{loc})$: *local* Dirichlet boundary data on $\Gamma_{loc} = \partial\Omega_{loc} \cap \partial\overline{\overline{\Omega}}$
2. $v_n \in \widetilde{V}_{\Omega_{nloc}}(\overline{\overline{\Omega}})$: nonlocal Dirichlet data in Ω_{nloc}
3. $s \in V'(\overline{\overline{\Omega}})$: forcing term over $\overline{\overline{\Omega}}$

We design two strategies to automatically convert w_l into a nonlocal volume constraint (either of Dirichlet or of Neumann type) on Ω_{loc}. As we show in the following section, the most important property of our strategies is their *asymptotic compatibility*, i.e.,

$$u_n \to u_l \ \text{ as } \ \delta \to 0 \quad \text{in} \ \ V(\overline{\overline{\Omega}}) \ \ \text{and} \ \ L^2(\overline{\overline{\Omega}}). \tag{21}$$

Here, u_n is the nonlocal solution corresponding to the proposed nonlocal volume constraints and u_l is the solution of the following Poisson's equation:

$$\begin{cases} -\Delta u_l = s \ \boldsymbol{x} \in \overline{\overline{\Omega}} \\ u_l = w_l \quad \boldsymbol{x} \in \Gamma_{loc} \\ u_l = v_n \quad \boldsymbol{x} \in \Gamma_{nloc}, \end{cases} \tag{22}$$

i.e., the solution of the local problem with boundary data as in **Assumption 1** on Ω_{loc} and with boundary data $v_n|_{\Gamma_{nloc}}$, with $\Gamma_{nloc} = \partial\Omega_{nloc} \cap \partial\overline{\overline{\Omega}}$. Note that, by prescribing the Dirichlet condition on Γ_{nloc}, we are assuming that $v_n|_{\Gamma_{nloc}}$ exists and is such that $v_n|_{\Gamma_{nloc}} \in H^{\frac{1}{2}}(\Gamma_D)$. We emphasize that we are not assuming $v_n \in H^1(\Omega_{nloc})$, but only that v_n has a well-defined trace on Γ_{nloc}.

3.1 Dirichlet-to-Dirichlet Strategy

The first proposed strategy, referred to as Dirichlet-to-Dirichlet (DtD) strategy, consists in using the local solution u_l of problem (22) as Dirichlet volume constraint for the nonlocal problem in Ω_{loc}. We summarize the procedure below:

1. Solve the local problem (22) to obtain u_l. Note that $u_l \in \widetilde{V}(\Omega_{loc})$.
2. Solve the (well-posed) nonlocal problem:

$$\begin{cases} -\mathcal{L}^{NL} u_n = s \quad \boldsymbol{x} \in \Omega \\ \qquad\quad u_n = u_l \ \boldsymbol{x} \in \Omega_{loc} \\ \qquad\quad u_n = v_n \ \boldsymbol{x} \in \Omega_{nloc}. \end{cases} \tag{23}$$

3.2 Dirichlet-to-Neumann Strategy

The second strategy, referred to as Dirichlet-to-Neumann (DtN) strategy, consists in using the local solution u_l of problem (22) to generate a Neumann volume constraint for the nonlocal problem in Ω_{loc}. We summarize the procedure below:

1. Solve the local problem (22) to obtain u_l. Note that $\mathcal{N}^{NL} u_l$ for $\boldsymbol{x} \in \Omega_{loc}$ is well-defined and belongs to $V'(\Omega_{loc})$.

2. Solve the (well-posed) nonlocal problem:

$$\begin{cases} -\mathcal{L}^{NL} u_n = s & \boldsymbol{x} \in \Omega \\ -\mathcal{N}^{NL} u_n = -\mathcal{N}^{NL} u_l & \boldsymbol{x} \in \Omega_{loc} \\ \qquad u_n = v_n & \boldsymbol{x} \in \Omega_{nloc}. \end{cases} \tag{24}$$

4 Convergence to the Local Limit

In this section, we study the limiting behavior of the solution as the nonlocal interactions vanish, i.e., as $\delta \to 0$, and we show that (21) holds true with a second-order convergence rate for both Poisson's and LPS models.

For both the Dirichlet-to-Dirichlet strategy and Dirichlet-to-Neumann, the following propositions provide bounds for the errors:

$$\begin{aligned} e_{E,NL} &= |||u_n - u_l|||, \quad & e_{E,LPS} &= |||\mathbf{u}_n - \mathbf{u}_l|||_{LPS}, \\ e_{0,NL} &= \|u_n - u_l\|_{0,\overline{\overline{\Omega}}}, \quad & e_{0,LPS} &= \|\mathbf{u}_n - \mathbf{u}_l\|_{0,\overline{\overline{\Omega}}}. \end{aligned} \tag{25}$$

Theorem 1 *Let* $\delta_0 \in (0, \infty)$ *and* $\mathcal{U}_l := \{u_l \in C^4(\overline{\overline{\Omega}}) : u_l \text{ solves (22) for } \delta \in (0, \delta_0]\}$ *be solutions to* (22)*. Then,*

$$e_{E,NL} = \mathcal{O}(\delta^2). \tag{26}$$

Proof We only prove (26) for the DtD strategy and refer the reader to [22] for the DtN strategy as the steps of the proof are the same. In fact, for DtN, the only difference with the approach presented in that paper is step **1** (solution of a local problem), where, instead of solving a mixed boundary condition Poisson's problem, we solve a fully Dirichlet problem.

By definition of u_n and u_l, we have

$$\begin{cases} -\mathcal{L} u_n = s = -\Delta u_l & \boldsymbol{x} \in \Omega \\ u_n = u_l & \boldsymbol{x} \in \Omega_{loc} \\ u_n = v_n & \boldsymbol{x} \in \Omega_{nloc}. \end{cases} \tag{27}$$

We introduce a nonlocal auxiliary problem for the local solution u_l, keeping in mind that v_n is compatible with the local solution.

$$\begin{cases} -\mathcal{L}u_l = s_l = -\displaystyle\int_{\overline{\overline{\Omega}}}(u_l(\boldsymbol{y}) - u_l(\boldsymbol{x}))\gamma(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y} & \boldsymbol{x}\in\Omega \\ u_l = u_l & \boldsymbol{x}\in\Omega_{loc} \\ u_l = v_n & \boldsymbol{x}\in\Omega_{nloc}. \end{cases} \tag{28}$$

In order to estimate $e_{E,NL}$, we first consider the pointwise difference $s(\boldsymbol{x}) - s_l(\boldsymbol{x})$. Property (4) implies that

$$|s(\boldsymbol{x}) - s_l(\boldsymbol{x})| = \left|\int_{\overline{\overline{\Omega}}}(u_l(\boldsymbol{y}) - u_l(\boldsymbol{x}))\gamma(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y} - \Delta u_l\right| = O(\delta^2). \tag{29}$$

Next, we consider the weak forms of (27) and (28) and use, in both of them, the test function $z\in V_{\Omega_I}(\overline{\overline{\Omega}})$; we have

$$\int_{\overline{\overline{\Omega}}}\int_{\overline{\overline{\Omega}}}(u_n(\boldsymbol{x}) - u_n(\boldsymbol{y}))(z(\boldsymbol{x}) - z(\boldsymbol{y}))\gamma(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}\,d\boldsymbol{x} = \int_\Omega s\,z\,d\boldsymbol{x}, \tag{30}$$

$$\int_{\overline{\overline{\Omega}}}\int_{\overline{\overline{\Omega}}}(u_l(\boldsymbol{x}) - u_l(\boldsymbol{y}))(z(\boldsymbol{x}) - z(\boldsymbol{y}))\gamma(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}\,d\boldsymbol{x} = \int_\Omega s_l\,z\,d\boldsymbol{x}. \tag{31}$$

Subtraction gives

$$\int_{\overline{\overline{\Omega}}}\int_{\overline{\overline{\Omega}}}(u_n(\boldsymbol{x}) - u_l(\boldsymbol{x}) - u_n(\boldsymbol{y}) + u_l(\boldsymbol{y}))(z(\boldsymbol{x}) - z(\boldsymbol{y}))\gamma(\boldsymbol{x},\boldsymbol{y})\,d\boldsymbol{y}\,d\boldsymbol{x} = \int_\Omega (s - s_l)\,z\,d\boldsymbol{x}.$$

To prove the error estimate, we then choose $z = u_n - u_l \in V_{\Omega_I}(\overline{\overline{\Omega}})$. We have

$$|||u_n - u_l|||^2 \le \int_\Omega (s - s_l)\,(u_n - u_l)\,d\boldsymbol{x} \le \|s - s_l\|_{0,\Omega}\|\widetilde{u}_n - u_l\|_{0,\Omega} \le O(\delta^2)C_{pn}|||u_n - u_l|||.$$

By dividing both sides by $|||u_n - u_l|||$, the error bound follows. □

Before addressing the error bound for the LPS model, we introduce the local problem corresponding to the operator $\mathcal{L}_l$ introduced in (11), i.e.,

$$\begin{cases} -\mathcal{L}_l\mathbf{u}_l = \mathbf{s} & \boldsymbol{x}\in\overline{\overline{\Omega}} \\ \mathbf{u}_l = \mathbf{w}_l & \boldsymbol{x}\in\Gamma_{loc} \\ \mathbf{u}_l = \mathbf{v}_n & \boldsymbol{x}\in\Gamma_{nloc}, \end{cases} \tag{32}$$

where $\mathbf{w}_l$ is the available local Dirichlet data on Γ_{loc}, and $\mathbf{v}_n$ is the available nonlocal Dirichlet volume constraint on $\Gamma_{nloc} = \partial\Omega_{nloc}\cap\partial\overline{\overline{\Omega}}$. As for the local Poisson's equation, we assume that the nonlocal Dirichlet data $\mathbf{v}_n$ has a well-defined trace on Γ_{nloc} and is compatible with the local solution. We can now state the following theorem, whose proof, based on (12), follows exactly the same steps used in Theorem 1 and is, hence, omitted.

Theorem 2 *Let* $\delta_0 \in (0,\infty)$ *and* $\mathcal{U}_l^{LPS} := \{\mathbf{u}_l \in C^4(\overline{\overline{\Omega}}) : \mathbf{u}_l$ solves (32) for $\delta \in (0,\delta_0]\}$ *be solutions to* (32). *Then,*

$$e_{E,LPS} = \mathcal{O}(\delta^2). \tag{33}$$

Remark 1 An immediate consequence of Theorem 1 implies that the convergence rate of e_0 is at least quadratic. This result can be obtained by applying the Poincaré inequality, i.e.

$$e_{0,NL} = \|u_n - u_l\|_{0,\overline{\overline{\Omega}}} \leq C_{pn}|||u_n - u_l||| = C_{pn}e_{E,NL} = \mathcal{O}(\delta^2).$$

Following the same arguments, we can also show that the same bound holds for the LPS model. In fact, paper [38] provides a Poincaré-type inequality associated with the LPS operator $\mathcal{L}^{LPS}$ with constant C_{pn}^{LPS}. Thus, as a consequence of Theorem 2, we have

$$e_{0,LPS} = \|\mathbf{u}_n - \mathbf{u}_l\|_{0,\overline{\overline{\Omega}}} \leq C_{pn}^{LPS}|||\mathbf{u}_n - \mathbf{u}_l||| = C_{pn}^{LPS}e_{E,LPS} = \mathcal{O}(\delta^2).$$

5 Numerical Tests

We report the results of several two-dimensional numerical tests that illustrate our theoretical results and highlight the efficacy of the proposed methods.

In all tests, we utilize a particle discretization of the strong form of the nonlocal Poisson's problem and the LPS model introduced in Sects. 2.1 and 2.2, respectively. The meshfree discretization method we use is based on an optimization-based quadrature rule developed and analyzed in [50, 54, 55, 59]. In this approach, we discretize the union of the domain and interaction domain, $\overline{\overline{\Omega}}$, by a collection of points

$$\chi_h = \{\boldsymbol{x}_i\}_{\{i=1,2,\cdots,M\}} \subset \overline{\overline{\Omega}}$$

and then solve for the solution $u_{(i)} \approx u_n(\boldsymbol{x}_i)$ at $\boldsymbol{x}_i \in \chi_h$ using a one-point quadrature rule. Although the method can be applied to more general grids, in all numerical tests below, we require χ_h to be a uniform Cartesian grid:

$$\chi_h := \{(k_{(1)}h, \cdots, k_{(d)}h)|\boldsymbol{k} = (k_{(i)}, \cdots, k_{(d)}) \in \mathbb{Z}^d\} \cap \overline{\overline{\Omega}}.$$

Here, h is the spatial grid size. To maintain an easily scalable implementation, in our δ-convergence studies [4], we assume h to be chosen such that the ratio $\frac{\delta}{h}$ is bounded by a constant as $\delta \to 0$. This meshfree discretization method based on optimization-based quadrature rules features simplicity in implementation and is asymptotically compatible, i.e., it is such that the nonlocal solution converges to its local counterpart as $\delta, h \to 0$. For further implementation details, we refer the interested reader to [29, 59].

5.1 Consistency Tests for the Nonlocal Poisson's Equation

Theorem 1 implies that when the data are smooth enough to have $\mathcal{L}^{NL} u_l = \Delta u_l$, then $u_n = u_l$. We use this observation to conduct a consistency test for the proposed method. Indeed, we consider local solutions u_l such that $\mathcal{L} u_l = \Delta u_l$ and expect to observe that the local and nonlocal solutions coincide (up to discretization error).

We refer to the two-dimensional configuration reported in Fig. 2. Here, $\Omega = (0, 1)^2$ and Ω_I is a layer of thickness δ surrounding the domain. We use two different configurations for the DtD and DtN strategy. For the former, we refer to the configuration on the left of Fig. 2, where $\Omega_I = \Omega_{loc}$, whereas for the latter we refer to the configuration on the right where Ω_{loc} only covers the right side of the interaction domain, i.e., $\Omega_{loc} = [1, 1 + \delta] \times [0, 1]$. In all our consistency tests, we use the constant kernel

$$\gamma(\boldsymbol{x}, \boldsymbol{y}) = \frac{4}{\pi \delta^4} \mathcal{X}_{B_\delta(\boldsymbol{x})}(\boldsymbol{y}) \tag{34}$$

and the following set of solutions:

- $f(\boldsymbol{x}) = 0$, $u_l(\boldsymbol{x}) = \boldsymbol{x}_1 + \boldsymbol{x}_2$ on $\partial\Omega$, $u_n(\boldsymbol{x}) = \boldsymbol{x}_1 + \boldsymbol{x}_2$ on Ω_{nloc}. Note that this solution corresponds to $u_l = \boldsymbol{x}_1 + \boldsymbol{x}_1$.

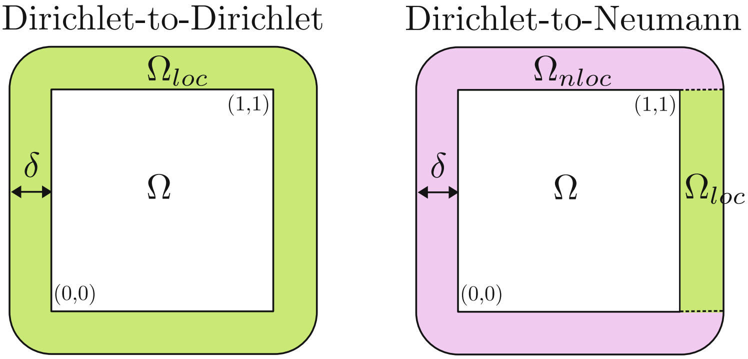


Fig. 2 Two-dimensional configuration utilized in the nonlocal Poisson's consistency and convergence tests for the DtD strategy (left) and DtN strategy (right)

- $f(\boldsymbol{x}) = -6(\boldsymbol{x}_1 + \boldsymbol{x}_2)$, $u_l(\boldsymbol{x}) = \boldsymbol{x}_1^3 + \boldsymbol{x}_2^3$ on $\partial\Omega$, $u_n(\boldsymbol{x}) = \boldsymbol{x}_1^3 + \boldsymbol{x}_2^3$ on Ω_{nloc}. Note that this solution corresponds to $u_l = \boldsymbol{x}_1^3 + \boldsymbol{x}_2^3$.

Consistently with our theory, in both cases and for both strategies (i.e., DtD and DtN), the nonlocal solution coincides with the local solution up to *machine precision*. In fact, we observe $e_0 \approx O(10^{-17})$. Note that this is possible because our meshfree discretization method can reproduce exactly both linear and cubic polynomials.

5.2 Convergence Tests for the Nonlocal Poisson's Equation

We test the convergence of u_n to the local solution u_l as $\delta \to 0$. For the same constant kernel defined in (34) and for the same configurations illustrated in Fig. 2, we consider the following set of solutions:

- $f(\boldsymbol{x}) = -2\sin(\boldsymbol{x}_1)\cos(\boldsymbol{x}_2)$, $u_l(\boldsymbol{x}) = \sin(\boldsymbol{x}_1)\cos(\boldsymbol{x}_2)$ for $\mathbf{x} \in \partial\Omega$, and $u_n(\boldsymbol{x}) = \sin(\boldsymbol{x}_1)\cos(\boldsymbol{x}_2)$ for $\boldsymbol{x} \in \Omega_{nloc}$; the corresponding local solution is $u_l(\boldsymbol{x}) = \sin(\boldsymbol{x}_1)\cos(\boldsymbol{x}_2)$.
- $f(\boldsymbol{x}) = -12(\boldsymbol{x}_1^2 + \boldsymbol{x}_2^2)$, $u_l(\boldsymbol{x}) = \boldsymbol{x}_1^4 + \boldsymbol{x}_2^4$ for $\boldsymbol{x} \in \partial\Omega$ and $u_n(\boldsymbol{x}) = \boldsymbol{x}_1^4 + \boldsymbol{x}_2^4$ for $\boldsymbol{x} \in \Omega_{nloc}$; the corresponding local solution is given by $u_l = \boldsymbol{x}_1^4 + \boldsymbol{x}_2^4$.

Convergence results are reported in Table 1 for the DtD strategy and in Table 2 for the DtN strategy. Here, we report, for decreasing values of δ, the L^2 norm of the difference between local and nonlocal solutions, i.e., e_0 and the corresponding rate of convergence. We recall that in our discretization scheme δ and the node spacing h are related, i.e., their ratio is constant and it is set to 2.5 for the sinusoidal solution and to 3.1 for the polynomial one. In both cases, the smallest h is set to 0.1 and then halved at every run. The observed *quadratic* rates are in alignment with our theory, see Remark 1. We point out that the faster converge of the DtD strategy is due to the fact that the nonlocal solution is closer (by construction) to the local one. In fact, they coincide on the interaction domain.

Table 1 For the nonlocal Poisson's equation, L^2-norm errors and convergence rates for the DtD strategy

Sinusoidal			Polynomial		
δ	e_0	Rate	δ	e_0	Rate
0.25	1.837e–4	–	0.31	9.571e–3	–
0.125	4.443e–5	2.0473	0.155	2.198e–3	2.1226
0.0625	1.098e–5	2.0174	0.0775	5.290e–3	2.0547
0.03125	2.730e–6	2.0071	0.0388	1.299e–4	2.0255

Table 2 For the nonlocal Poisson's equation, L^2-norm errors and convergence rates for the DtN strategy

Sinusoidal			Polynomial		
δ	e_0	Rate	δ	e_0	Rate
0.25	2.551e–4	–	0.31	1.094e–2	–
0.125	7.257e–5	1.8136	0.155	2.929e–3	1.9014
0.0625	1.953e–5	1.9455	0.0775	7.720e–4	1.9239
0.03125	5.069e–6	1.8941	0.0388	1.990e–4	1.9561

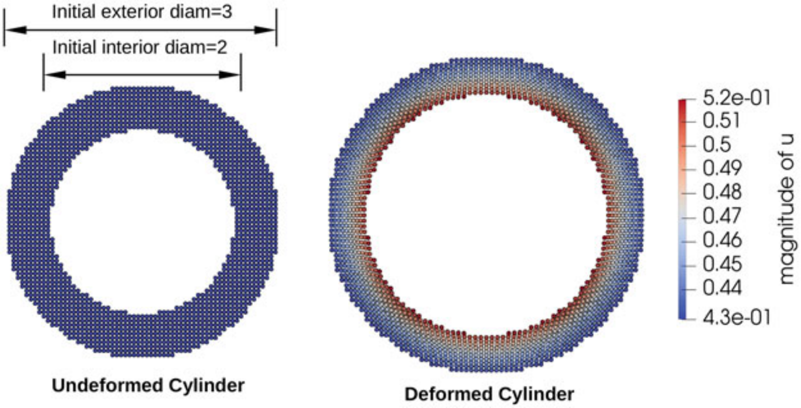


Fig. 3 Two-dimensional hollow cylinder problem settings

5.3 Numerical Tests for the LPS Model

We consider the LPS model introduced in Sect. 2.2, and we test consistency and convergence with respect to δ of both strategies. In all our tests, we consider the deformation of a hollow cylinder as illustrated in Fig. 3 and refer the two-dimensional configurations reported in Fig. 4 for details on the domain parameters. Specifically, we set $\Omega = B_{1.5}(\mathbf{0}) \setminus B_1(\mathbf{0})$. The interaction domain is then defined as a layer of thickness 2δ surrounding the disc, both inside and outside. For the DtD strategy we use the configuration on the left where $\Omega_{loc} = \Omega_I$, i.e., we assume that only local boundary conditions are available. For the DtN strategy, we consider the configuration on the right where Ω_{loc} only corresponds to the inner portion of the interaction domain, i.e., $\Omega_{loc} = B_1(\mathbf{0}) \setminus B_{1-2\delta}(\mathbf{0})$.

To test the consistency of both procedures, we consider the linear function $\mathbf{u}_l = [10\boldsymbol{x}_1 + 2\boldsymbol{x}_2, 3\boldsymbol{x}_1 + 4\boldsymbol{x}_2]$. This function is such that $\mathcal{L}^{LPS}\mathbf{u}_l = \mathcal{L}_l\mathbf{u}_l$, where $\mathcal{L}^{LPS}$ and $\mathcal{L}_l$ are defined as in (10) and (11), respectively. Thus, as for the nonlocal Poisson's model, we expect the nonlocal solution obtained with both the DtD and DtN procedures to be such that $\mathbf{u}_n = \mathbf{u}_l$. Our results indicate, once again, that the two solutions are identical, up to *machine precision*, i.e., $e_0 = \mathcal{O}(10^{-17})$.

To test the convergence with respect to δ, we consider an analytic solution of the local Navier equation (32). Under a plane strain assumption and subject to an

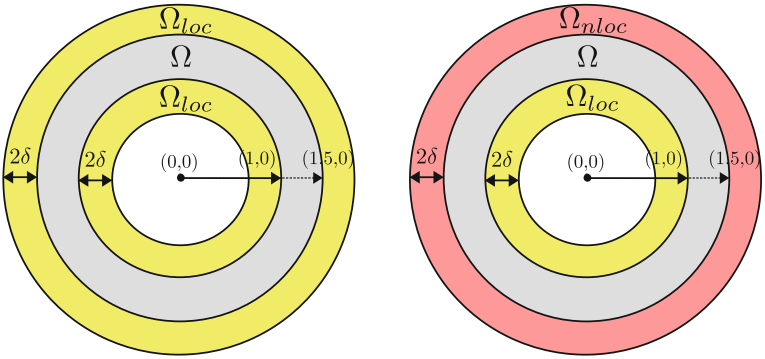


Fig. 4 Two-dimensional configuration utilized in the LPS consistency and convergence tests for the DtD strategy (left) and DtN strategy (right)

Table 3 For the LPS model, L^2-norm errors and convergence rates for the DtD strategy and different values of Poisson's ratio

$\nu = 0.3$			$\nu = 0.49$		
δ	e_0	Rate	δ	e_0	Rate
0.3	4.547e–6	–	0.3	3.253e–5	–
0.15	7.698e–7	2.5625	0.15	4.836e–6	2.7498
0.075	1.714e–7	2.1673	0.075	1.002e–6	2.2711
0.0375	4.053e–8	2.0801	0.0375	2.291e–7	2.1291

internal pressure $p_0 = 0.1$, the classical, local displacement solution for the hollow cylinder is given by

$$\mathbf{u}_l = \left[A\boldsymbol{x}_1 + \frac{B\boldsymbol{x}_1}{\boldsymbol{x}_1^2 + \boldsymbol{x}_2^2}, A\boldsymbol{x}_2 + \frac{B\boldsymbol{x}_2}{\boldsymbol{x}_1^2 + \boldsymbol{x}_2^2} \right],$$

where

$$A = \frac{(1+\nu)(1-2\nu)p_0 R_0^2}{K(R_1^2 - R_0^2)}, \quad B = \frac{(1+\nu)p_0 R_0^2 R_1^2}{K(R_1^2 - R_0^2)}.$$

$R_0 = 1$ and $R_1 = 1.5$ are the interior and exterior radius of the (undeformed) hollow cylinder. We report the results of our tests in Table 3 for the DtD strategy and in Table 4 for the DtN strategy. In both cases, we consider two values of Poisson's ratio $\nu = 0.3$ and 0.49, respectively. Also, in this case, the ratio between δ and h is fixed and set to 3.2; the coarser computational domain is such that $h = 0.0937$. The node spacing is then halved at each run of the convergence test. The L^2-norm errors show a *quadratic* convergence rate, confirming our theoretical predictions in Remark 1.

Table 4 For the LPS model, L^2-norm errors and convergence rates for the DtN strategy and different values of Poisson's ratio

$\nu = 0.3$			$\nu = 0.49$		
δ	e_0	Rate	δ	e_0	Rate
0.3	7.651e–6	–	0.3	2.460e–4	–
0.15	2.025e–6	1.9179	0.15	7.133e–5	1.7863
0.075	4.900e–7	2.0470	0.075	1.694e–5	2.0737
0.0375	1.111e–7	2.1412	0.0375	3.824e–6	2.1478

6 Conclusion

In this work, we introduced a technique to automatically convert local boundary conditions into nonlocal volume constraints. A first approximation to the nonlocal solution is provided by the computation of the corresponding local solution, for which local boundary data are available. The local solution is then used to define either Dirichlet or Neumann nonlocal volume constraints. The latter guarantee that the nonlocal problem is well-posed and that its corresponding solution is physically consistent, i.e., it converges quadratically to the local solution as the nonlocality vanishes. Our conversion method does not have any geometry or dimensionality constraints and is inexpensive compared to the computational cost incurred in when solving nonlocal problems. The theoretical quadratic convergence with respect to the horizon δ is illustrated by several two-dimensional numerical experiments conducted by meshfree discretization. The consistency, convergence, and effectiveness of our approach are demonstrated for both scalar nonlocal Poisson's problems and nonlocal mechanics problems (namely, for the linear peridynamic solid model).

This work sets the groundwork for the deployment of nonlocal models at the engineering and industry level where the use of such models is often hindered by the technical difficulties that arise when dealing with the lack of volume constraints necessary for the well-posedness and numerical solution of nonlocal equations.

Acknowledgments M. D'Elia is supported by the Sandia National Laboratories Laboratory Directed Research and Development program. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. Y. Yu would like to acknowledge support by the National Science Foundation under award DMS 1753031. Portions of this research were conducted on Lehigh University's Research Computing infrastructure partially supported by NSF Award 2019035.

This work, SAND2021-7745 R, describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the chapter do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

1. B. Alali, M. Gunzburger, Peridynamics and material interfaces. J. Elasticity **120**, 225–248 (2015)
2. E. Aulisa, G. Capodaglio, A. Chierici, M. D'Elia, Efficient quadrature rules for finite element discretizations of nonlocal equations. Numerical Methods for Partial Differential Equations (2021)
3. D. Benson, S. Wheatcraft, M. Meerschaert, Application of a fractional advection-dispersion equation. Water Resour. Res. **36**, 1403–1412 (2000)
4. F. Bobaru, M. Yang, L.F. Alves, S.A. Silling, E. Askari, J. Xu, Convergence, adaptive refinement, and scaling in 1D peridynamics. Int. J. Numer. Methods Eng. **77**, 852–877 (2009)
5. A.F. Bower, *Applied Mechanics of Solids* (CRC Press, Boca Raton, 2009)
6. A.A. Buades, B. Coll, J. Morel, Image denoising methods. a new nonlocal principle. SIAM Rev. **52**, 113–147 (2010)
7. N. Burch, M. D'Elia, R. Lehoucq, The exit-time problem for a Markov jump process. Eur. Phys. J. Spec. Top. **223**, 3257–3271 (2014)
8. O. Burkovska, M. Gunzburger, On a nonlocal Cahn–Hilliard model permitting sharp interfaces. Mathematical Models and Methods in Applied Sciences **31**(09), 1749–1786 (2021)
9. O. Burkovska, C. Glusa, M. D'Elia, An optimization-based approach to parameter learning for fractional type nonlocal models. Computers & Mathematics with Applications (2021)
10. G. Capodaglio, M. D'Elia, P. Bochev, M. Gunzburger, An energy-based coupling approach to nonlocal interface problems. Comput. Fluids **207**, 104593 (2019)
11. G. Capodaglio, M. D'Elia, M. Gunzburger, P. Bochev, M. Klar, C. Vollmann, A general framework for substructuring-based domain decomposition methods for models having nonlocal interactions. Numerical Methods for Partial Differential Equations (2020)
12. P. Clark Di Leoni, T.A. Zaki, G. Karniadakis, C. Meneveau, Two-point stress–strain-rate correlation structure and non-local eddy viscosity in turbulent flows. J. Fluid Mech. **914**, A6 (2021). https://doi.org/10.1017/jfm.2020.977
13. O. Defterli, M. D'Elia, Q. Du, M. Gunzburger, R. Lehoucq, M.M. Meerschaert, Fractional diffusion on bounded domains. Fract. Calculus Appl. Analy. **18**, 342–360 (2015)
14. A. Delgoshaie, D. Meyer, P. Jenny, H. Tchelepi, Non-local formulation for multiscale flow in porous media. J. Hydrol. **531**, 649–654 (2015)
15. M. D'Elia, M. Gulian, Analysis of anisotropic nonlocal diffusion models: Well-posedness of fractional problems for anomalous transport (2021). Preprint arXiv:2101.04289
16. M. D'Elia, M. Gunzburger, Identification of the diffusion parameter in nonlocal steady diffusion problems. Appl. Math. Optim. **73**, 227–249 (2016)
17. M. D'Elia, P. Bochev, D. Littlewood, M. Perego, Optimization-based coupling of local and nonlocal models: Applications to peridynamics, in *Handbook of Nonlocal Continuum Mechanics for Materials and Structures* (Springer, Cham, 2017)
18. M. D'Elia, Q. Du, M. Gunzburger, R. Lehoucq, Nonlocal convection-diffusion problems on bounded domains and finite-range jump processes. Comput. Methods Appl. Math. **29**, 71–103 (2017)
19. M. D'Elia, Q. Du, C. Glusa, M. Gunzburger, X. Tian, Z. Zhou, Numerical methods for nonlocal and fractional models. Acta Numer. **29**, 1–124 (2020)
20. M. D'Elia, C. Flores, X. Li, P. Radu, Y. Yu, Helmholtz-Hodge decompositions in the nonlocal framework. Well-posedness analysis and applications. J. Peridyn. Nonlocal Model. **2**, 401–418 (2020)
21. M. D'Elia, M. Gunzburger, C. Vollmann, A cookbook for approximating Euclidean balls and for quadrature rules in finite element methods for nonlocal problems. Mathematical Models and Methods in Applied Sciences **31**(08), 1505–1567 (2021)

22. M. D'Elia, X. Tian, Y. Yue, A physically consistent, flexible, and efficient strategy to convert local boundary conditions into nonlocal volume constraints. SIAM J. Sci. Comput. **42**(4), A1935–A1949 (2020)
23. M. D'Elia, M. Gulian, H. Olson, G.E. Karniadakis, Towards a unified theory of fractional and nonlocal vector calculus. Fract. Calculus Appl. Analy. **24**, 5 (2021)
24. M. D'Elia, D. Littlewood, J. Trageser, M. Perego, P. Bochev, An optimization-based strategy for peridynamic-FEM coupling and for the prescription of nonlocal boundary conditions (2021). Preprint arXiv:2110.04420
25. M. D'Elia, J.C.D.L. Reyes, A. Miniguano-Trujillo, Bilevel parameter learning for nonlocal image denoising models. J. Math. Imaging Vis. **63**(6), 753–775 (2021)
26. Q. Du, M. Gunzburger, R. Lehoucq, K. Zhou, Analysis and approximation of nonlocal diffusion problems with volume constraints. SIAM Rev. **54**, 667–696 (2012)
27. Q. Du, M. Gunzburger, R.B. Lehoucq, K. Zhou, A nonlocal vector calculus, nonlocal volume–constrained problems, and nonlocal balance laws. Math. Models Methods Appl. Sci. **23**, 493–540 (2013)
28. E. Emmrich, O. Weckner, et al., On the well-posedness of the linear peridynamic model and its convergence towards the Navier equation of linear elasticity. Commun. Math. Sci. **5**, 851–864 (2007)
29. Y. Fan, X. Tian, X. Yang, X. Li, C. Webster, Y. Yue, An asymptotically compatible probabilistic collocation method for randomly heterogeneous nonlocal problems. J. Comput. Phys. **111376** (2022)
30. P. Fife, *Some Nonclassical Trends in Parabolic and Parabolic-Like Evolutions*. Vehicular Ad Hoc Networks (Springer, New York, 2003), pp. 153–191
31. M. Foss, P. Radu, Y. Yue, Convergence analysis and numerical studies for linearly elastic peridynamics with dirichlet-type boundary conditions. J Peridyn Nonlocal Model, 1–36 (2022)
32. G. Gilboa, S. Osher, Nonlocal linear image regularization and supervised segmentation. Multiscale Model. Simul. **6**, 595–630 (2007)
33. Y.D. Ha, F. Bobaru, Characteristics of dynamic brittle fracture captured with peridynamics. Eng. Fract. Mechan. **78**, 1156–1168 (2011)
34. R.B. Lehoucq, S.A. Silling, Force flux and the peridynamic stress tensor. J. Mechan. Phys. Solids **56**, 1566–1577 (2008)
35. M. Meerschaert, A. Sikorskii, *Stochastic Models for Fractional Calculus*. Studies in Mathematics, Gruyter (2012)
36. T. Mengesha, Q. Du, Analysis of a scalar nonlocal peridynamic model with a sign changing kernel. Discrete Continuous Dyn. Syst.-B **18**, 1415–1437 (2013)
37. T. Mengesha, Q. Du, The bond-based peridynamic system with Dirichlet-type volume constraint. Proc. Roy. Soc. Edinburgh Sect. A **144**, 161–186 (2014)
38. T. Mengesha, Q. Du, Nonlocal constrained value problems for a linear peridynamic Navier equation. J. Elasticity **116**, 27–51 (2014)
39. R. Metzler, J. Klafter, The random walk's guide to anomalous diffusion: a fractional dynamics approach. Phys. Rep. **339**, 1–77 (2000)
40. G. Pang, L. Lu, G.E. Karniadakis, fPINNs: Fractional physics-informed neural networks. SIAM J. Sci. Comput. **41**, A2603–A2626 (2019)
41. G. Pang, M. D'Elia, M. Parks, G.E. Karniadakis, nPINNs: nonlocal Physics-Informed Neural Networks for a parametrized nonlocal universal Laplacian operator. Algorithms and applications. J. Comput. Phys. **422**, 109760 (2020)
42. M. Pasetto, Enhanced Meshfree Methods for Numerical Solution of Local and Nonlocal Theories of Solid Mechanics, PhD thesis, UC San Diego, 2019
43. A. Schekochihin, S. Cowley, T. Yousef, MHD turbulence: Nonlocal, anisotropic, nonuniversal?, in *In IUTAM Symposium on Computational Physics and New Perspectives in Turbulence* (Springer, Dordrecht, 2008), pp. 347–354
44. R. Schumer, D. Benson, M. Meerschaert, S. Wheatcraft, Eulerian derivation of the fractional advection-dispersion equation. J. Contaminant Hydrol. **48**, 69–88 (2001)

45. R. Schumer, D. Benson, M. Meerschaert, B. Baeumer, Multiscaling fractional advection-dispersion equations and their solutions. Water Resour. Res. **39**, 1022–1032 (2003)
46. P. Seleson, M. Gunzburger, M.L. Parks, Interface problems in nonlocal diffusion and sharp transitions between local and nonlocal domains. Comput. Methods Appl. Mechan. Eng. **266**, 185–204 (2013)
47. S. Silling, Reformulation of elasticity theory for discontinuities and long-range forces. J. Mechan. Phys. Solids **48**, 175–209 (2000)
48. S.A. Silling, E. Askari, A meshfree method based on the peridynamic model of solid mechanics. Comput. Struct. **83**, 1526–1535 (2005)
49. X. Tian, Q. Du, Asymptotically compatible schemes and applications to robust discretization of nonlocal models. SIAM J. Num. Analy. **52**, 1641–1665 (2014)
50. N. Trask, H. You, Y. Yu, M.L. Parks, An asymptotically compatible meshfree quadrature rule for nonlocal problems with applications to peridynamics. Comput. Methods Appl. Mechan. Eng. **343**, 151–165 (2019)
51. H. Wang, K. Wang, T. Sircar, A direct $O(N \log^2 N)$ finite difference method for fractional diffusion equations. J. Comput. Phys. **229**, 8095–8104 (2010)
52. X. Xu, M. D'Elia, J.T. Foster, A machine-learning framework for peridynamic material models with physical constraints. Comput. Methods Appl. Mech. Eng. **386**, 114062 (2021)
53. X. Xu, C. Glusa, M. D'Elia, J.T. Foster, A FETI approach to domain decomposition for meshfree discretizations of nonlocal problems. Comput. Methods Appl. Mech. Eng. **387**, 114148 (2021)
54. H. You, X. Lu, N. Trask, Y. Yu, An asymptotically compatible approach for Neumann-type boundary condition on nonlocal problems. ESAIM: Math. Modell. Num. Analy. **54**, 1373–1413 (2020)
55. H. You, Y. Yu, D. Kamensky, An asymptotically compatible formulation for local-to-nonlocal coupling problems without overlapping regions. Comput. Methods Appl. Mechan. Eng. **366**, 113038 (2020)
56. H. You, Y. Yu, S. Silling, M. D'Elia, Data-driven learning of nonlocal models: From high-fidelity simulations to constitutive laws, in *Proceedings of the AAAI Spring Symposium*, (MLPS, 2021)
57. H. You, Y. Yue, N. Trask, M. Gulian, M. D'Elia, Data-driven learning of nonlocal physics from high-fidelity synthetic data. Comput. Methods Appl. Mech. Eng. **374**, 113553 (2021)
58. Y. Yu, F.F. Bargos, H. You, M.L. Parks, M.L. Bittencourt, G.E. Karniadakis, A partitioned coupling framework for peridynamics and classical theory: analysis and simulations. Comput. Methods Appl. Mechan. Eng. **340**, 905–931 (2018)
59. Y. Yu, H. You, N. Trask, An asymptotically compatible treatment of traction loading in linearly elastic peridynamic fracture. Comput. Methods Appl. Mechan. Eng. **377**, 113691 (2021)

Existence of Global Solutions for 2D Fluid–Elastic Interaction with Small Data

Karoline Disser and Michelle Luckas

List of Definitions

$\Omega_{\mathrm{F}} \subset \mathbb{R}^2$	Fluid domain
$\Omega_{\mathrm{S}} \subset \mathbb{R}^2$	Solid domain
$u : (0, \infty) \times \Omega_{\mathrm{F}} \to \mathbb{R}^2$	Fluid velocity
$p : (0, \infty) \times \Omega_{\mathrm{F}} \to \mathbb{R}$	Fluid pressure
$\xi : (0, \infty) \times \Omega_{\mathrm{S}} \to \mathbb{R}^2$	Elastic displacement
$E : (0, \infty) \to \mathbb{R}$	Energy of the coupled system
$K : (0, \infty) \to \mathbb{R}$	Higher order quantity of the coupled system
X_T	Solution space for the fluid velocity up to time T
$\tilde{X}_T$	Higher regularity solution space for the fluid velocity up to time T
Y_T	Solution space for the fluid pressure up to time T
$\tilde{Y}_T$	Higher regularity solution space for the fluid pressure up to time T
Z_T	Solution space for the elastic deformation up to time T
$\tilde{Z}_T$	Higher regularity solution space for the elastic deformation up to time T

K. Disser · M. Luckas (✉)
Institut für Mathematik, Universität Kassel, Kassel, Germany
e-mail: kdisser@mathematik.uni-kassel.de; mluckas@mathematik.uni-kassel.de

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_9

1 Introduction

We study a model for the dynamics of a linearly elastic body immersed in an incompressible viscous fluid, in a two-dimensional setting. There are several difficulties in the analysis of systems of this type, like the mixed parabolic–hyperbolic character of the linearized problem and the fact that it is of free-boundary type. In the mathematical literature, many different techniques have been developed for dealing with these issues. We refer to [3, 5, 6, 9, 13, 17, 18] and the references therein for results on the existence and uniqueness of weak and strong solutions, also for related models as in the case of elastic shells and compressible fluids. It remains an essentially open question to characterize the long-time behaviour of solutions to these systems or prove stability of equilibria. For results in this direction, on related models, we refer to [1, 8, 12]. Recently, for the first time, for fully coupled non-damped fluid–linear elasticity systems, in [4, 18], the existence of unique solutions that conserve regularity over time was shown in a three-dimensional setting. This gives rise to the hope of finding global strong solutions, even if it is for small data. Our main findings include the identification of a weaker functional analytic setting, which only includes norms subject to global a priori estimates.

At the same time, however, our system is simpler in that we interpret the idea of *geometric linearization* to mean that the fluid domain Ω_{F} on which the equations need to be solved remains constant through time, cf. [11]. More precisely, let $\Omega \subset \mathbb{R}^2$ be a bounded $\mathrm{C}^{2,1}$-domain that contains the solid body in the $\mathrm{C}^{2,1}$-domain $\Omega_{\mathrm{S}} \subset \Omega$ and a viscous fluid in the domain Ω_{F} such that $\overline{\Omega_{\mathrm{S}}} \cup \Omega_{\mathrm{F}} = \Omega$. The fluid–solid interface corresponds to the boundary $\partial\Omega_{\mathrm{S}}$. We consider the system

$$\begin{cases} \dot{u} + (u \cdot \nabla)u - \operatorname{div}(\sigma(u, p)) = 0 & \text{in } (0, T) \times \Omega_{\mathrm{F}}, \\ \operatorname{div}(u) = 0 & \text{in } (0, T) \times \Omega_{\mathrm{F}}, \\ \sigma(u, p)n = \Sigma(\xi)n & \text{on } (0, T) \times \partial\Omega_{\mathrm{S}}, \\ u = \dot{\xi} & \text{on } (0, T) \times \partial\Omega_{\mathrm{S}}, \\ u = 0 & \text{on } (0, T) \times \partial\Omega, \\ \ddot{\xi} - \operatorname{div}(\Sigma(\xi)) = 0 & \text{in } (0, T) \times \Omega_{\mathrm{S}}, \\ u(0) = u_0 & \text{in } \Omega_{\mathrm{F}}, \\ \xi(0) = \xi_0 & \text{in } \Omega_{\mathrm{S}}, \\ \dot{\xi}(0) = \xi_1 & \text{in } \Omega_{\mathrm{S}}. \end{cases} \tag{1}$$

The unknowns are the fluid velocity $u\colon (0, T) \times \Omega_{\mathrm{F}} \to \mathbb{R}^2$, fluid pressure $p\colon (0, T) \times \Omega_{\mathrm{F}} \to \mathbb{R}^2$ and elastic displacement $\xi\colon (0, T) \times \Omega_{\mathrm{S}} \to \mathbb{R}^2$. We denote the fluid and solid stress tensors, respectively, by

$$\sigma(u, p) := 2\nu\varepsilon(u) - p\mathrm{Id}, \qquad \Sigma(\xi) := 2\lambda_1\varepsilon(\xi) + \lambda_2 \operatorname{div}(\xi)\mathrm{Id},$$

with the symmetric gradient

$$\varepsilon(v) := \frac{1}{2}\left(\nabla v + (\nabla v)^T\right),$$

and viscosity and Lamé constants $\nu, \lambda_1, \lambda_2 > 0$.

Note that compared to the standard model considered, e.g. in [4], system (1) has no non-linearity in the boundary condition and it is not a free-boundary problem. On the other hand, the linearized system remains the same and the fluid non-linearity is still strong due to the substantial coupling to the elasticity equations. We refer to [14] for the proof of existence of weak solutions to this system.

We prove the existence of a global strong solution in the case that the initial data are sufficiently small. This smallness assumption in the two-dimensional setting may correspond to the fact that linear elasticity is a model for small deformations only. As far as we know, these are the first global results for a model of this kind. It will be the subject of future work to extend them as far as possible.

The chapter is organized as follows: in Sect. 2, we prove local existence and uniqueness of solutions to system (1) and recall important preliminary results. In Sect. 3, we show how, for small data, the solution extends to be global in time. As a corollary, we obtain the convergence

$$\lim_{t\to\infty} \|u(t)\|_{\mathrm{H}^1(\Omega_{\mathrm{F}})} = 0.$$

The Appendix contains the proofs of several auxiliary estimates.

2 Local Existence of Solutions

We first establish the existence of solutions (u, p, ξ) to (1) up to a (possibly small) time $T > 0$ in the spaces

$$\begin{aligned} X_T &:= \mathrm{L}^2(\mathrm{H}^2(\Omega_{\mathrm{F}})) \cap \mathrm{H}^1(\mathrm{H}^1(\Omega_{\mathrm{F}})) \cap \mathrm{C}^1(\mathrm{L}^2(\Omega_{\mathrm{F}})), \\ Y_T &:= \mathrm{L}^2(\mathrm{H}^1(\Omega_{\mathrm{F}})), \\ Z_T &:= \mathrm{C}^0(\mathrm{H}^2(\Omega_{\mathrm{S}})) \cap \mathrm{C}^1(\mathrm{H}^1(\Omega_{\mathrm{S}})) \cap \mathrm{C}^2(\mathrm{L}^2(\Omega_{\mathrm{S}})), \end{aligned}$$

respectively, where we omit the time interval $(0, T)$ for the Sobolev spaces and $[0, T]$ for the spaces of continuous functions whenever possible. We also show that this implies the additional regularity

$$\begin{aligned} u &\in \mathrm{C}^0(\mathrm{H}^2(\Omega_{\mathrm{F}})) \cap X_T, \\ p &\in \mathrm{C}^0(\mathrm{H}^1(\Omega_{\mathrm{F}})). \end{aligned} \tag{2}$$

The equations in (1) are thus all strongly satisfied in $\mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{F}))$ and this corresponds to the requirement that the initial data

$$u_0 \in \mathrm{H}^2(\Omega_\mathrm{F}),\ \xi_0 \in \mathrm{H}^2(\Omega_\mathrm{S}),\ \xi_1 \in \mathrm{H}^1(\Omega_\mathrm{S}) \tag{3}$$

are such that there are $u_1,\ p_0 \in \mathrm{H}^1(\Omega_\mathrm{F})$ and $\xi_2 \in \mathrm{L}^2(\Omega_\mathrm{S})$ that satisfy the compatibility conditions

$$\begin{cases} u_1 + (u_0 \cdot \nabla)u_0 - \operatorname{div}(\sigma(u_0, p_0)) = 0 & \text{in } \Omega_\mathrm{F}, \\ \operatorname{div}(u_0) = 0 & \text{in } \Omega_\mathrm{F}, \\ \operatorname{div}(u_1) = 0 & \text{in } \Omega_\mathrm{F}, \\ \sigma(u_0, p_0)n = \Sigma(\xi_0)n & \text{on } \partial\Omega_\mathrm{S}, \\ u_0 = \xi_1 & \text{on } \partial\Omega_\mathrm{S}, \\ u_0 = 0 & \text{on } \partial\Omega, \\ u_1 = 0 & \text{on } \partial\Omega, \\ \xi_2 - \operatorname{div}(\Sigma(\xi_0)) = 0 & \text{in } \Omega_\mathrm{S}. \end{cases} \tag{4}$$

Our local existence result is the following.

Theorem 1 *Let the initial data u_0, ξ_0 and ξ_1 be given such that (3) and (4) are satisfied. Then there exists a time*

$$T = T\left(\|u_0\|_{\mathrm{H}^1(\Omega_\mathrm{F})}, \|u_1\|_{\mathrm{L}^2(\Omega_\mathrm{F})}, \|\varepsilon(\xi_0)\|_{\mathrm{L}^2(\Omega_\mathrm{S})}, \|\xi_1\|_{\mathrm{H}^1(\Omega_\mathrm{S})}, \|\xi_2\|_{\mathrm{L}^2(\Omega_\mathrm{S})}\right) > 0$$

such that the system (1) admits a unique solution

$$u \in \mathrm{C}^0(\mathrm{H}^2(\Omega_\mathrm{F})) \cap X_T,\ p \in \mathrm{C}^0(\mathrm{H}^1(\Omega_\mathrm{F})) \text{ and } \xi \in Z_T.$$

The proof of Theorem 1 is divided into four steps.

Step 1: Linearization and Preliminary Results
A main ingredient in the proof of Theorem 1 is a recent result of Boulakia, Guerrero and Takahashi on the existence of solutions to the linearized equations,

$$\begin{cases} \dot{u} - \operatorname{div}(\sigma(u,p)) = f & \text{in } (0,T) \times \Omega_{\mathrm{F}}, \\ \operatorname{div}(u) = 0 & \text{in } (0,T) \times \Omega_{\mathrm{F}}, \\ \sigma(u,p)n = \Sigma(\xi)n & \text{on } (0,T) \times \partial\Omega_{\mathrm{S}}, \\ u = \dot{\xi} & \text{on } (0,T) \times \partial\Omega_{\mathrm{S}}, \\ u = 0 & \text{on } (0,T) \times \partial\Omega, \\ \ddot{\xi} - \operatorname{div}(\Sigma(\xi)) = 0 & \text{in } (0,T) \times \Omega_{\mathrm{S}}, \\ u(0) = u_0 & \text{in } \Omega_{\mathrm{F}}, \\ \xi(0) = \xi_0 & \text{in } \Omega_{\mathrm{S}}, \\ \dot{\xi}(0) = \xi_1 & \text{in } \Omega_{\mathrm{S}}, \end{cases} \tag{5}$$

in a more regular setting. This was shown in three spatial dimensions but can be transferred to the two-dimensional situation.

We define the auxiliary spaces

$$\begin{aligned} \tilde{X}_T &:= \mathrm{L}^2(\mathrm{H}^{5/2+1/16}(\Omega_{\mathrm{F}})) \cap \mathrm{H}^1(\mathrm{H}^2(\Omega_{\mathrm{F}})) \cap \mathrm{H}^2(\mathrm{H}^1(\Omega_{\mathrm{F}})), \\ \tilde{Y}_T &:= \mathrm{L}^2(\mathrm{H}^{3/2+1/16}(\Omega_{\mathrm{F}})) \cap \mathrm{H}^1(\mathrm{H}^1(\Omega_{\mathrm{F}})), \\ \tilde{Z}_T &:= \mathrm{L}^2(\mathrm{H}^{5/2+1/16}(\Omega_{\mathrm{S}})) \cap \mathrm{C}^1(\mathrm{H}^{3/2+1/16}(\Omega_{\mathrm{S}})) \\ &\quad \cap \mathrm{C}^2(\mathrm{H}^{1/2+1/16}(\Omega_{\mathrm{S}})) \cap \mathrm{C}^3(\mathrm{H}^{-1/2+1/16}(\Omega_{\mathrm{S}})). \end{aligned}$$

Then, the following holds.

Theorem 2 ([4, Theorem 1.5]) *Let*

$$\begin{aligned} &(u_0,\ u_1,\ p_0,\ \xi_0,\ \xi_1,\ \xi_2,\ f) \\ \in\ &\mathrm{H}^{5/2+1/16}(\Omega_{\mathrm{F}}) \times \mathrm{H}^1(\Omega_{\mathrm{F}}) \times \mathrm{H}^{3/2+1/16}(\Omega_{\mathrm{F}}) \\ &\times \mathrm{H}^{5/2+1/16}(\Omega_{\mathrm{S}}) \times \mathrm{H}^{3/2+1/16}(\Omega_{\mathrm{S}}) \times \mathrm{H}^{1/2+1/16}(\Omega_{\mathrm{S}}) \\ &\times \left(\mathrm{L}^2(\mathrm{H}^{1/2+1/16}(\Omega_{\mathrm{F}})) \cap \mathrm{H}^1(\mathrm{L}^2(\Omega_{\mathrm{F}}))\right) \end{aligned} \tag{6}$$

be such that the compatibility conditions in (4) are satisfied, where the non-linear term $(u_0 \cdot \nabla)u_0$ is replaced by $-f(0)$. Then, for every $T > 0$, the linear system (5) admits a unique solution $(u, p, \xi) \in \tilde{X}_T \times \tilde{Y}_T \times \tilde{Z}_T$.

The proof uses a hidden regularity result for the Lamé system (cf. [15]) and the following optimal elliptic estimates for the stationary Stokes and Lamé systems which we will also need throughout this chapter.

Theorem 3 ([10, Theorem 7.5]) *Let $D \subset \mathbb{R}^2$ be a domain with boundary $\partial D = \Gamma_0 \cup \Gamma_n$ of class $\mathrm{C}^{2,1}$, and let $0 \le s \le 1$. Consider given $f \in \mathrm{H}^s(D)$ and $g \in \mathrm{H}^{1/2+s}(\Gamma_n)$. If the pair $(v, q) \in \mathrm{H}^2(D) \times \mathrm{H}^1(D)$ solves*

$$\begin{cases} -\operatorname{div}(\sigma(v,q)) = f \;\; in\; D, \\ \operatorname{div}(v) = 0 \;\; in\; D, \\ v = 0 \;\; on\; \Gamma_0, \\ \sigma(v,q)n = g \;\; on\; \Gamma_n, \end{cases}$$

then it satisfies the estimate

$$\|v\|_{\mathrm{H}^{2+s}(D)} + \|q\|_{\mathrm{H}^{1+s}(D)} \le C\left(\|f\|_{\mathrm{H}^{s}(D)} + \|g\|_{\mathrm{H}^{1/2+s}(\Gamma_n)}\right).$$

Theorem 4 *Let $D \subset \mathbb{R}^2$ be a domain with boundary ∂D of class $\mathrm{C}^{2,1}$, and let $0 \le s \le 2$. Consider given $f \in \mathrm{H}^s(D)$ and $g \in \mathrm{H}^{1/2+s}(\partial D)$. If $\eta \in \mathrm{H}^1(D)$ solves*

$$\begin{cases} -\operatorname{div}(\Sigma(\eta)) = f \;\; in\; D, \\ \eta = g \;\; on\; \partial D, \end{cases}$$

then it satisfies the estimate

$$\|\eta\|_{\mathrm{H}^{1+s}(D)} \le C\left(\|f\|_{\mathrm{H}^{-1+s}(D)} + \|g\|_{\mathrm{H}^{1/2+s}(\partial D)}\right).$$

Step 2: A Priori Estimates and Approximation

A key step in our argument is to reduce the regularity needed in Theorem 2 to norms with global a priori estimates. First, we do this in the linearized setting, using an approximation argument.

Lemma 1 *Let*

$$\begin{aligned} &(u_0,\, u_1,\, p_0,\, \xi_0,\, \xi_1,\, \xi_2,\, f) \\ \in\; &\mathrm{H}^2(\Omega_\mathrm{F}) \times \mathrm{H}^1(\Omega_\mathrm{F}) \times \mathrm{H}^1(\Omega_\mathrm{F}) \times \mathrm{H}^2(\Omega_\mathrm{S}) \times \mathrm{H}^1(\Omega_\mathrm{S}) \times \mathrm{L}^2(\Omega_\mathrm{S}) \\ &\times \left(\mathrm{L}^2(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{F}))\right) \end{aligned} \tag{7}$$

be such that the compatibility conditions (4) are satisfied (with $(u_0 \cdot \nabla)u_0$ replaced by $-f(0)$). Then, the linear system (5) admits a unique solution $(u, p, \xi) \in X_T \times Y_T \times Z_T$ satisfying

$$\begin{aligned} &\|u(t)\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} + \|\dot{\xi}(t)\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})} + \int_{\Omega_{\mathrm{S}}} \Sigma(\xi) : \varepsilon(\xi)(t)\,\mathrm{d}y \\ &+ \int_0^t 4\nu \|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})}\,\mathrm{d}s \\ = \ &\|u_0\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} + \|\xi_1\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})} + \int_{\Omega_{\mathrm{S}}} \Sigma(\xi_0) : \varepsilon(\xi_0)\,\mathrm{d}y \\ &+2\int_0^t \int_{\Omega_{\mathrm{F}}} f \cdot u\,\mathrm{d}y\mathrm{d}s \end{aligned} \tag{8}$$

and

$$\begin{aligned} &\|\dot{u}(t)\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} + \|\ddot{\xi}(t)\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})} + \int_{\Omega_{\mathrm{S}}} \Sigma(\dot{\xi}) : \varepsilon(\dot{\xi})(t)\,\mathrm{d}y \\ &+ \int_0^t 4\nu \|\varepsilon(\dot{u}(s))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})}\,\mathrm{d}s \\ = \ &\|u_1\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} + \|\xi_2\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})} + \int_{\Omega_{\mathrm{S}}} \Sigma(\xi_1) : \varepsilon(\xi_1)\,\mathrm{d}y \\ &+2\int_0^t \langle \dot{u}(s), \dot{f}(s)\rangle_{1/2-1/16}\,\mathrm{d}s. \end{aligned} \tag{9}$$

Proof We start by considering the solutions $(u, p, \xi) \in \tilde{X}_T \times \tilde{Y}_T \times \tilde{Z}_T$ from Theorem 2 in the case of regular initial data and right-hand side f satisfying (6) and (4). Then, we use an approximation argument to come back to less regular data.

First, we multiply the differential equations for the fluid and the elastic part with u and $\dot{\xi}$, respectively, and use integration by parts to obtain the energy equality (8). To obtain higher order estimates, we use that

$$(\dot{u}, \dot{p}) \in \big(\mathrm{L}^2(\mathrm{H}^2(\Omega_{\mathrm{F}})) \cap \mathrm{H}^1(\mathrm{L}^2(\Omega_{\mathrm{F}}))\big) \times \mathrm{L}^2(\mathrm{H}^1(\Omega_{\mathrm{F}}))$$

solves the linear system

$$\begin{cases} \dot{U} - \mathrm{div}(\sigma(U, P)) = \dot{f} & \text{in } (0, T) \times \Omega_{\mathrm{F}}, \\ \mathrm{div}(U) = 0 & \text{in } (0, T) \times \Omega_{\mathrm{F}}, \\ \sigma(U, P)n = \Sigma(\dot{\xi})n & \text{on } (0, T) \times \partial\Omega_{\mathrm{S}}, \\ U = 0 & \text{on } (0, T) \times \partial\Omega, \\ U(0) = u_1 & \text{in } \Omega_{\mathrm{F}}. \end{cases}$$

Testing with $\dot{u}$ yields

$$\begin{aligned} &\frac{1}{2}\|\dot{u}(t)\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \int_0^t 2\nu\|\varepsilon(\dot{u}(s))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})}\,\mathrm{d}s \\ &-\int_0^t\int_{\partial\Omega_\mathrm{S}} \sigma(\dot{u},\dot{p})(s)n\cdot\dot{u}(s)\,\mathrm{d}S(y)\mathrm{d}s \\ =\; &\frac{1}{2}\|u_1\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \int_0^t\int_{\Omega_\mathrm{F}} \dot{f}(s)\cdot\dot{u}(s)\,\mathrm{d}y\,\mathrm{d}s. \end{aligned} \tag{10}$$

For the elastic part, we note that

$$\dot{\xi} \in \mathrm{C}^0(\mathrm{H}^{3/2+1/16}(\Omega_\mathrm{S})) \cap \mathrm{C}^1(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{S})) \cap \mathrm{C}^2(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{S}))$$

satisfies

$$\operatorname{div}(\Sigma(\dot{\xi})) = \dddot{\xi} \quad \text{in } \mathrm{C}^0(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{F}))$$

and

$$\Sigma(\dot{\xi})n = \sigma(\dot{u},\dot{p})n \quad \text{in } \mathrm{L}^2(\mathrm{L}^2(\partial\Omega_\mathrm{S})).$$

Hence we obtain

$$\begin{aligned} &\frac{1}{2}\|\ddot{\xi}(t)\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})} + \frac{1}{2}\int_{\Omega_\mathrm{S}} \Sigma(\dot{\xi}) : \varepsilon(\dot{\xi})(t)\,\mathrm{d}y \\ &+\int_0^t\int_{\partial\Omega_\mathrm{S}} \sigma(\dot{u},\dot{p})(s)n\cdot\ddot{\xi}(s)\,\mathrm{d}S(y)\mathrm{d}s \\ =\; &\frac{1}{2}\|\xi_2\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})} + \frac{1}{2}\int_{\Omega_\mathrm{S}} \Sigma(\xi_1) : \varepsilon(\xi_1)\,\mathrm{d}y. \end{aligned} \tag{11}$$

By using Korn's second inequality, we obtain that

$$\|u(t)\|_{\mathrm{H}^1(\Omega_\mathrm{F})} \le C\left(\|u(t)\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \|\varepsilon(u(t))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})}\right)^{1/2} \tag{12}$$

and

$$\|\xi(t)\|_{\mathrm{H}^1(\Omega_\mathrm{S})} \le C\left(\|\xi(t)\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})} + \|\varepsilon(\xi(t))\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})}\right)^{1/2}.$$

Moreover, we observe that

$$\begin{aligned}\int_{\Omega_{\mathrm{S}}} \Sigma(\xi(t)) : \varepsilon(\xi(t))\,\mathrm{d}y &= \textstyle\int_{\Omega_{\mathrm{S}}} (2\lambda_1 \varepsilon(\xi(t)) + \lambda_2 \operatorname{div}(\xi(t))\mathrm{Id}) : \varepsilon(\xi(t))\,\mathrm{d}y \\ &= 2\lambda_1 \|\varepsilon(\xi(t))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})} + \textstyle\int_{\Omega_{\mathrm{S}}} \lambda_2 \operatorname{div}(\xi(t))^2\,\mathrm{d}y.\end{aligned}$$

Therefore,

$$c\|\varepsilon(\xi(t))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})} \leq \int_{\Omega_{\mathrm{S}}} \Sigma(\xi(t)) : \varepsilon(\xi(t))\,\mathrm{d}y \leq C\|\varepsilon(\xi(t))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})}, \tag{13}$$

and

$$c\|\dot{\xi}(t)\|^2_{\mathrm{H}^1(\Omega_{\mathrm{S}})} \leq \|\dot{\xi}(t)\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})} + \int_{\Omega_{\mathrm{S}}} \Sigma(\dot{\xi}(t)) : \varepsilon(\dot{\xi}(t))\,\mathrm{d}y \leq C\|\dot{\xi}(t)\|^2_{\mathrm{H}^1(\Omega_{\mathrm{S}})}. \tag{14}$$

By combining (12), (13), (14) and (8) and using Young's inequality, it follows that

$$\begin{aligned} &\|u\|^2_{\mathrm{H}^1(\mathrm{H}^1(\Omega_{\mathrm{F}}))\cap \mathrm{C}^1(\mathrm{L}^2(\Omega_{\mathrm{F}}))} + \|\varepsilon(\xi)\|^2_{\mathrm{C}^0(\mathrm{L}^2(\Omega_{\mathrm{S}}))} + \|\dot{\xi}\|^2_{\mathrm{C}^0(\mathrm{H}^1(\Omega_{\mathrm{S}}))\cap \mathrm{C}^1(\mathrm{L}^2(\Omega_{\mathrm{F}}))} \\ \leq\ & C\Big(\|u_0\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} + \|u_1\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} + \|\varepsilon(\xi_0)\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})} + \|\xi_1\|^2_{\mathrm{H}^1(\Omega_{\mathrm{S}})} \\ &\quad + \|\xi_2\|^2_{\mathrm{L}^2(\Omega_{\mathrm{S}})} + \|f\|^2_{\mathrm{L}^2(\mathrm{L}^2(\Omega_{\mathrm{F}}))\cap \mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_{\mathrm{F}}))}\Big). \end{aligned} \tag{15}$$

Now, let us consider less regular initial data, i.e. let $(u_0, u_1, p_0, \xi_0, \xi_1, \xi_2, f)$ satisfy (7) and (4). By Lemma 7, there exists a sequence

$$\begin{aligned} &\left(u_0^n, u_1^n, p_0^n, \xi_0^n, \xi_1^n, \xi_2^n, f^n\right) \\ \in\ & \mathrm{H}^{5/2+1/16}(\Omega_{\mathrm{F}}) \times \mathrm{H}^1(\Omega_{\mathrm{F}}) \times \mathrm{H}^{3/2+1/16}(\Omega_{\mathrm{F}}) \times \mathrm{H}^{5/2+1/16}(\Omega_{\mathrm{S}}) \\ & \times \mathrm{H}^{3/2+1/16}(\Omega_{\mathrm{S}}) \times \mathrm{H}^{1/2+1/16}(\Omega_{\mathrm{S}}) \times \mathrm{C}^\infty(\mathrm{H}^{1/2+1/16}(\Omega_{\mathrm{F}})) \end{aligned}$$

satisfying the compatibility conditions (4) for all $n \in \mathbb{N}$ and converging to

$$(u_0, u_1, p_0, \xi_0, \xi_1, \xi_2, f)$$

in the norm of

$$\begin{aligned} &\mathrm{H}^2(\Omega_{\mathrm{F}}) \times \mathrm{H}^1(\Omega_{\mathrm{F}}) \times \mathrm{H}^1(\Omega_{\mathrm{F}}) \times \mathrm{H}^2(\Omega_{\mathrm{S}}) \times \mathrm{H}^1(\Omega_{\mathrm{S}}) \times \mathrm{L}^2(\Omega_{\mathrm{S}}) \\ &\qquad \times \left(\mathrm{L}^2(\mathrm{H}^{1/2+1/16}(\Omega_{\mathrm{F}})) \cap \mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_{\mathrm{F}}))\right). \end{aligned}$$

By Theorem 2, we find solutions $(u^n, p^n, \xi^n) \in \tilde{X}_T \times \tilde{Y}_T \times \tilde{Z}_T$ to the linear system (5) and data $(u_0^n, u_1^n, p_0^n, \xi_0^n, \xi_1^n, \xi_2^n, f^n)$. By repeating the calculations for (15) for the difference $(u^n - u^m, p^n - p^m, \xi^n - \xi^m)$ for any $n, m \in \mathbb{N}$, we obtain the estimate

$$\begin{aligned}
&\|u^n - u^m\|^2_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))} + \|\varepsilon(\xi^n - \xi^m)\|^2_{\mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{S}))}\\
&\quad + \|\dot\xi^n - \dot\xi^m\|^2_{\mathrm{C}^0(\mathrm{H}^1(\Omega_\mathrm{S}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))}\\
\le{}& C\Big(\|u_0^n - u_0^m\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \|u_1^n - u_1^m\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \|\varepsilon(\xi_0^n - \xi_0^m)\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})}\\
&\quad + \|\xi_1^n - \xi_1^m\|^2_{\mathrm{H}^1(\Omega_\mathrm{S})} + \|\xi_2^n - \xi_2^m\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})}\\
&\quad + \|f^n - f^m\|^2_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))\cap\mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{F}))}\Big).
\end{aligned} \tag{16}$$

Hence, it follows from the convergence of the data $\left(u_0^n,\ u_1^n,\ p_0^n,\ \xi_0^n,\ \xi_1^n,\ \xi_2^n,\ f^n\right)$ that (u_n, ξ_n) is a Cauchy sequence in

$$\Big(\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))\Big) \times \Big(\mathrm{C}^1(\mathrm{H}^1(\Omega_\mathrm{S}))\cap\mathrm{C}^2(\mathrm{L}^2(\Omega_\mathrm{S}))\Big).$$

Moreover, by Theorem 3 with $s = 0$ and a Trace theorem,

$$\begin{aligned}
&\|u^n - u^m\|_{\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{F}))} + \|p^n - p^m\|_{\mathrm{L}^2(\mathrm{H}^1(\Omega_\mathrm{F}))}\\
\le{}& C\Big(\|\dot u^n - \dot u^m\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))} + \|f^n - f^m\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))}\\
&\quad + \|\Sigma(\xi^n - \xi^m)\|_{\mathrm{L}^2(\mathrm{H}^{1/2}(\partial\Omega_\mathrm{S}))}\Big)\\
\le{}& C\Big(T^{1/2}\|\dot u^n - \dot u^m\|_{\mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{F}))} + \|f^n - f^m\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))}\\
&\quad + T^{1/2}\|\xi^n - \xi^m\|_{\mathrm{C}^0(\mathrm{H}^2(\Omega_\mathrm{S}))}\Big).
\end{aligned}$$

Similarly, Theorem 4 with $s = 1$ implies

$$\begin{aligned}
&\|\xi^n - \xi^m\|_{\mathrm{C}^0(\mathrm{H}^2(\Omega_\mathrm{S}))}\\
\le{}& C\left(\|\ddot\xi^n - \ddot\xi^m\|_{\mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{S}))} + \|\xi_0^n - \xi_0^m\|_{\mathrm{H}^2(\Omega_\mathrm{S})} + T^{1/2}\|u^n - u^m\|_{\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{F}))}\right).
\end{aligned}$$

Combining the last two estimates yields

$$\begin{aligned}
&\|u^n - u^m\|_{\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{F}))} + \|p^n - p^m\|_{\mathrm{L}^2(\mathrm{H}^1(\Omega_\mathrm{F}))} + \|\xi^n - \xi^m\|_{\mathrm{C}^0(\mathrm{H}^2(\Omega_\mathrm{S}))}\\
\le{}& C\Big(T^{1/2}\|\dot u^n - \dot u^m\|_{\mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{F}))} + \|f^n - f^m\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))}\\
&\quad + (1 + T^{1/2})\|\ddot\xi_n - \ddot\xi_m\|_{\mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{S}))} + (1 + T^{1/2})\|\xi_0^n - \xi_0^m\|_{\mathrm{H}^2(\Omega_\mathrm{S})}\\
&\quad + (T^{1/2} + T)\|u^n - u^m\|_{\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{F}))}\Big).
\end{aligned}$$

Consequently, if we choose $T > 0$ small enough such that $C(T^{1/2} + T) < 1$ and absorb the last term on the right-hand side, we obtain that (u^n, p^n, ξ^n) is also a Cauchy sequence in

$$\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{F})) \times \mathrm{L}^2(\mathrm{H}^1(\Omega_\mathrm{F})) \times \mathrm{C}^0(\mathrm{H}^2(\Omega_\mathrm{S}))$$

and therefore in $X_T \times Y_T \times Z_T$. Hence, (u^n, p^n, ξ^n) converges to some $(u, p, \xi) \in X_T \times Y_T \times Z_T$, which is by construction a strong solution to the linear system (5) for the less regular data $(u_0, u_1, p_0, \xi_0, \xi_1, \xi_2, f)$. The uniqueness of solutions to this system follows from the energy equality (8). This concludes the proof of Lemma 1. □

Step 3: Non-Linear Problem

Now, we are in the position to prove the existence and uniqueness of solutions in Theorem 1 based on a fixed-point argument. We give the proof in some detail as the choice of norms is special. Consider given data

$$\begin{aligned}&(u_0, u_1, p_0, \xi_0, \xi_1, \xi_2)\\ &\in \mathrm{H}^2(\Omega_\mathrm{F}) \times \mathrm{H}^1(\Omega_\mathrm{F}) \times \mathrm{H}^1(\Omega_\mathrm{F}) \times \mathrm{H}^2(\Omega_\mathrm{S}) \times \mathrm{H}^1(\Omega_\mathrm{S}) \times \mathrm{L}^2(\Omega_\mathrm{S})\end{aligned}$$

such that the compatibility conditions (4) are satisfied. For some

$$M = M\left(\|u_0\|_{\mathrm{H}^1(\Omega_\mathrm{F})}, \|u_1\|_{\mathrm{L}^2(\Omega_\mathrm{F})}, \|\varepsilon(\xi_0)\|_{\mathrm{L}^2(\Omega_\mathrm{S})}, \|\xi_1\|_{\mathrm{H}^1(\Omega_\mathrm{S})}, \|\xi_2\|_{\mathrm{L}^2(\Omega_\mathrm{S})}\right) > 0,$$

we set

$$X_T^{0,M} := \left\{v \in X_T : v(0) = u_0,\ \dot{v}(0) = u_1,\ \|v\|^2_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))} \le M\right\}.$$

Note that it follows from Lemma 1 and the estimate (15) for

$$f := (u_0 \cdot \nabla)u_0 \in \mathrm{L}^2(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{F}))$$

that $X_T^{0,M} \neq \emptyset$ if we choose $M > 0$ large enough. For given $\tilde{u} \in X_T^{0,M}$, Lemma 5 (a) tells us that

$$\tilde{f} := (\tilde{u} \cdot \nabla)\tilde{u} \in \mathrm{L}^2(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{F})),$$

so we can apply Lemma 1 to obtain a solution (u, p, ξ) of the linear system (5) with $f = \tilde{f}$. We want to show that the map $S : \tilde{u} \to u$ is a contraction from $X_T^{0,M}$ to $X_T^{0,M}$ and thus admits a unique fixed point. Lemma 1 shows that $u \in X_T$ attains the correct initial values. To obtain

$$\|u\|^2_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))} \le M$$

for $T > 0$ sufficiently small, first we observe that again by Lemma 1,

$$\begin{aligned}&\|u\|^2_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))} + \|\varepsilon(\xi)\|^2_{\mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{S}))} + \|\dot{\xi}\|^2_{\mathrm{C}^0(\mathrm{H}^1(\Omega_\mathrm{S}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{S}))}\\ \leq\ & C\Big(\|u_0\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \|u_1\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \|\varepsilon(\xi_0)\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})} + \|\xi_1\|^2_{\mathrm{H}^1(\Omega_\mathrm{S})} + \|\xi_2\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})}\\ &\quad + \textstyle\int_0^T\int_{\Omega_\mathrm{F}}(\tilde{u}\cdot\nabla)\tilde{u}\cdot u\,\mathrm{d}y\mathrm{d}s + \int_0^T\int_{\Omega_\mathrm{F}}\Big((\dot{\tilde{u}}\cdot\nabla)\tilde{u} + (\tilde{u}\cdot\nabla)\dot{\tilde{u}}\Big)\cdot\dot{u}\,\mathrm{d}y\mathrm{d}s\Big).\end{aligned}$$

Now, using Lemma 6, we estimate

$$\begin{aligned}&\int_0^T\int_{\Omega_\mathrm{F}}(\tilde{u}\cdot\nabla)\tilde{u}\cdot u\,\mathrm{d}y\mathrm{d}s + \int_0^T\int_{\Omega_\mathrm{F}}\Big((\dot{\tilde{u}}\cdot\nabla)\tilde{u} + (\tilde{u}\cdot\nabla)\dot{\tilde{u}}\Big)\cdot\dot{u}\,\mathrm{d}y\mathrm{d}s\\ \leq\ & CT^\alpha\Big(\|\tilde{u}\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} + \|u_0\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\Big)\\ &\qquad\times\|\tilde{u}\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))}\|u\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))},\end{aligned}$$

and hence by Young's inequality,

$$\begin{aligned}&\|u\|^2_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))} + \|\varepsilon(\xi)\|^2_{\mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{S}))} + \|\dot{\xi}\|^2_{\mathrm{C}^0(\mathrm{H}^1(\Omega_\mathrm{S}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{S}))}\\ \leq\ & C\Big(\|u_0\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \|u_1\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \|\varepsilon(\xi_0)\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})} + \|\xi_1\|^2_{\mathrm{H}^1(\Omega_\mathrm{S})}\\ &\quad + \|\xi_2\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})} + T^\alpha\Big(M^2 + \|u_0\|^2_{\mathrm{H}^1(\Omega_\mathrm{F})}\Big)M^2\Big).\end{aligned}\tag{17}$$

Thus, for

$$M = M\Big(\|u_0\|_{\mathrm{L}^2(\Omega_\mathrm{F})}, \|u_1\|_{\mathrm{L}^2(\Omega_\mathrm{F})}, \|\varepsilon(\xi_0)\|_{\mathrm{L}^2(\Omega_\mathrm{S})}, \|\xi_1\|_{\mathrm{H}^1(\Omega_\mathrm{S})}, \|\xi_2\|_{\mathrm{L}^2(\Omega_\mathrm{S})}\Big) > 0$$

sufficiently large and $T = T(M, \|u_0\|_{\mathrm{H}^1(\Omega_\mathrm{F})}) > 0$ sufficiently small, we obtain

$$\|u\|^2_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))} \leq M$$

and consequently $u \in X_T^{0,M}$, such that $S : X_T^{0,M} \to X_T^{0,M}$ is well-defined. Now, in order to prove that S is also a contraction, let $(u^1, p^1, \xi^1),\ (u^2, p^2, \xi^2) \in X_T \times Y_T \times Z_T$ denote the solutions of (5) corresponding to some $\tilde{u}^1,\ \tilde{u}^2 \in X_T^{0,M}$, respectively. By repeating the calculations from (17) for the difference $u^1 - u^2$ and using that $\tilde{u}^1(0) - \tilde{u}^2(0) = 0$, we obtain that

$$\begin{aligned}
&\|u^1 - u^2\|^2_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))} + \|\varepsilon(\xi^1 - \xi^2)\|^2_{\mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{S}))}\\
&+\|\dot{\xi}^1 - \dot{\xi}^2\|^2_{\mathrm{C}^0(\mathrm{H}^1(\Omega_\mathrm{S}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{S}))} \qquad (18)\\
\le\ & CT^\alpha\left(M^2 + \|u_0\|^2_{\mathrm{H}^1(\Omega_\mathrm{F})}\right)\|\tilde{u}^1 - \tilde{u}^2\|^2_{X_T}.
\end{aligned}$$

Furthermore, using Theorem 3 with $s = 0$, Theorem 4 with $s = 1$ and Lemma 5 (b) yields

$$\begin{aligned}
&\|u^1 - u^2\|_{\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{F}))}\\
\le\ & C\Big(\|\dot{u}^1 - \dot{u}^2\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))}\\
&+\|\tilde{u}^1\cdot\nabla(\tilde{u}^1 - \tilde{u}^2)\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))} + \|(\tilde{u}^1 - \tilde{u}^2)\cdot\nabla\tilde{u}^2\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))}\\
&+\|\Sigma(\xi^1 - \xi^2)\|_{\mathrm{L}^2(\mathrm{H}^{1/2}(\partial\Omega_\mathrm{S}))}\Big)\\
\le\ & C\Big(T^{1/2}\|u^1 - u^2\|_{\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))} \qquad (19)\\
&+T^\alpha\left(\|\tilde{u}^1\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} + \|\tilde{u}^2\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} + 2\|u_0\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right)\|\tilde{u}^1 - \tilde{u}^2\|_{X_T}\\
&+\|\xi^1 - \xi^2\|_{\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{S}))}\Big)\\
\le\ & C\Big(T^{1/2}\|u^1 - u^2\|_{\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))}\\
&+T^\alpha(M + \|u_0\|_{\mathrm{H}^1(\Omega_\mathrm{F})})\|\tilde{u}^1 - \tilde{u}^2\|_{X_T}\\
&+T^{1/2}\|\dot{\xi}^1 - \dot{\xi}^2\|_{\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{S}))} + T\|u^1 - u^2\|_{\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{F}))}\Big).
\end{aligned}$$

Combining (18) and (19) for $T = T\left(M,\ \|u_0\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right) > 0$ sufficiently small, we obtain that

$$\|u^1 - u^2\|_{X_T} \le \frac{1}{2}\|\tilde{u}^1 - \tilde{u}^2\|_{X_T},$$

and hence $S : X_T^{0,M} \to X_T^{0,M}$ is a contraction. Therefore, S admits a unique fixed point $u \in X_T$, which together with the corresponding $p \in Y_T$ and $\xi \in Z_T$ forms the unique solution to the non-linear system (1).

Step 4: Additional Regularity
Finally, we remark that $(u, p, \xi) \in X_T \times Y_T \times Z_T$ also implies that

$$u \in \mathrm{C}^0(\mathrm{H}^2(\Omega_\mathrm{F})) \text{ and } p \in \mathrm{C}^0(\mathrm{H}^1(\Omega_\mathrm{F})). \qquad (20)$$

By Lemma 5,

$$(u \cdot \nabla)u \in \mathrm{L}^2(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{F})),$$

so it follows from [16, Theorem 3.1] that $(u \cdot \nabla)u \in \mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{F}))$. Hence, (2) follows from Theorem 3 with right-hand side $f = -\dot{u} - (u \cdot \nabla)u \in \mathrm{C}^0(\mathrm{L}^2(\Omega_\mathrm{F}))$ and Neumann boundary data $\Sigma(\xi)n \in \mathrm{C}^0(\mathrm{H}^{3/2}(\partial\Omega_\mathrm{S}))$. This concludes the proof of Theorem 1. □

3 Existence of Global Solutions for Small Data

We want to use the structure of the problem to show that in a setting that guarantees small deformation, the unique solution exists globally. To this end, we define the energy

$$E(t) := \|u(t)\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \|\dot{\xi}(t)\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})} + \int_{\Omega_\mathrm{S}} \Sigma(\xi) : \varepsilon(\xi)(t)\,\mathrm{d}y,$$

associated with system (1), and a corresponding higher order quantity

$$K(t) := \|\dot{u}(t)\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} + \|\ddot{\xi}(t)\|^2_{\mathrm{L}^2(\Omega_\mathrm{S})} + \int_{\Omega_\mathrm{S}} \Sigma(\dot{\xi}) : \varepsilon(\dot{\xi})(t)\,\mathrm{d}y.$$

We show that the lifespan $T > 0$ of the local solution given in Theorem 1 can be controlled by E and K and that E and K can be bounded also in the non-linear setting, if the initial data are sufficiently small. This implies the following result on global existence.

Theorem 5 *There exist constants $C_E > 0$ and $C_K > 0$ such that for initial data*

$$\begin{aligned} d &:= (u_0, u_1, p_0, \xi_0, \xi_1, \xi_2) \\ &\in \mathrm{H}^2(\Omega_\mathrm{F}) \times \mathrm{H}^1(\Omega_\mathrm{F}) \times \mathrm{H}^1(\Omega_\mathrm{F}) \times \mathrm{H}^2(\Omega_\mathrm{S}) \times \mathrm{H}^1(\Omega_\mathrm{S}) \times \mathrm{L}^2(\Omega_\mathrm{S}) \end{aligned}$$

satisfying the compatibility conditions (4) and the bounds

$$E(0) \le C_E, \qquad K(0) \le C_K, \tag{21}$$

the corresponding unique solution (u, p, ξ) to (1) exists up to any time $T > 0$.

Proof From Theorem 1, it follows that (1) admits a unique solution

$$
\begin{aligned}
(u, p, \xi) \in & \Big(\mathrm{C}^0(0, T; \mathrm{H}^2(\Omega_\mathrm{F})) \cap \mathrm{H}^1(0, T; \mathrm{H}^1(\Omega_\mathrm{F})) \cap \mathrm{C}^1(0, T; \mathrm{L}^2(\Omega_\mathrm{F}))\Big) \\
& \times \mathrm{C}^0(0, T; \mathrm{H}^1(\Omega_\mathrm{F})) \qquad (22) \\
& \times \Big(\mathrm{C}^0(0, T; \mathrm{H}^2(\Omega_\mathrm{S})) \cap \mathrm{C}^1(0, T; \mathrm{H}^1(\Omega_\mathrm{S})) \cap \mathrm{C}^2(0, T; \mathrm{L}^2(\Omega_\mathrm{S}))\Big)
\end{aligned}
$$

up to some time

$$
T\left(\|u_0\|_{\mathrm{H}^1(\Omega_\mathrm{F})}, \|u_1\|_{\mathrm{L}^2(\Omega_\mathrm{F})}, \|\varepsilon(\xi_0)\|_{\mathrm{L}^2(\Omega_\mathrm{S})}, \|\xi_1\|_{\mathrm{H}^1(\Omega_\mathrm{S})}, \|\xi_2\|_{\mathrm{L}^2(\Omega_\mathrm{S})}\right) > 0.
$$

First, we show that if condition (21) is satisfied for suitable $C_E, C_K > 0$, then

$$
E(t) \leq E(0) \text{ and } K(t) \leq K(0) \text{ for all } t \in [0, T).
$$

Note that by Korn's and Poincaré's inequalities, there exist constants $c_1, c_2 > 0$ such that

$$
c_1 \|\varepsilon(v)\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} \leq \|v\|^2_{\mathrm{H}^1(\Omega_\mathrm{F})} \leq c_2 \|\varepsilon(v)\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} \qquad (23)
$$

holds for all $v \in \mathrm{H}^1(\Omega_\mathrm{F})$ with partially vanishing trace at the boundary $\partial\Omega_\mathrm{F}$. In particular, this is true for $u(t)$ and $\dot{u}(t)$.

As in Lemma 1, we obtain the energy equality

$$
E(t) + \int_0^t 4\nu \|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} \,\mathrm{d}s = E(0) - 2\int_0^t \int_{\Omega_\mathrm{F}} (u \cdot \nabla)u \cdot u \,\mathrm{d}y\mathrm{d}s. \qquad (24)
$$

For the second term on the right-hand side of (24), we use Hölder's inequality, the embedding $\mathrm{H}^{1/2}(\Omega_\mathrm{F}) \hookrightarrow \mathrm{L}^4(\Omega_\mathrm{F})$ and interpolation to estimate

$$
\begin{aligned}
2\textstyle\int_{\Omega_\mathrm{F}} [(u \cdot \nabla)u \cdot u](s) \,\mathrm{d}y &\leq C\|u(s)\|^2_{\mathrm{L}^4(\Omega_\mathrm{F})} \|\nabla u(s)\|_{\mathrm{L}^2(\Omega_\mathrm{F})} \\
&\leq C\|u(s)\|^2_{\mathrm{H}^{1/2}(\Omega_\mathrm{F})} \|u(s)\|_{\mathrm{H}^1(\Omega_\mathrm{F})} \qquad (25) \\
&\leq C\|u(s)\|_{\mathrm{L}^2(\Omega_\mathrm{F})} \|u(s)\|^2_{\mathrm{H}^1(\Omega_\mathrm{F})} \\
&\leq \hat{C} E(s)^{1/2} \|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})}
\end{aligned}
$$

for some fixed $\hat{C} > 0$ that depends only on Ω_F. This leads to

$$
E(t) + \int_0^t (4\nu - \hat{C} E(s)^{1/2}) \|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} \,\mathrm{d}s \leq E(0). \qquad (26)
$$

If we choose the initial data small enough such that

$$4\nu - \hat{C}E(0)^{1/2} > 0, \tag{27}$$

then we can show that it follows that

$$E(t) \leq E(0) \quad \text{for all } t \in [0, T) : \tag{28}$$

Assume to the contrary that there is a time $t_0 \in (0, T)$ such that $E(t_0) > E(0)$. Because of (27), there is some $\tilde{E} > E(0)$, which still satisfies

$$4\nu - \hat{C}\tilde{E}^{1/2} > 0.$$

As E is continuous in time, we find a time $t_1 \in (0, T)$ such that $E(t_1) > E(0)$ and

$$E(t) \leq \tilde{E} \quad \text{for all } t \leq t_1.$$

But then

$$E(t_1) + \underbrace{\int_0^{t_1} \left(4\nu - \hat{C}\tilde{E})^{1/2}\right) \|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} \,\mathrm{d}s}_{\geq 0} \leq E(0)$$

and hence $E(t_1) \leq E(0)$, which is a contradiction. Moreover, it follows from (26) and (28) that

$$\int_0^t \|\varepsilon(u(s))\|_{\mathrm{L}^2(\Omega_{\mathrm{F}})} \,\mathrm{d}s \leq \frac{E(0)}{4\nu - \hat{C}E(0)^{1/2}}. \tag{29}$$

Next, we want to derive a similar result for the higher order quantity K. By Lemma 1,

$$\begin{aligned} K(t) + \int_0^t 4\nu \|\varepsilon(\dot{u}(s))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} \,\mathrm{d}s \\ = K(0) - 2\int_0^t \int_{\Omega_{\mathrm{F}}} ((\dot{u} \cdot \nabla)u + (u \cdot \nabla)\dot{u}) \cdot \dot{u} \,\mathrm{d}y\mathrm{d}s. \end{aligned} \tag{30}$$

First, we estimate

$$\begin{aligned} 2\textstyle\int_{\Omega_{\mathrm{F}}} [(\dot{u} \cdot \nabla)u \cdot \dot{u}](s) \,\mathrm{d}y &\leq C\|\dot{u}(s)\|^2_{\mathrm{L}^4(\Omega_{\mathrm{F}})} \|\nabla u(s)\|_{\mathrm{L}^2(\Omega_{\mathrm{F}})} \\ &\leq C\|\varepsilon(u(s))\|_{\mathrm{L}^2(\Omega_{\mathrm{F}})} \|\varepsilon(\dot{u}(s))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})}, \end{aligned} \tag{31}$$

and similarly,

$$2\int_{\Omega_F}[(u\cdot\nabla)\dot{u}\cdot\dot{u}](s)\,\mathrm{d}y \le C\|\varepsilon(u(s))\|_{\mathrm{L}^2(\Omega_F)}\|\varepsilon(\dot{u}(s))\|^2_{\mathrm{L}^2(\Omega_F)}. \tag{32}$$

Moreover, using the differentiated energy equality, (25) and (28), we have

$$\begin{aligned}4\nu\|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_F)} &= -\dot{E}(s) - 2\int_{\Omega_F}[(u\cdot\nabla)u\cdot u](s)\,\mathrm{d}y\\ &\le C\left(E(s)+K(s)\right)+\hat{C}E(s)^{1/2}\|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_F)}\\ &\le C\left(E(0)+K(s)\right)+\hat{C}E(0)^{1/2}\|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_F)}\end{aligned}$$

for all $s\in[0,T)$. Because of (27), it follows that

$$\|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_F)} \le C\left(E(0)+K(s)\right). \tag{33}$$

Combining (31), (32) and (33) leads to

$$K(t)+\int_0^t\left(4\nu-\tilde{C}\left(E(0)+K(s)\right)^{1/2}\right)\|\varepsilon(\dot{u}(s))\|^2_{\mathrm{L}^2(\Omega_F)}\,\mathrm{d}s \le K(0) \tag{34}$$

for some fixed $\tilde{C}>0$. If we choose the initial data small enough such that

$$4\nu-\tilde{C}\left(E(0)+K(0)\right)^{1/2}>0, \tag{35}$$

then it follows by a similar argumentation as before that

$$K(t)\le K(0)\quad \text{for all } t\in[0,T). \tag{36}$$

Moreover, by using (34) and (36), we obtain that

$$\int_0^t\|\varepsilon(\dot{u})(s)\|_{\mathrm{L}^2(\Omega_F)}\,\mathrm{d}s \le \frac{K(0)}{4\nu-\tilde{C}\left(E(0)+K(0)\right)^{1/2}}. \tag{37}$$

Hence, we have obtained bounds for $E(t)$ and $K(t)$ for all $t\in[0,T)$. From (33), it follows that also $\|u(t)\|_{\mathrm{H}^1(\Omega_F)}$ is bounded by

$$\|u(t)\|_{\mathrm{H}^1(\Omega_F)} \le C\|\varepsilon(u(t))\|_{\mathrm{L}^2(\Omega_F)} \le C\left(E(0)+K(0)\right)^{1/2}$$

for all $t\in[0,T)$. This yields a bound

$$\begin{aligned}\max\left\{\|u(t)\|_{\mathrm{H}^1(\Omega_F)},\|\dot{u}(t)\|_{\mathrm{L}^2(\Omega_F)},\|\varepsilon(\xi(t))\|_{\mathrm{L}^2(\Omega_S)},\|\dot{\xi}(t)\|_{\mathrm{H}^1(\Omega_S)},\|\ddot{\xi}(t)\|_{\mathrm{L}^2(\Omega_S)}\right\}\\ \le M\end{aligned}$$

for all $t \in [0, T)$. Since the lifespan T of the local solution given in Theorem 1 depends decreasingly on the corresponding norms of the initial data, we find a time T_0 up to which the local solution exists if the initial data satisfies

$$\begin{aligned}\max &\left\{\|u_0\|_{\mathrm{H}^1(\Omega_\mathrm{F})}, \|u_1\|_{\mathrm{L}^2(\Omega_\mathrm{F})}, \|\varepsilon(\xi_0)\|_{\mathrm{L}^2(\Omega_\mathrm{F})}, \|\xi_1\|_{\mathrm{H}^1(\Omega_\mathrm{F})}, \|\xi_2\|_{\mathrm{L}^2(\Omega_\mathrm{F})}\right\} \\ &\leq M. \end{aligned} \tag{38}$$

Since $u \in \mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))$, we can choose $t_0 \in (T_0/2, T_0)$ such that $\dot{u}(t_0) \in \mathrm{H}^1(\Omega_\mathrm{F})$ and set

$$\tilde{d} := \left(\tilde{u}_0, \tilde{u}_1, \tilde{p}_0, \tilde{\xi}_0, \tilde{\xi}_1, \tilde{\xi}_2\right) := \left(u(t_0), \dot{u}(t_0), p(t_0), \xi(t_0), \dot{\xi}(t_0), \ddot{\xi}(t_0)\right).$$

Then, $\tilde{d}$ satisfies the conditions (3) and (4) of Theorem 1 and also (38) such that we obtain a local solution $(\tilde{u}, \tilde{p}, \tilde{\xi})$ on $[0, T_0)$. Due to uniqueness, it follows that $(u(t), p(t), \xi(t))$ and $(\tilde{u}(t - t_0), \tilde{p}(t - t_0), \tilde{\xi}(t - t_0))$ coincide on $[t_0, T_0)$, and hence we can extend $(u(t), p(t), \xi(t))$ up to $[0, T_0 + t_0)$. Since $E(t_0) \leq E(0)$ and $K(t_0) \leq K(0)$, we can repeat the calculations from above for $(\tilde{u}, \tilde{p}, \tilde{\xi})$ to extend the bounds (28), (29), (36) and (37) up to $[0, T_0 + t_0)$. Next, we can choose a suitable $t_1 \in (T_0/2 + t_0, T_0 + t_0)$ and repeat this procedure arbitrarily often to obtain a global solution which satisfies (28), (29), (36) and (37) for all $t > 0$. This proves the existence of a global solution, provided that $E(0)$ and $K(0)$ are sufficiently small such that (27) and (35) are fulfilled. □

Corollary 1 *Consider initial data $(u_0, u_1, p_0, \xi_0, \xi_1, \xi_2)$ satisfying the conditions of Theorem 5. Then, the corresponding global solution (u, p, ξ) to (1) satisfies*

$$\lim_{t\to\infty} \|u(t)\|_{\mathrm{H}^1(\Omega_\mathrm{F})} = 0.$$

Proof As the global solution (u, p, ξ) satisfies (29) and (37) for all $t > 0$, it follows that

$$\int_0^\infty \|\varepsilon(u(t))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})}\, \mathrm{d}t < \infty, \quad \int_0^\infty \|\varepsilon(\dot{u}(t))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})}\, \mathrm{d}t < \infty. \tag{39}$$

Now, let $\delta > 0$. Then, (39) implies that we can find a time $T_\delta > 0$ such that

$$\int_{T_\delta}^\infty \|\varepsilon(u(t))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})}\, \mathrm{d}t < \frac{\delta}{2}, \quad \int_{T_\delta}^\infty \|\varepsilon(\dot{u}(t))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})}\, \mathrm{d}t < \frac{\delta}{2}. \tag{40}$$

Moreover, this shows that there exists another time $t_\delta^1 \geq T_\delta$ such that

$$\|\varepsilon(u(t_\delta^1))\|^2_{\mathrm{L}^2(\Omega_\mathrm{F})} \leq \delta. \tag{41}$$

Now, for $t \geq t_\delta^1$, (23), the fundamental theorem of calculus, Young's inequality and (40) and (41) imply that

$$\begin{aligned}\|u(t)\|^2_{\mathrm{H}^1(\Omega_{\mathrm{F}})} &\leq c_2\|\varepsilon(u(t))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})}\\ &= c_2\|\varepsilon(u(t_\delta^1))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} + c_2\int_{t_\delta^1}^t\int_{\Omega_{\mathrm{F}}} 2\varepsilon(u(s)) : \varepsilon(\dot{u}(s))\,\mathrm{d}y\mathrm{d}s\\ &\leq c_2\|\varepsilon(u(t_\delta^1))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})} + c_2\int_{t_\delta^1}^t \|\varepsilon(u(s))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})}\,\mathrm{d}s\\ &\quad +c_2\int_{t_\delta^1}^t \|\varepsilon(\dot{u}(s))\|^2_{\mathrm{L}^2(\Omega_{\mathrm{F}})}\,\mathrm{d}s\\ &\leq 2c_2\delta.\end{aligned}$$

As $\delta > 0$ was arbitrary, it follows that

$$\lim_{t\to\infty}\|u(t)\|_{\mathrm{H}^1(\Omega_{\mathrm{F}})} = 0. \qquad \square$$

Acknowledgments Michelle Luckas would like to thank the “Studienstiftung des deutschen Volkes” for academic and financial support.

Appendix

The Appendix contains the proof of several auxiliary estimates and an approximation argument.

Definition of Spaces and Auxiliary Estimates

Given a Banach space X, $T > 0$ and $0 < s < 1$, for $f \in \mathrm{L}^2(0, T; X)$, we define

$$[f]_{s,(0,T),X} := \left(\int_0^T\int_0^T \frac{\|f(t_1,\cdot) - f(t_2,\cdot)\|_X^2}{|t_1 - t_2|^{2s+1}}\,\mathrm{d}t_1\mathrm{d}t_2\right)^{1/2}.$$

We denote by $\mathrm{H}^s(0, T; X)$ the Sobolev–Slobodeckii spaces with norms

$$\|f\|_{\mathrm{H}^s(0,T;X)} := \begin{cases}\left(\|f\|^2_{\mathrm{L}^2(0,T;X)} + [f]^2_{s,(0,T),X}\right)^{1/2} & \text{if } 0 < s < 1,\\ \left(\|f\|^2_{\mathrm{H}^1(0,T;X)} + [\dot{f}]^2_{s,(0,T),X}\right)^{1/2} & \text{if } 1 < s < 2.\end{cases}$$

Lemma 2 ([4, Corollary A.3]) *Let $\frac{1}{2} < \sigma \leq 1$ and $0 < s < \sigma$. Then, there exists a constant $C > 0$ independent of T such that*

$$\|f\|_{\mathrm{H}^s(0,T;X)} \leq CT^{\sigma-s}\|f\|_{\mathrm{H}^\sigma(0,T;X)}$$

holds for all $f \in \mathrm{H}^\sigma(0, T; X)$ with $f(0, \cdot) = 0$.

For general $f \in \mathrm{H}^\sigma(0, T; X)$, the preceding lemma implies that

$$\begin{aligned}\|f\|_{\mathrm{H}^s(0,T;X)} &\leq \|f - f(0,\cdot)\|_{\mathrm{H}^s(0,T;X)} + \|f(0,\cdot)\|_{\mathrm{H}^s(0,T;X)} \\ &\leq CT^{\sigma-s}\|f - f(0,\cdot)\|_{\mathrm{H}^\sigma(0,T;X)} + T^{1/2}\|f(0,\cdot)\|_X \qquad \text{(A.1)} \\ &\leq CT^{\sigma-s}\|f\|_{\mathrm{H}^\sigma(0,T;X)} + \left(CT^{1/2+\sigma-s} + T^{1/2}\right)\|f(0,\cdot)\|_X.\end{aligned}$$

Lemma 3 ([4, Lemma A.5])

(a) Let $0 \leq s \leq 1$, $\sigma_1, \sigma_2 \geq 0$, and set $\sigma := s\sigma_1 + (1-s)\sigma_2$. Then,

$$\mathrm{H}^1(\mathrm{H}^{\sigma_1}(\Omega_\mathrm{F})) \cap \mathrm{L}^2(\mathrm{H}^{\sigma_2}(\Omega_\mathrm{F})) \hookrightarrow \mathrm{H}^s(\mathrm{H}^\sigma(\Omega_\mathrm{F})),$$

and there exists a constant $C > 0$ independent of T such that

$$\|v\|_{\mathrm{H}^s(\mathrm{H}^\sigma(\Omega_\mathrm{F}))} \leq C\|v\|^s_{\mathrm{H}^1(\mathrm{H}^{\sigma_1}(\Omega_\mathrm{F}))}\|v\|^{1-s}_{\mathrm{L}^2(\mathrm{H}^{\sigma_2}(\Omega_\mathrm{F}))}$$

for all $v \in \mathrm{H}^1(\mathrm{H}^{\sigma_1}(\Omega_\mathrm{F})) \cap \mathrm{L}^2(\mathrm{H}^{\sigma_2}(\Omega_\mathrm{F}))$.

(b) Let $1 \leq s \leq 2$, $\sigma_1, \sigma_2 \geq 0$, and set $\sigma := (s-1)\sigma_1 + (2-s)\sigma_2$. Then,

$$\mathrm{H}^2(\mathrm{H}^{\sigma_1}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{\sigma_2}(\Omega_\mathrm{F})) \hookrightarrow \mathrm{H}^s(\mathrm{H}^\sigma(\Omega_\mathrm{F})),$$

and there exists a constant $C > 0$ independent of T such that

$$\|v\|_{\mathrm{H}^s(\mathrm{H}^\sigma(\Omega_\mathrm{F}))} \leq C\|v\|^s_{\mathrm{H}^2(\mathrm{H}^{\sigma_1}(\Omega_\mathrm{F}))}\|v\|^{1-s}_{\mathrm{H}^1(\mathrm{H}^{\sigma_2}(\Omega_\mathrm{F}))}$$

for all $v \in \mathrm{H}^2(\mathrm{H}^{\sigma_1}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{\sigma_2}(\Omega_\mathrm{F}))$.

We recall some Sobolev embeddings on the interval $(0, T)$ to clarify the dependence of the appearing constants on the interval length $T > 0$.

Lemma 4

(a) Let $s \in (0, 1/2)$, and set $q := \frac{2}{1-2s}$. Then, $\mathrm{H}^s(0, T) \hookrightarrow \mathrm{L}^q(0, T)$, and there exists a constant $C > 0$ independent of T such that

$$\|f\|_{\mathrm{L}^q(0,T)} \leq C\left(T^{-s}\|f\|_{\mathrm{L}^2(0,T)} + \|f\|_{\mathrm{H}^s(0,T)}\right)$$

holds for all $f \in \mathrm{H}^s(0, T)$.

(b) Let $s \in (1/2, 1)$. Then, $\mathrm{H}^s(0, T) \hookrightarrow \mathrm{C}^0(0, T)$, and there exists a constant $C > 0$ independent of T such that

$$\|f\|_{\mathrm{C}^0(0,T)} \leq C\left(T^{-1/2}\|f\|_{\mathrm{L}^2(0,T)} + T^{s-1/2}\|f\|_{\mathrm{H}^s(0,T)}\right)$$

holds for all $f \in \mathrm{H}^s(0, T)$.

Proof After rescaling a given function $f \in \mathrm{H}^s(0, T)$ to

$$\tilde{f}(\tau) := T^{1/2} f(T\tau), \quad \tau \in (0, 1),$$

the estimates in (a) and (b) can be shown to follow from the corresponding embeddings on the interval (0, 1), cf. [7, Theorem 5.4], [7, Theorem 6.7] and [7, Theorem 8.2]. □

In the sequel, we will often use an estimate obtained by combining (A.1) and Lemma 4 (a). To avoid repetition and to shorten the following proofs, we will now once explain this procedure in detail.

Let $s \in (0, 1/2)$, $\sigma \in (1/2, 1)$ and $f \in \mathrm{H}^\sigma(0, T; X)$ for some Banach space X. Then, for $q := \frac{2}{1-2s}$, Lemma 4 (a) implies that

$$\| \|f\|_X \|_{\mathrm{L}^q(0,T)} \leq C\left(T^{-s}\, \| \|f\|_X \|_{\mathrm{L}^2(0,T)} + \| \|f\|_X \|_{\mathrm{H}^s(0,T)}\right).$$

Now, we apply (A.1) to both terms on the right-hand side to obtain

$$T^{-s}\, \| \|f\|_X \|_{\mathrm{L}^2(0,T)} \leq C\left(T^{\sigma-s}\|f\|_{\mathrm{H}^\sigma(0,T;X)} + (T^{1/2+\sigma-s} + T^{1/2-s})\|f(0)\|_X\right)$$

and

$$\| \|f\|_X \|_{\mathrm{H}^s(0,T)} \leq C\left(T^{\sigma-s}\|f\|_{\mathrm{H}^\sigma(0,T;X)} + (T^{1/2+\sigma-s} + T^{1/2})\|f(0)\|_X\right).$$

Note that every appearing exponent of T is positive. Consequently, we can choose $\alpha > 0$ such that

$$\| \|f\|_X \|_{\mathrm{L}^q(0,T)} \leq CT^\alpha \left(\|f\|_{\mathrm{H}^\sigma(0,T;X)} + \|f(0)\|_X\right). \tag{A.2}$$

Estimates on $(u \cdot \nabla)u$

We provide estimates on the non-linear term $u \cdot \nabla u$ in some detail as the choice of norms for our arguments is special and the time dependence of embedding constants is non-trivial.

Lemma 5 *Let* $u,\ v \in X_T$.

(a) Then,

$$(u \cdot \nabla)u \in \mathrm{L}^2(\mathrm{H}^{3/4}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{-1/4}(\Omega_\mathrm{F})).$$

(b) There exist some $C,\ \alpha > 0$ *such that*

$$\begin{aligned}
&\|(u \cdot \nabla)v\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))}\\
\leq\ & CT^\alpha \min\Big\{\Big(\|u\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} + \|u(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\Big)\\
&\qquad \times \Big(\|v\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} + \|v(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\Big)^{1/2} \|v\|_{X_T}^{1/2},\\
&\qquad \|u\|_{X_T}^{1/2}\Big(\|u\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} + \|u(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\Big)^{1/2}\\
&\qquad \times \Big(\|v\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} + \|v(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\Big)\Big\}.
\end{aligned}$$

Proof

(a) To show $(u \cdot \nabla)u \in \mathrm{L}^2(\mathrm{H}^{3/4}(\Omega_\mathrm{F}))$, first we use interpolation to estimate

$$\begin{aligned}
&\|(u \cdot \nabla)u\|_{\mathrm{H}^{3/4}(\Omega_\mathrm{F})}\\
\leq\ & C\|(u \cdot \nabla)u\|_{\mathrm{H}^1(\Omega_\mathrm{F})}^{3/4}\|(u \cdot \nabla)u\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{1/4}\\
\leq\ & C\|(u \cdot \nabla)u\|_{\mathrm{L}^2(\Omega_\mathrm{F})} + C\||\nabla u||\nabla u|\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{3/4}\|(u \cdot \nabla)u\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{1/4}\\
&+C\||u||\nabla^2 u|\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{3/4}\|(u \cdot \nabla)u\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{1/4}.
\end{aligned}$$

Now, Hölder's inequality together with the embeddings $\mathrm{H}^{1/2}(\Omega_\mathrm{F}) \hookrightarrow \mathrm{L}^4(\Omega_\mathrm{F})$ and $\mathrm{H}^{9/8}(\Omega_\mathrm{F}) \hookrightarrow \mathrm{L}^\infty(\Omega_\mathrm{F})$ and interpolation yields

$$\begin{aligned}
&\||\nabla u||\nabla u|\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{3/4}\|(u \cdot \nabla)u\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{1/4} + C\||u||\nabla^2 u|\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{3/4}\|(u \cdot \nabla)u\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{1/4}\\
\leq\ & C\|\nabla u\|_{\mathrm{L}^4(\Omega_\mathrm{F})}^{3/2}\|u\|_{\mathrm{L}^4(\Omega_\mathrm{F})}^{1/4}\|\nabla u\|_{\mathrm{L}^4(\Omega_\mathrm{F})}^{1/4}\\
&+C\|u\|_{\mathrm{L}^\infty(\Omega_\mathrm{F})}^{3/4}\|\nabla^2 u\|_{\mathrm{L}^2(\Omega_\mathrm{F})}^{3/4}\|u\|_{\mathrm{L}^4(\Omega_\mathrm{F})}^{1/4}\|\nabla u\|_{\mathrm{L}^4(\Omega_\mathrm{F})}^{1/4}
\end{aligned}$$

$$\begin{aligned}
&\leq C\|u\|^{7/4}_{\mathrm{H}^{3/2}(\Omega_\mathrm{F})}\|u\|^{1/4}_{\mathrm{H}^{1/2}(\Omega_\mathrm{F})}\\
&\quad +C\|u\|^{3/4}_{\mathrm{H}^{9/8}(\Omega_\mathrm{F})}\|u\|^{3/4}_{\mathrm{H}^{2}(\Omega_\mathrm{F})}\|u\|^{1/4}_{\mathrm{H}^{1/2}(\Omega_\mathrm{F})}\|u\|^{1/4}_{\mathrm{H}^{3/2}(\Omega_\mathrm{F})}\\
&\leq C\|u\|^{1/8}_{\mathrm{L}^{2}(\Omega_\mathrm{F})}\|u\|_{\mathrm{H}^{1}(\Omega_\mathrm{F})}\|u\|^{7/8}_{\mathrm{H}^{2}(\Omega_\mathrm{F})} + C\|u\|^{1/8}_{\mathrm{L}^{2}(\Omega_\mathrm{F})}\|u\|^{29/32}_{\mathrm{H}^{1}(\Omega_\mathrm{F})}\|u\|^{31/32}_{\mathrm{H}^{2}(\Omega_\mathrm{F})}\\
&\leq C\|u\|_{\mathrm{H}^{1}(\Omega_\mathrm{F})}\|u\|_{\mathrm{H}^{2}(\Omega_\mathrm{F})}.
\end{aligned}$$

Similarly, we can estimate

$$\|(u\cdot\nabla)u\|_{\mathrm{L}^{2}(\Omega_\mathrm{F})} \leq C\|u\|_{\mathrm{H}^{1}(\Omega_\mathrm{F})}\|u\|_{\mathrm{H}^{2}(\Omega_\mathrm{F})}.$$

From these estimates follows by applying Hölder's inequality on $(0,T)$ and using the embedding $\mathrm{H}^1(0,T)\hookrightarrow \mathrm{L}^\infty(0,T)$ that

$$\begin{aligned}
\|(u\cdot\nabla)u\|_{\mathrm{L}^{2}(\mathrm{H}^{3/4}(\Omega_\mathrm{F}))} &\leq C\left\|\|u\|_{\mathrm{H}^{1}(\Omega_\mathrm{F})}\|u\|_{\mathrm{H}^{2}(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^2(0,T)}\\
&\leq C\left\|\|u\|_{\mathrm{H}^{1}(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^\infty(0,T)}\left\|\|u\|_{\mathrm{H}^{2}(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^2(0,T)}\\
&\leq C(T)\|u\|^2_{\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{F}))\cap\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}\\
&\leq C(T)\|u\|^2_{X_T}
\end{aligned}$$

and hence $(u\cdot\nabla)u\in\mathrm{L}^2(\mathrm{H}^{3/4}(\Omega_\mathrm{F}))$. To show $(u\cdot\nabla)u\in\mathrm{H}^1(\mathrm{H}^{-1/4}(\Omega_\mathrm{F}))$, note that

$$\mathrm{H}^{-1/4}(\Omega_\mathrm{F}) = (\mathrm{H}^{1/4}(\Omega_\mathrm{F}))^*$$

by [19, Theorem 4.8.2], [2, Theorem 1.1]. Now, for $v\in\mathrm{H}^{1/4}(\Omega_\mathrm{F})$, we use Hölder's inequality together with the embeddings $\mathrm{H}^{1/2}(\Omega_\mathrm{F})\hookrightarrow\mathrm{L}^4(\Omega_\mathrm{F})$, $\mathrm{H}^{1/4}(\Omega_\mathrm{F})\hookrightarrow\mathrm{L}^{8/3}(\Omega_\mathrm{F})$ and $\mathrm{H}^{3/4}(\Omega_\mathrm{F})\hookrightarrow\mathrm{L}^{8}(\Omega_\mathrm{F})$ to estimate

$$\begin{aligned}
&\int_{\Omega_\mathrm{F}}((\dot{u}\cdot\nabla)u+(u\cdot\nabla)\dot{u})\cdot v\,\mathrm{d}y\\
&\leq C\left(\|\dot{u}\|_{\mathrm{L}^{4}(\Omega_\mathrm{F})}\|\nabla u\|_{\mathrm{L}^{8/3}(\Omega_\mathrm{F})}+\|u\|_{\mathrm{L}^{8}(\Omega_\mathrm{F})}\|\nabla\dot{u}\|_{\mathrm{L}^{2}(\Omega_\mathrm{F})}\right)\|v\|_{\mathrm{L}^{8/3}(\Omega_\mathrm{F})}\\
&\leq C\left(\|\dot{u}\|_{\mathrm{H}^{1/2}(\Omega_\mathrm{F})}\|\nabla u\|_{\mathrm{H}^{1/4}(\Omega_\mathrm{F})}+\|u\|_{\mathrm{H}^{3/4}(\Omega_\mathrm{F})}\|\nabla\dot{u}\|_{\mathrm{L}^{2}(\Omega_\mathrm{F})}\right)\|v\|_{\mathrm{H}^{1/4}(\Omega_\mathrm{F})}\\
&\leq C\left(\|\dot{u}\|_{\mathrm{H}^{1/2}(\Omega_\mathrm{F})}\|u\|_{\mathrm{H}^{5/4}(\Omega_\mathrm{F})}+\|u\|_{\mathrm{H}^{3/4}(\Omega_\mathrm{F})}\|\dot{u}\|_{\mathrm{H}^{1}(\Omega_\mathrm{F})}\right)\|v\|_{\mathrm{H}^{1/4}(\Omega_\mathrm{F})}.
\end{aligned}$$

By applying Hölder's inequality on (0,T) and the embeddings $\mathrm{H}^{1/4}(0,T)\hookrightarrow\mathrm{L}^4(0,T)$ and $\mathrm{H}^{3/4}(0,T)\hookrightarrow\mathrm{L}^\infty(0,T)$ together with Lemma 3, we obtain

$$\begin{aligned}
&\Big\| \int_{\Omega_\mathrm{F}} ((\dot{u} \cdot \nabla)u + (u \cdot \nabla)\dot{u}) \cdot v \,\mathrm{d}y \Big\|_{\mathrm{L}^2(0,T)} \\
&\leq C\Big(\Big\| \|\dot{u}\|_{\mathrm{H}^{1/2}(\Omega_\mathrm{F})} \Big\|_{\mathrm{L}^4(0,T)} \Big\| \|u\|_{\mathrm{H}^{5/4}(\Omega_\mathrm{F})} \Big\|_{\mathrm{L}^4(0,T)} \\
&\qquad + \Big\| \|u\|_{\mathrm{H}^{3/4}(\Omega_\mathrm{F})} \Big\|_{\mathrm{L}^\infty(0,T)} \Big\| \|\dot{u}\|_{\mathrm{H}^1(\Omega_\mathrm{F})} \Big\|_{\mathrm{L}^2(0,T)} \Big) \|v\|_{\mathrm{H}^{1/4}(\Omega_\mathrm{F})} \\
&\leq C(T)\Big(\|u\|_{\mathrm{H}^{5/4}(\mathrm{H}^{1/2}(\Omega_\mathrm{F}))} \|u\|_{\mathrm{H}^{1/4}(\mathrm{H}^{5/4}(\Omega_\mathrm{F}))} \\
&\qquad + \|u\|_{\mathrm{H}^{3/4}(\mathrm{H}^{3/4}(\Omega_\mathrm{F}))} \|u\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} \Big) \|v\|_{\mathrm{H}^{1/4}(\Omega_\mathrm{F})} \\
&\leq C(T)\|u\|^2_{\mathrm{L}^2(\mathrm{H}^2(\Omega_\mathrm{F}))\cap \mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} \|v\|_{\mathrm{H}^{1/4}(\Omega_\mathrm{F})} \\
&\leq C(T)\|u\|^2_{X_T} \|v\|_{\mathrm{H}^{1/4}(\Omega_\mathrm{F})}.
\end{aligned}$$

Similarly, we can also estimate

$$\Big\| \int_{\Omega_\mathrm{F}} (u \cdot \nabla)u \cdot v \,\mathrm{d}y \Big\|_{\mathrm{L}^2(0,T)} \leq C(T)\|u\|^2_{X_T} \|v\|_{\mathrm{H}^{1/4}(\Omega_\mathrm{F})},$$

so we conclude that $(u \cdot \nabla)u \in \mathrm{H}^1(\mathrm{H}^{-1/4}(\Omega_\mathrm{F}))$.

(b) We use again Hölder's inequality on Ω_F, the embedding $\mathrm{H}^{3/2}(\Omega_\mathrm{F}) \hookrightarrow \mathrm{L}^\infty(\Omega_\mathrm{F})$ and interpolation to estimate

$$\begin{aligned}
\|(u \cdot \nabla)v\|_{\mathrm{L}^2(\Omega_\mathrm{F})} &\leq C\|u\|_{\mathrm{L}^\infty(\Omega_\mathrm{F})} \|\nabla v\|_{\mathrm{L}^2(\Omega_\mathrm{F})} \\
&\leq C\|u\|_{\mathrm{H}^{3/2}(\Omega_\mathrm{F})} \|v\|_{\mathrm{H}^1(\Omega_\mathrm{F})} \\
&\leq C\|u\|^{1/2}_{\mathrm{H}^2(\Omega_\mathrm{F})} \|u\|^{1/2}_{\mathrm{H}^1(\Omega_\mathrm{F})} \|v\|_{\mathrm{H}^1(\Omega_\mathrm{F})}.
\end{aligned}$$

Now, Hölder's inequality on $(0, T)$ together with (A.2) for $q = 6$, $s = 1/3$ and $\sigma = 1$ implies

$$\begin{aligned}
&\|(u \cdot \nabla)v\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))} \\
&\leq C \Big\| \|u\|^{1/2}_{\mathrm{H}^2(\Omega_\mathrm{F})} \Big\|_{\mathrm{L}^4(0,T)} \Big\| \|u\|^{1/2}_{\mathrm{H}^1(\Omega_\mathrm{F})} \| \Big\|_{\mathrm{L}^{12}(0,T)} \Big\| \|v\|_{\mathrm{H}^1(\Omega_\mathrm{F})} \Big\|_{\mathrm{L}^6(0,T)} \\
&\leq C \Big\| \|u\|_{\mathrm{H}^2(\Omega_\mathrm{F})} \Big\|^{1/2}_{\mathrm{L}^2(0,T)} \Big\| \|u\|_{\mathrm{H}^1(\Omega_\mathrm{F})} \| \Big\|^{1/2}_{\mathrm{L}^6(0,T)} \Big\| \|v\|_{\mathrm{H}^1(\Omega_\mathrm{F})} \Big\|_{\mathrm{L}^6(0,T)} \\
&\leq CT^\alpha \|u\|^{1/2}_{X_T} \Big(\|u\|_{\mathrm{H}^1(\mathrm{H}^1((\Omega_\mathrm{F}))} + \|u(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})} \| \Big)^{1/2} \\
&\qquad\qquad \times \Big(\|v\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} + \|v(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})} \Big).
\end{aligned}$$

Similarly, we obtain together with $\mathrm{H}^{1/2}(\Omega_\mathrm{F}) \hookrightarrow \mathrm{L}^4(\Omega_\mathrm{F})$ that

$$\begin{aligned}\|(u\cdot\nabla)v\|_{\mathrm{L}^2(\Omega_\mathrm{F})} &\le C\|u\|_{\mathrm{L}^4(\Omega_\mathrm{F})}\|\nabla v\|_{\mathrm{L}^4(\Omega_\mathrm{F})}\\ &\le C\|u\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\|v\|^{1/2}_{\mathrm{H}^1(\Omega_\mathrm{F})}\|v\|^{1/2}_{\mathrm{H}^2(\Omega_\mathrm{F})}\end{aligned}$$

and hence

$$\begin{aligned}&\|(u\cdot\nabla)v\|_{\mathrm{L}^2(\mathrm{L}^2(\Omega_\mathrm{F}))}\\ &\le CT^\alpha\left(\|u\|_{\mathrm{H}^1(\mathrm{H}^1((\Omega_\mathrm{F}))} + \|u(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right)\\ &\qquad\times\left(\|v\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))} + \|v(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right)^{1/2}\|v\|^{1/2}_{X_T}.\end{aligned}$$

□

Lemma 6 *Let* $u,\ v,\ w \in X_T$. *Then, there exists some* $\alpha > 0$ *such that*

$$\begin{aligned}&\int_0^t\int_{\Omega_\mathrm{F}} |(u\cdot\nabla)v\cdot w|\,\mathrm{d}y\mathrm{d}s + \int_0^t\int_{\Omega_\mathrm{F}} |(\dot u\cdot\nabla)v\cdot\dot w| + |(u\cdot\nabla)\dot v\cdot\dot w|\,\mathrm{d}y\mathrm{d}s\\ &\le CT^\alpha\Big(\|u\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))}\|v\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}\\ &\qquad+\|u(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\|v\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}\\ &\qquad+\|v(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\|u\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))}\Big)\|w\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))\cap\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))}.\end{aligned}$$

Proof For the second term, we use Hölder's inequality, the embedding $\mathrm{H}^{1/2}(\Omega_\mathrm{F}) \hookrightarrow \mathrm{L}^4(\Omega_\mathrm{F})$ and interpolation to estimate

$$\begin{aligned}&\int_0^t\int_{\Omega_\mathrm{F}} |(\dot u\cdot\nabla)v\cdot\dot w|\,\mathrm{d}y\mathrm{d}s\\ &\le C\int_0^t \|\dot u\|_{\mathrm{L}^4(\Omega_\mathrm{F})}\|\nabla v\|_{\mathrm{L}^2(\Omega_\mathrm{F})}\|\dot w\|_{\mathrm{L}^4(\Omega_\mathrm{F})}\,\mathrm{d}s\\ &\le C\int_0^t \|\dot u\|_{\mathrm{H}^{1/2}(\Omega_\mathrm{F})}\|\nabla v\|_{\mathrm{L}^2(\Omega_\mathrm{F})}\|\dot w\|_{\mathrm{H}^{1/2}(\Omega_\mathrm{F})}\,\mathrm{d}s\\ &\le C\int_0^t \|\dot u\|^{1/2}_{\mathrm{L}^2(\Omega_\mathrm{F})}\|\dot u\|^{1/2}_{\mathrm{H}^1(\Omega_\mathrm{F})}\|v\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\|\dot w\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\,\mathrm{d}s.\end{aligned}$$

Now, we apply again Hölder's inequality on $(0,T)$ together with the embedding $\mathrm{C}^0(0,T) \hookrightarrow \mathrm{L}^4(0,T)$ and (A.2) for $q = 8$, $s = 3/8$ and $\sigma = 1$ and obtain

$$\begin{aligned}
&\int_0^t \int_{\Omega_\mathrm{F}} |(\dot{u}\cdot\nabla)v\cdot\dot{w}|\,\mathrm{d}y\mathrm{d}s\\
&\le C\left\|\|\dot{u}\|^{1/2}_{\mathrm{L}^2(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^8(0,T)}\left\|\|\dot{u}\|^{1/2}_{\mathrm{H}^1(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^4(0,T)}\left\|\|v\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^8(0,T)}\|\dot{w}\|_{\mathrm{L}^2(\mathrm{H}^1(\Omega_\mathrm{F}))}\\
&\le C\left\|\|\dot{u}\|_{\mathrm{L}^2(\Omega_\mathrm{F})}\right\|^{1/2}_{\mathrm{L}^4(0,T)}\left\|\|\dot{u}\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right\|^{1/2}_{\mathrm{L}^2(0,T)}\left\|\|v\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^8(0,T)}\|w\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}\\
&\le CT^\alpha\|u\|^{1/2}_{\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))}\|u\|^{1/2}_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}\\
&\qquad\times\left(\|v\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}+\|v(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right)\|w\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}\,.
\end{aligned}$$

We can estimate the first term similarly. For the last term, we make use of the same tools to estimate

$$\begin{aligned}
&\int_0^t \int_{\Omega_\mathrm{F}} |(u\cdot\nabla)\dot{v}\cdot\dot{w}|\,\mathrm{d}y\mathrm{d}s\\
&\le C\int_0^t \|u\|_{\mathrm{L}^4(\Omega_\mathrm{F})}\|\nabla\dot{v}\|_{\mathrm{L}^2(\Omega_\mathrm{F})}\|\dot{w}\|_{\mathrm{L}^4(\Omega_\mathrm{F})}\,\mathrm{d}s\\
&\le C\int_0^t \|u\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\|\nabla\dot{v}\|_{\mathrm{L}^2(\Omega_\mathrm{F})}\|\dot{w}\|^{1/2}_{\mathrm{H}^1(\Omega_\mathrm{F})}\|\dot{w}\|^{1/2}_{\mathrm{L}^2(\Omega_\mathrm{F})}\,\mathrm{d}s\\
&\le C\left\|\|u\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^8(0,T)}\left\|\|\nabla\dot{v}\|_{\mathrm{L}^2(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^2(0,T)}\\
&\qquad\times\left\|\|\dot{w}\|^{1/2}_{\mathrm{H}^1(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^4(0,T)}\left\|\|\dot{w}\|^{1/2}_{\mathrm{L}^2(\Omega_\mathrm{F})}\right\|_{\mathrm{L}^8(0,T)}\\
&\le CT^\alpha\left(\|u\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}+\|u(0)\|_{\mathrm{H}^1(\Omega_\mathrm{F})}\right)\|v\|_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}\\
&\qquad\times\|w\|^{1/2}_{\mathrm{H}^1(\mathrm{H}^1(\Omega_\mathrm{F}))}\|w\|^{1/2}_{\mathrm{C}^1(\mathrm{L}^2(\Omega_\mathrm{F}))}.
\end{aligned}$$

□

Approximation of Data

We define

$$A := \mathrm{H}^2(\Omega_\mathrm{F})\times\mathrm{H}^1(\Omega_\mathrm{F})\times\mathrm{H}^1(\Omega_\mathrm{F})\times\mathrm{H}^2(\Omega_\mathrm{S})\times\mathrm{H}^1(\Omega_\mathrm{S})\times\mathrm{L}^2(\Omega_\mathrm{S})$$

and

$$\begin{aligned}
\tilde{A} := {}&\mathrm{H}^{5/2+1/16}(\Omega_\mathrm{F})\times\mathrm{H}^1(\Omega_\mathrm{F})\times\mathrm{H}^{3/2+1/16}(\Omega_\mathrm{F})\\
&\times\mathrm{H}^{5/2+1/16}(\Omega_\mathrm{S})\times\mathrm{H}^{3/2+1/16}(\Omega_\mathrm{S})\times\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{S}).
\end{aligned}$$

Now, we want to show the following approximation result:

Lemma 7 *For given*

$$\begin{aligned} d &:= (u_0, u_1, p_0, \xi_0, \xi_1, \xi_2, f) \\ &\in A \times \left(\mathrm{L}^2(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{F}))\right) \end{aligned}$$

satisfying (4), there exists a sequence

$$d_n := (u_0^n, u_1^n, p_0^n, \xi_0^n, \xi_1^n, \xi_2^n, f^n) \in \tilde{A} \times \left(\mathrm{C}^\infty(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F}))\right),$$

which satisfies (4) for all $n \in \mathbb{N}$ and which converges to d in

$$A \times \left(\mathrm{L}^2(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{F}))\right).$$

Proof To construct such a sequence, we proceed in the following steps:

(1) As $u_1 \in \mathrm{H}^1(\Omega_\mathrm{F})$ with $\mathrm{div}(u_1) = 0$ in Ω_F and $u_1|_{\partial\Omega} = 0$ is already satisfied, we set $u_1^n := u_1$ for all $n \in \mathbb{N}$.

(2) Since $\xi_2 \in \mathrm{L}^2(\Omega_\mathrm{S})$ and $\mathrm{C}_0^\infty(\Omega_\mathrm{S})$ is dense in $\mathrm{L}^2(\Omega_\mathrm{S})$, we can find a sequence $(\hat{\xi}_2^n) \subset \mathrm{C}_0^\infty(\Omega_\mathrm{S})$ such that $\lim_{n\to\infty} \hat{\xi}_2^n = \xi_2$ in $\mathrm{L}^2(\Omega_\mathrm{S})$. To modify this sequence such that it satisfies the compatibility condition on $\partial\Omega_\mathrm{S}$, we first define the sets

$$(\partial\Omega_\mathrm{S})^n := \left\{ y \in \Omega_\mathrm{S} : \mathrm{dist}(y, \partial\Omega_\mathrm{S}) < \frac{1}{2^n} \right\}$$

for $n \in \mathbb{N}$. Then, we can find a sequence $(\varphi^n) \subset \mathrm{C}^\infty(\Omega_\mathrm{S})$ such that

$$\varphi^n(y) = \begin{cases} 1 & \text{if } y \in (\partial\Omega_\mathrm{S})^{n+1}, \\ 0 & \text{if } y \in \Omega_\mathrm{S} \setminus (\partial\Omega_\mathrm{S})^n. \end{cases}$$

Now, let $u_1^E \in \mathrm{H}^1(\Omega)$ denote an extension of u_1 to Ω, and set

$$\xi_2^n := \hat{\xi}_2^n + \varphi^n u_1^E \in \mathrm{H}^1(\Omega_\mathrm{S}).$$

Then, $\xi_2^n|_{\partial\Omega_\mathrm{S}} = u_1|_{\partial\Omega_\mathrm{S}}$ and

$$\|\varphi^n u_1^E\|_{\mathrm{L}^2(\Omega_\mathrm{S})} \le C\|\varphi^n\|_{\mathrm{L}^3(\Omega_\mathrm{S})}\|u_1^E\|_{\mathrm{L}^6(\Omega_\mathrm{S})} \le C|(\partial\Omega_\mathrm{S})^n|^{1/3}\|u_1^E\|_{\mathrm{H}^1(\Omega_\mathrm{S})} \to 0,$$

so we get that $\lim_{n\to\infty} \xi_2^n = \xi_2$ in $\mathrm{L}^2(\Omega_\mathrm{S})$.

(3) Since $\xi_0|_{\partial\Omega_S} \in H^{3/2}(\partial\Omega_S)$, we can choose a sequence $(g^n) \subset H^{2+1/16}(\partial\Omega_S)$ such that $\lim_{n\to\infty} g^n = \xi_0|_{\partial\Omega_S}$ in $H^{3/2}(\partial\Omega_S)$. Because of

$$\operatorname{div}(\Sigma(\xi_0)) = \xi_2,$$

we can construct a sequence $(\xi_0^n) \subset H^{5/2+1/16}(\Omega_S)$ which satisfies $\lim_{n\to\infty} \xi_0^n = \xi_0$ in $H^2(\Omega_S)$ by solving the Dirichlet problem

$$\begin{cases} \operatorname{div}(\Sigma(\xi_0^n)) = \xi_2^n \text{ in } \Omega_S, \\ \qquad\qquad \xi_0^n = g^n \text{ on } \partial\Omega_S, \end{cases}$$

and using Theorem 4 for both $s = 3/2 + 1/16$ and $s = 1$.

(4) Since $H^{1/2+1/16}(\Omega_F) \hookrightarrow H^{-1/2+1/16}(\Omega_F)$ is dense, [16, Theorem 2.1] implies that we find a sequence $(\tilde{f}^n) \subset C^\infty(H^{1/2+1/16}(\Omega_F))$ such that $\lim_{n\to\infty} \tilde{f}^n = f$ in $L^2(H^{1/2+1/16}(\Omega_F)) \cap H^1(H^{-1/2+1/16}(\Omega_F))$. Since d solves

$$\begin{cases} \operatorname{div}(\sigma(u_0, p_0)) = u_1 - f(0) & \text{in } \Omega_F, \\ \operatorname{div}(u_0) = 0 & \text{in } \Omega_F, \\ \sigma(u_0, p_0)n = \Sigma(\xi_0)n & \text{on } \partial\Omega_S, \\ u_0 = 0 & \text{on } \partial\Omega, \end{cases}$$

integration by parts shows that

$$\int_{\Omega_F} f(0)\,\mathrm{d}y - \int_{\Omega_F} u_1\,\mathrm{d}y = -\int_{\partial\Omega_S} \sigma(u_0, p_0)N\,\mathrm{d}S(y) = \int_{\partial\Omega_S} \Sigma(\xi_0)n\,\mathrm{d}S(y).$$

Therefore, we can modify $(\tilde{f}^n)$ by adding suitable constants to obtain a sequence $(f^n) \subset C^\infty(H^{1/2+1/16}(\Omega_F))$ such that

$$\int_{\Omega_F} f^n(0)\,\mathrm{d}y = \int_{\partial\Omega_S} \Sigma(\xi_0^n)n\,\mathrm{d}S(y) + \int_{\Omega_F} u_1^n\,\mathrm{d}y \tag{A.3}$$

and still $\lim_{n\to\infty} f^n = f$ in $L^2(H^{1/2+1/16}(\Omega_F)) \cap H^1(H^{-1/2+1/16}(\Omega_F))$. Moreover, then [16, Theorem 3.1] together with

$$\left(H^{1/2+1/16}(\Omega_F), H^{-1/2+1/16}(\Omega_F)\right)_{1/2} \hookrightarrow L^2(\Omega_F)$$

implies that

$$\|f - f^n\|_{C^0(L^2(\Omega_F))} \le C\|f - f^n\|_{L^2(H^{1/2+1/16}(\Omega_F))\cap H^1(H^{-1/2+1/16}(\Omega_F)} \to 0,$$

so in particular $\lim_{n\to\infty} f^n(0) = f(0)$ in $\mathrm{L}^2(\Omega_\mathrm{F})$.

(5) Next, we consider the Stokes problem

$$\begin{cases} \operatorname{div}(\sigma(u_0^n, p_0^n)) = u_1^n - f^n(0) & \text{in } \Omega_\mathrm{F}, \\ \operatorname{div}(u_0^n) = 0 & \text{in } \Omega_\mathrm{F}, \\ \sigma(u_0^n, p_0^n)n = \Sigma(\xi_0^n)n & \text{on } \partial\Omega_\mathrm{S}, \\ u_0^n = 0 & \text{on } \partial\Omega, \end{cases}$$

for $n \in \mathbb{N}$. Note that $f^n \in \mathrm{C}^\infty(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F}))$ implies $f^n(0) \in \mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F})$. Because of (A.3) together with $u_1^n \in \mathrm{H}^1(\Omega_\mathrm{F})$ and $\Sigma(\xi_0^n)n \in \mathrm{H}^{1+1/16}(\partial\Omega_\mathrm{S})$, we find a sequence of solutions $(u_0^n, p_0^n) \subset \mathrm{H}^{5/2+1/16}(\Omega_\mathrm{F}) \times \mathrm{H}^{3/2+1/16}(\Omega_\mathrm{F})$ by using Theorem 3 for $s = 1/2 + 1/16$. Since

$$\lim_{n\to\infty} u_1^n = u_1 \text{ in } \mathrm{L}^2(\Omega_\mathrm{F}), \quad \lim_{n\to\infty} \Sigma(\xi_0^n)n = \Sigma(\xi_0)n \text{ in } \mathrm{H}^{1/2}(\partial\Omega_\mathrm{S})$$
$$\text{and } \lim_{n\to\infty} f^n(0) = f(0) \text{ in } \mathrm{L}^2(\Omega_\mathrm{F}),$$

for $s = 0$, Theorem 3 implies $\lim_{n\to\infty} u_0^n = u_0$ in $\mathrm{H}^2(\Omega_\mathrm{F})$ and $\lim_{n\to\infty} p_0^n = p_0$ in $\mathrm{H}^1(\Omega_\mathrm{F})$.

(6) Finally, we set $h := \operatorname{div}(\Sigma(\xi_1)) \in \mathrm{H}^{-1}(\Omega_\mathrm{S}))$ and consider the elliptic problem

$$\begin{cases} \operatorname{div}(\Sigma(\xi_1)) = h & \text{in } \mathrm{H}^{-1}(\Omega_\mathrm{S}), \\ \xi_1 = u_0 & \text{on } \partial\Omega_\mathrm{S}. \end{cases}$$

Now, choose some sequence $(h^n) \subset \mathrm{L}^2(\Omega_\mathrm{S})$ such that $\lim_{n\to\infty} h^n = h$ in $\mathrm{H}^{-1}(\Omega_\mathrm{S})$, and consider the elliptic problems

$$\begin{cases} \operatorname{div}(\Sigma(\xi_1^n)) = h^n & \text{in } \Omega_\mathrm{S}, \\ \xi_1^n = u_0^n & \text{on } \partial\Omega_\mathrm{S}. \end{cases}$$

Since it follows from step 5 that $(u_0^n|_{\partial\Omega_\mathrm{S}}) \subset \mathrm{H}^{2+1/16}(\partial\Omega_\mathrm{S})$ and $\lim_{n\to\infty} u_0^n|_{\partial\Omega_\mathrm{S}} = u_0|_{\partial\Omega_\mathrm{S}}$ in $\mathrm{H}^{3/2}(\partial\Omega_\mathrm{S})$, we can use Theorem 4 for both $s = 1$ and $s = 0$ and obtain a sequence of solutions $(\xi_1^n) \subset \mathrm{H}^2(\Omega_\mathrm{S})$ such that $\lim_{n\to\infty} \xi_1^n = \xi_1$ in $\mathrm{H}^1(\Omega_\mathrm{S})$.

Consequently, we have found a compatible sequence

$$(d_n) := (u_0^n, u_1^n, p_0^n, \xi_0^n, \xi_1^n, \xi_2^n, f^n)$$

approximating d in

$$A \times \left(\mathrm{L}^2(\mathrm{H}^{1/2+1/16}(\Omega_\mathrm{F})) \cap \mathrm{H}^1(\mathrm{H}^{-1/2+1/16}(\Omega_\mathrm{F}))\right).$$

□

References

1. G. Avalos, P.G. Geredeli, J.T. Webster, A linearized viscous, compressible flow-plate interaction with non-dissipative coupling. J. Math. Anal. Appl. **477**(1), 334–356 (2019)
2. S. Bechtel, M. Egert, Interpolation theory for Sobolev functions with partially vanishing trace on irregular open sets. J. Fourier Anal. Appl. **25**(5), 2733–2781 (2019)
3. M. Boulakia, Existence of weak solutions for the three-dimensional motion of an elastic structure in an incompressible fluid. J. Math. Fluid Mech. **9**(2), 262–294 (2007)
4. M. Boulakia, S. Guerrero, T. Takahashi, Well-posedness for the coupling between a viscous incompressible fluid and an elastic structure. Nonlinearity **32**, 3548–3592 (2019)
5. D. Coutand, S. Shkoller, Motion of an elastic solid inside an incompressible viscous fluid. Arch. Ration. Mech. Anal. **176**(1), 25–102 (2005)
6. D. Coutand, S. Shkoller, The interaction between quasilinear elastodynamics and the Navier-Stokes equations. Arch. Ration. Mech. Anal. **179**(3), 303–352 (2006)
7. E. Di Nezza, G. Palatucci, E. Valdinoci, Hitchhiker's guide to the fractional Sobolev spaces. Bull. Sci. Math. **136**, 512–573 (2012)
8. C. Grandmont, M. Hillairet, Existence of global strong solutions to a beam-fluid interaction system. Arch. Ration. Mech. Anal. **220**(3), 1283–1333 (2016)
9. C. Grandmont, M. Hillairet, J. Lequeurre, Existence of local strong solutions to fluid-beam and fluid-rod interaction systems. Ann. Inst. H. Poincaré Anal. Non Linéaire **36**(4), 1105–1149 (2019)
10. G. Grubb, V. Solonnikov, Boundary Value Problems for the Nonstationary Navier-Stokes Equations treated by Pseudo-Differential Methods. Math. Scand. **69**, 217–290 (1991)
11. P. Haupt, *Continuum Mechanics and Theory of Materials* (Springer, Berlin, 2002)
12. M. Ignatova, I. Kukavica, I. Lasiecka, A. Tuffaha, Small data global existence for a fluid-structure model. Nonlinearity **30**(2), 848–898 (2017)
13. I. Kukavica, A. Tuffaha, Solutions to a fluid-structure interaction free boundary problem. Discrete Contin. Dyn. Syst. **32**(4), 1355–1389 (2012)
14. I. Kukavica, A. Tuffaha, M. Ziane, Strong solutions to a Navier-Stokes-Lamé system on a domain with a nonflat boundary. Nonlinearity **24**, 159–176 (2011)
15. I. Lasiecka, J.L. Lions, R. Triggiani, Non-homogeneous boundary value problems for second order hyperbolic operators. J. Math. Pures Appl. **65**, 149–192 (1986)
16. J.L. Lions, E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications* (Springer, Berlin, 1972)
17. B. Muha, S. Čanić, Fluid-structure interaction between an incompressible, viscous 3D fluid and an elastic shell with nonlinear Koiter membrane energy. Interfaces Free Bound. **17**(4), 465–495 (2015)
18. J.P. Raymond, M. Vanninathan, A fluid-structure model coupling the Navier-Stokes equations and the Lamé system. J. Math. Pures Appl. **102**, 546–596 (2014)
19. H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators* (North-Holland Publishing Company, Amsterdam, 1978)

Doubly Nonlocal Cahn–Hilliard Equations

Well-posedness and Asymptotic Behavior

Mikil D. Foss, Petronela Radu, and Laura White

1 Introduction

Over the past two decades, nonlocal theories have brought significant advances in continuum mechanics [28], biological aggregation [29], thermal diffusion [5], and image processing [20]. The nonlocality in these integro-differential models is exhibited through integral operators, thus successfully capturing discontinuous behavior of material structures or predicting singular phenomena (such as dynamic fracture modeled by peridynamics [28]). Nonlocal models are also suitable candidates for models with abrupt transitions in the material, such as in phase transitions [3].

In this chapter, we will focus on a doubly nonlocal counterpart of the well-studied Cahn–Hilliard phase-field model for sharp interface limits that appear in the separation of two phases [8, 31]. Cahn–Hilliard systems have been heavily investigated due to their numerous applications in a variety of fields: spinodal decomposition [7], diblock copolymer [9, 23], image inpainting (the process of reconstructing lost parts of images) [4], multiphase fluid flows [2], micro-structures with elastic inhomogeneity [33], tumor growth simulation [21, 32], and topology optimization [14, 34]. An overview of some of these applications can be found in [24].

In [19], the authors derived a macroscopic equation for phase segregation phenomena that gives rise to a nonlocal version of the Cahn–Hilliard equation. Although numerical solutions to the classical equation show good agreement with

M. D. Foss · P. Radu (✉)
Department of Mathematics, University of Nebraska-Lincoln, Lincoln, NE, USA
e-mail: mfoss@unl.edu; pradu@unl.edu

L. White
NASA Langley Research Center, Hampton, VA, USA
e-mail: laura.white@nasa.gov

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_10

experiments, in the classical framework the macroscopic model cannot be obtained from microscopic models for interacting particles, in contrast with the diffusion equation which can be derived from idealized microscopic models in appropriate limits [12, 27]. In order to provide a macroscopic equation for phase separation, the authors of [19] considered long-range interactions which are described by an integral equation, thus deriving a Cahn–Hilliard equation with a single nonlocality. This Cahn–Hilliard equation describes a deterministic process in which the particle transport obeys Fick's law. However, when considering phase separation in a multiscale heterogeneous environment, non-Fickian behavior of the chemical potential is observed. This observation motivates a *fully* nonlocal formulation where integral operators replace all differential operators, giving rise to a doubly nonlocal Cahn–Hilliard equation [6, 16, 17].

In the classical setting, the Cahn–Hilliard equation can be obtained by letting the chemical potential μ in a binary mixture be given as the variational derivative of the energy $\mathcal{E}$ associated with the relative difference of the two phases u, to obtain

$$\mu = \frac{\partial \mathcal{E}}{\partial u} = F'(u) - \varepsilon^2 \Delta u. \tag{1}$$

Above, $0 < \varepsilon^2 \ll 1$ is the coefficient of gradient energy. The potential F is usually a singular logarithmic function, or the double-well potential $F(u) = (u^2 - 1)^2$ defined over $[-1, 1]$, with the values ± 1 denoting the pure two phases and the values in between corresponding to transition phases. Then, using the net flux $\mathcal{J} = -M\nabla\mu$, with M denoting the mobility parameter (here considered constant), the conservation law for u yields the second equation of the classical Cahn–Hilliard system

$$\frac{\partial u}{\partial t} = -\nabla \cdot \mathcal{J} = \nabla \cdot (M\nabla\mu) = M\Delta\mu.$$

The boundary value problem associated with the Cahn–Hilliard system for $\Omega \subset \mathbb{R}^N$, a bounded domain with sufficiently smooth boundary $\partial\Omega$, usually assumes a natural boundary condition for u and a no-flux boundary condition for the chemical potential μ so that

$$\nabla u \cdot n = 0 \quad \text{and} \quad \mathcal{J} \cdot n = -M\nabla\mu \cdot n = 0 \quad \text{on} \quad \partial\Omega,$$

where n is the outward pointing unit normal vector to $\partial\Omega$.

Thus, the initial boundary value problem on Ω (which could be taken as $\mathbb{R}^n$, in which case no boundary conditions are imposed) becomes

$$\begin{cases} \dfrac{\partial u}{\partial t} = M\Delta[F'(u) - \varepsilon^2 \Delta u], & (x,t) \in \Omega \times \mathbb{R}^+, \\ n \cdot \nabla u = n \cdot M\nabla[F'(u) - \varepsilon^2 \Delta u] = 0, & (x,t) \in \partial\Omega \times \mathbb{R}^+, \\ u(x,u) = u_0(x), & x \in \Omega. \end{cases} \tag{2}$$

Two important properties of (2) are the decrease in energy and conservation of mass, which are shown in the sequel to hold even in the doubly nonlocal setting.

The (singly) nonlocal Cahn–Hilliard system of [19] is given by

$$\frac{\partial u}{\partial t} = \nabla \cdot (M \nabla \mu), \quad \text{in } \Omega \times (0, T), \tag{3}$$

where the chemical potential μ is defined by

$$\mu = u \int_{\Omega} \varepsilon J_{\varepsilon}(x-y)dy + \frac{1}{\varepsilon} F'(u) - J * u, \quad \text{in } \Omega \times (0, T).$$

Above, $J_{\varepsilon}(x) = \varepsilon^{N} J(\varepsilon^{-1}x)$ and $J : \mathbb{R}^N \to \mathbb{R}$ is a convolution kernel such that $J(x) = J(-x)$. A formal asymptotic analysis shows that the interface evolution problems associated with (3) as $\varepsilon \to 0$ are similar to the ones associated with the standard Cahn–Hilliard equation. Moreover, the nonlocal equation can be viewed as the conserved gradient flow of the first variation of the free energy functional

$$\mathcal{N}(u) = \int_{\Omega \times \Omega} \frac{\varepsilon}{4} J_{\varepsilon}(x-y)|u(x)-u(y)|^2 dxdy + \int_{\Omega} \frac{1}{\varepsilon} F(u) dx. \tag{4}$$

Van der Waals noted that $\mathcal{E}$ in (1) can be considered a local approximation of (4) [30]. Therefore, the nonlocal Cahn–Hilliard equation appears to be justified and more general than the classical one. This work focuses on a *doubly* nonlocal Cahn–Hilliard system (see Sect. 3), which replaces all differential operators with integral operators. The doubly nonlocal system was first introduced and studied in [16, 17] where well-posedness and regularity of solutions were proved with one of the interaction kernels chosen to be singular. It is known that in such cases, compact embedding and regularity results are available, which are lacking for the current framework with integrable kernels.

Contributions of This Work and Organization of Results

To offer a self-contained presentation, we introduce the nonlocal operator framework in Sect. 2. The analysis of Sect. 3 focuses on the long time behavior of solutions for the doubly nonlocal linearized system with time-dependent coefficients for which we derive *explicit decay rates for the L^p norms of the solution*, for all $p \geq 1$. Section 4 focuses on the steady-state system. First, well-posedness is proved in the linear setting. We then establish *regularity and higher integrability properties of solutions in a nonlinear setting*. The nonlinearities permitted by the arguments have a general structure but are short of satisfying some of the physically relevant assumptions. Some open problems and future directions are presented in Sect. 5.

2 Nonlocal Vector Calculus

In order to present and study the doubly nonlocal formulation of (2), we introduce a framework of nonlocal operators, together with some identities and other properties. This setting is based on [13], although some definitions are slightly modified.

The basic idea of nonlocality relies on incorporating values of a function in a neighborhood around a point through (usually) an integral operator. In some applications, the interactions are within a finite range, called a horizon, which mathematically is expressed through the size of a kernel's support. The nonlocal operators defined below incorporate a symmetric integrable kernel $\alpha : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ (i.e., $\alpha(x, y) = \alpha(y, x)$), which captures the interactions between nearby points x and y. The nonlocal operators are defined as follows:

Definition 1 Let $\Omega \subset \mathbb{R}^N$ be an open set (possibly the entire $\mathbb{R}^n$), and let the α-interaction collar $\Gamma \subset \mathbb{R}^N \setminus \Omega$ be a "boundary" domain that contains all points of nonzero interaction, i.e.,

$$\Gamma_\alpha := \{y \in \mathbb{R}^n \setminus \Omega | \exists\, x \in \Omega \text{ such that } \alpha(x, y) \neq 0\}. \tag{5}$$

For α and $\Gamma_\alpha = \Gamma$ thus given, we define the following:

(i) **Nonlocal Divergence.** Given a vector field $\nu : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$, the nonlocal divergence operator is defined as

$$\mathcal{D}_\alpha[\nu](x) := \int_{\Omega\cup\Gamma} (\nu(x, y) - \nu(y, x))\, \alpha(x, y) dy, \qquad \text{for } x \in \Omega. \tag{6}$$

(ii) **Nonlocal Gradient.** Given a scalar function $w : \mathbb{R}^N \to \mathbb{R}$, the nonlocal (two-point) gradient operator is defined as

$$\mathcal{G}_\alpha[w](x, y) := (w(y) - w(x))\alpha(x, y) dy, \qquad \text{for } x, y \in \mathbb{R}^N. \tag{7}$$

(iii) **Nonlocal Laplacian.** Given $w : \Omega \cup \Gamma \to \mathbb{R}$, the nonlocal Laplacian operator is defined for $x \in \mathbb{R}^N$ as

$$\mathcal{L}_{\alpha^2}[w](x) := \mathcal{D}_\alpha[\mathcal{G}_\alpha[w]](x) = 2 \int_{\Omega\cup\Gamma} (w(y) - w(x))\alpha^2(x, y) dy. \tag{8}$$

(iv) **Nonlocal Normal.** Given a vector field $\nu : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$, the nonlocal (interaction) operator is defined as

$$\mathcal{N}_\alpha[\nu](x) := \int_{\Omega\cup\Gamma} (\nu(y, x) - \nu(x, y))\, \alpha(x, y) dy, \qquad \text{for } x \in \Gamma. \tag{9}$$

Equation (8) introduces the nonlocal Laplace operator that was considered in peridynamic models [28] and nonlocal diffusion [26]. Note that the definitions of the nonlocal Laplacian and nonlocal normal are formally the same; however, the Laplacian is defined for points inside Ω, while the normal is considered only for points in the boundary Γ.

Remark 1 Note that the definitions (6) and (8) need the kernels α and α^2 to be integrable. However, one can extend these definitions by using principal value operators, thus allowing nonintegrable kernels.

The definitions above form the skeleton of a nonlocal framework in which nonlocal counterparts for integration by parts and Green's identities hold, as follows:

Proposition 1 *Let $u, v : \Omega \cup \Gamma \to \mathbb{R}$ and $\nu : (\Omega \cup \Gamma)^2 \to \mathbb{R}$. Then, for $\alpha \in L^2((\Omega \cup \Gamma)^2)$ with α^2 symmetric ($\alpha^2(x, y) = \alpha^2(y, x)$), the following hold:*

(i)

$$\int_{\Omega} v\mathcal{D}_{\alpha}[\nu]\,dx + \int_{\Omega\cup\Gamma}\int_{\Omega\cup\Gamma} \nu\,\mathcal{G}_{\alpha}[v]\,dxdy = \int_{\Gamma} \mathcal{N}_{\alpha}[\nu]v\,dx. \tag{10}$$

(ii)

$$\int_{\Omega} v\mathcal{L}_{\alpha^2}[u]\,dx + \int_{\Omega\cup\Gamma}\int_{\Omega\cup\Gamma} \mathcal{G}_{\alpha}[u]\,\mathcal{G}_{\alpha}[v]\,dxdy = \int_{\Gamma} \mathcal{N}_{\alpha}[\mathcal{G}_{\alpha}[u]]v\,dx. \tag{11}$$

(iii) For a function u such that $\mathcal{N}_{\alpha}[\mathcal{G}_{\alpha}[u]] = 0$ on Γ, we have

$$\int_{\Omega} \mathcal{L}_{\alpha^2}[u]\,dx = 0. \tag{12}$$

The proofs of statements (i) and (ii) are available in [13], while (iii) is a simple consequence of (ii).

3 Asymptotic Behavior of Solutions to Doubly Nonlocal Cahn–Hilliard Systems

Consider the Cauchy problem for the doubly nonlocal Cahn–Hilliard system on the entire space:

$$\begin{cases} u_t(x,t) &= \mathcal{L}_K[\mu](x,t), \quad (x,t) \in \mathbb{R}^N \times (0,\infty) \\ \mu(x,t) &= -\mathcal{L}_J[u](x,t) + F'(u(x,t)), \quad (x,t) \in \mathbb{R}^N \times (0,\infty) \\ u(x,0) &= u_0(x), \quad x \in \mathbb{R}^N, \end{cases} \tag{13}$$

where $\mathcal{L}$ is the nonlocal Laplacian defined in (8) with $K(x, y)$ and $J(x, y)$ as symmetric kernels. Here, the kernel K encodes physical properties of the environment, in which the mass is being transported, while J represents the long-range Kac (symmetric) potentials.

We begin by showing that in the nonlocal setting, the energy is nonincreasing and the mass is conserved, as it is the case with the classical system as well.

Proposition 2 *Let J be a nonnegative kernel that in addition is*

(i) Radial: $J(x,y) = J(|x-y|)$ (with abuse of notation)
(ii) Integrable: $J(\cdot) \in L^1(\mathbb{R})$

Then, the energy associated with (13)

$$\begin{aligned}\mathcal{E}(t) := &\frac{1}{2}\int\int |u(y,t)-u(x,t)|^2 J(|y-x|)\,dxdy + \int F(u(x,t))\,dx \\ = &\frac{1}{2}\int\int |\mathcal{G}_{\sqrt{J}}[u](x,y,t)|^2\,dxdy + \int F(u(x,t))\,dx\end{aligned} \tag{14}$$

satisfies $\dfrac{d}{dt}\mathcal{E}(t) \leq 0$. *Moreover, if u is a solution to* (2)*, then the total mass is constant, i.e.,* $\dfrac{d}{dt}\displaystyle\int u\,dx = 0$.

Proof Differentiating the energy and performing an integration by parts (apply Proposition 1 with $\Omega = \mathbb{R}^N$, $\Gamma = \emptyset$) gives

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(t) &= \int\int \mathcal{G}_{\sqrt{J}}[u]\,\mathcal{G}_{\sqrt{J}}[u_t]\,dxdy + \int F'(u)u_t dx \\ &= \int \big(-\mathcal{L}_J[u] + F'(u)\big)\,u_t\,dx \\ &= \int \mu\frac{\partial u}{\partial t}dx = \int \mu\mathcal{L}_K[u]dx = -\int |\mathcal{G}_{\sqrt{K}}[\mu]|^2 dx.\end{aligned}$$

To show the conservation of mass, we apply part (iii) of Proposition 1 as follows:

$$\frac{d}{dt}\int u dx = \int \frac{\partial u}{\partial t}dx = \int \mathcal{L}_K[\mu]dx = 0.$$

□

The above theorem verifies physical properties that hold for the two-phase system. Indeed, as time increases, the two materials become more separated which in return requires less energy. The theorem holds on bounded domains as well, under the no-flux boundary condition.

3.1 Decay Estimates for the Linearized System with Time-Dependent Coefficients

In this subsection, we produce a fundamental solution to the linear doubly nonlocal Cahn–Hilliard equation and establish that this fundamental solution can be decomposed into a smooth part and a rapidly decaying part. The argument adapted from [26] leads directly into the study of asymptotic behavior of solutions for this equation.

Taking $F'(u) = \alpha(t)u$ and $T > 0$ fixed, we rewrite (13) as

$$\begin{cases} \dfrac{\partial u}{\partial t} = \mathcal{L}_K(\mu) & (t,x) \in [0,T] \times \mathbb{R}^N, \\ \mu = \mathcal{L}_J(u) + \alpha(t)u & (t,x) \in [0,T] \times \mathbb{R}^N. \end{cases} \tag{15}$$

Equation (15) is a linearization of (13) and a natural starting point for the investigations of the fully nonlinear case. For the results of this section, we make the following assumptions on J and K:

(A1) $J, K \in C(\mathbb{R}^N)$ are nonnegative radial functions with $J(0), K(0) > 0$ and $\displaystyle\int J(x)dx = \int K(x) = 1$.

(A2) J and K are smooth and $J, K \in \mathcal{S}(\mathbb{R}^N)$, the space of rapidly decreasing functions.

(A3) $\widehat{J}(\xi) \sim 1 - \xi^2 + |\xi|^3$ and $\widehat{K}(\xi) \sim 1 - \xi^2 + |\xi|^3$, for ξ close to 0.

(A4) $\widehat{J}(\xi), \widehat{K}(\xi) \in L^1(\mathbb{R}^N)$,

where $\widehat{F}$ denotes the Fourier transform of the function F.

First, we show that the fundamental solution of (15) can be split into a smooth part (which will have rational decay) and a rough part that exhibits exponential decay. This will allow us to establish explicit decay rates for (15) in terms of α.

Lemma 1 *For $\alpha \in L^1(0,T)$, $T > 0$, and J and K satisfying* (1)–(4) *above, the fundamental solution u satisfying*

$$\begin{cases} u_t = \mathcal{L}_K[\mu] & (t,x) \in [0,T] \times \mathbb{R}^N, \\ \mu = -\mathcal{L}_J[u] + \alpha(t)u & (t,x) \in [0,T] \times \mathbb{R}^N, \\ u(0) = \delta_0 & (t,x) \in [0,T] \times \mathbb{R}^N \end{cases} \tag{16}$$

can be decomposed as

$$S(t,x) = e^{-t-\int_0^t \alpha(s)ds}\delta_0 + H(t,x), \tag{17}$$

with $H(t,x)$ a smooth C^∞ function. Moreover, if u is the solution with initial condition u_0, then

$$u(t,x) = (S * u_0)(t,x) = \int S(t, x-y)u_0(y)dy$$
$$= e^{-t-\int_0^t \alpha(s)ds} u_0 + \int H(t, x-y)u_0(y)dy.$$

Proof Applying the Fourier transform to (16), we obtain

$$\begin{cases} \widehat{u}_t &= -\widehat{\mu} + \widehat{\mu}\widehat{K} = -\widehat{\mu}(1-\widehat{K}) \\ \widehat{\mu} &= \widehat{u} - \widehat{J}\widehat{u} + \alpha(t)\widehat{u} = \widehat{u}(1-\widehat{J}+\alpha(t)). \end{cases}$$

The initial data implies $\widehat{u_0} = \widehat{\delta_0} = 1$, and hence

$$\widehat{u}(t) = e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds}$$
$$= e^{-t-\int_0^t \alpha(s)ds} + e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds} - e^{-t-\int_0^t \alpha(s)ds}.$$

Applying the inverse Fourier transform yields the first part of the lemma. In order to show that

$$H(t,x) = \int \left(e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds} - e^{-t-\int_0^t \alpha(s)ds} \right) e^{ix\cdot\xi} d\xi$$

is smooth, define

$$A(t,\xi) = -\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds$$

and

$$B(t) = -t - \int_0^t \alpha(s)ds.$$

Thus,

$$A(t,\xi) - B(t) = \widehat{K}t + \widehat{J}(1-\widehat{K})t + \widehat{K}\int_0^t \alpha(s)ds.$$

Factoring $e^{B(t)}$ and using the analytic expansion around the origin of the exponential function give

$$e^{A(t,\xi)} - e^{B(t)} = e^{B(t)} \left(A(t,\xi) - B(t) + \frac{(A(t,\xi)-B(t))^2}{2} + \frac{(A(t,\xi)-B(t))^3}{3!} \right.$$
$$\left. + \mathcal{O}(|A(t,\xi) - B(t)|^4) \right).$$

By assumption (2), $\widehat{J}, \widehat{K} \in \mathcal{S}(\mathbb{R}^N)$. Thus, $\widehat{J}, \widehat{K} \to 0$ as $\xi \to \infty$, which implies

$$\xi^k (A(t,\xi) - B(t)) \to 0 \quad \text{as } \xi \to \infty,$$

for any k nonnegative integer. Hence, $H(t,x)$ is smooth. For the second part, observe that $S * u_0$ is a solution with initial data u_0. □

Next, we will use the fundamental solution given by (17) to derive an explicit decay rate for the L^p-norm of a positive solution.

Theorem 1 *Assume J and K satisfy* (1)–(4) *and that $u \geq 0$ is a solution for* (16). *Assume, also, that $\alpha \in L^1(\mathbb{R})$ and $\alpha(t) > 0$ for all t. Then, for any $p \geq 1$, there exists a constant $c(p, J, K) \geq 0$ such that*

$$H(t,x) = \int \left(e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds} - e^{-t-\int_0^t \alpha(s)ds} \right) e^{ix\cdot\xi} d\xi$$

satisfies

$$\|H(t,x)\|_{L^p(\mathbb{R}^N)} \leq c(p,J,K) \max \left\{ t^{-N/4}, \left(\int_0^t \alpha(s)ds \right)^{-N/2} \right\}^{(1-\frac{1}{p})} \tag{18}$$

for all $t > 0$ sufficiently large.

Proof The estimate will be obtained by interpolating the estimates obtained for $p = 1$ and $p = \infty$.

Control of $p = \infty$-Norm By using the Hausdorff–Young inequality, we obtain

$$\|H(x,t)\|_{L^\infty(\mathbb{R}^N)} \leq \int |e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds} - e^{-t-\int_0^t \alpha(s)ds}| d\xi.$$

Observe that the symmetry of J and K guarantees that $\widehat{J}$ and $\widehat{K}$ are real-valued. Choose $R > 0$ such that

$$|\widehat{J}(\xi)| \leq 1 - \frac{|\xi|^2}{2} \quad \text{and} \quad |\widehat{K}(\xi)| \leq 1 - \frac{|\xi|^2}{2}$$

for all $|\xi| \leq R$.

Since $\widehat{J}, \widehat{K} \in L^1(\mathbb{R}^N)$ and R is fixed, there exist $\delta = \delta(J,K)$, $0 < \delta < 1$, with

$$|\widehat{J}(\xi)| \leq 1 - \delta \quad \text{and} \quad |\widehat{K}(\xi)| \leq 1 - \delta, \quad \text{for all } |\xi| \geq R.$$

For any pair of real numbers a and b, the following inequality holds:

$$|e^a - e^b| \leq |a - b| \max\{e^a, e^b\}.$$

We deduce that, for each $|\xi| \geq R$,

$$\begin{aligned}
&|e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds} - e^{-t-\int_0^t \alpha(s)ds}| \\
&\quad \leq \max\{e^{-t-\int_0^t \alpha(s)ds}, e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds}\} \left|\widehat{K}t + \widehat{J}(1-\widehat{K})t + \widehat{K}\int_0^t \alpha(s)ds\right| \\
&\quad \leq e^{-t\delta^2-\delta\int_0^t \alpha(s)ds} \left|\widehat{K}t + \widehat{J}(1-\widehat{K})t + \widehat{K}\int_0^t \alpha(s)ds\right|.
\end{aligned}$$

Consequently, the following integral decays exponentially in time:

$$\begin{aligned}
|I_1(t)| := &\left|\int_{|\xi|\geq R} e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K}ds} - e^{-t-\int_0^t \alpha(s)ds} d\xi\right| \\
&\leq e^{-t\delta^2-\delta\int_0^t \alpha(s)ds} \int_{|\xi|\geq R} \left|\widehat{K}t + \widehat{J}(1-\widehat{K})t + \widehat{K}\int_0^t \alpha(s)ds\right| d\xi.
\end{aligned}$$

It remains to verify that

$$I_2(t) := \int_{|\xi|\leq R} e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds} - e^{-t-\int_0^t \alpha(s)ds} d\xi$$

decays in time. For this, we first have the bound

$$\begin{aligned}
|I_2(t)| &\leq \int_{|\xi|\leq R} e^{-\int_0^t (1-\widehat{J}+\alpha(s))(1-\widehat{K})ds} d\xi + C(R)e^{-t-\int_0^t \alpha(s)ds} \\
&\leq \int_{|\xi|\leq R} e^{-t\frac{|\xi|^4}{4}-\frac{|\xi|^2}{2}\int_0^t \alpha(s)ds} d\xi + C(R)e^{-t-\int_0^t \alpha(s)ds}.
\end{aligned}$$

In the sequel, we will use the following exponential inequality, a consequence of Jensen's inequality:

$$e^{\frac{a+b}{2}} \leq \frac{e^a + e^b}{2}.$$

With $a = -t\dfrac{|\xi|^4}{2}$ and $b = -|\xi|^2 \displaystyle\int_0^t \alpha(s)ds$, we obtain

$$\begin{aligned}
|I_2(t)| &\leq \frac{1}{2}\int_{|\xi|\leq R} e^{-t\frac{|\xi|^4}{2}} d\xi + \frac{1}{2}\int_{|\xi|\leq R} e^{-|\xi|^2\int_0^t \alpha(s)ds} d\xi + C(R)e^{-t-\int_0^t \alpha(s)ds} \\
&= t^{-N/4}\frac{2^{-N/4}}{2}\int_{|\eta|\leq R(t/2)^{1/4}} e^{-|\eta|^4} d\eta
\end{aligned}$$

$$+\left(\int_0^t \alpha(s)ds\right)^{-N/2} \frac{2^{-N/2}}{2} \int_{|\eta|\le R\sqrt{\int_0^t \alpha(s)ds}} e^{-|\eta|^2 d\eta} + C(R)e^{-t-\int_0^t \alpha(s)ds}$$
$$\le C \max\left\{t^{-N/4}, \left(\int_0^t \alpha(s)ds\right)^{-N/2}\right\}.$$

This concludes the proof for the L^∞-norm; more precisely, we have that

$$\|H(x,t)\|_{L^\infty(\mathbb{R}^N)} \le C \max\left\{t^{-N/4}, \left(\int_0^t \alpha(s)ds\right)^{-N/2}\right\}. \tag{19}$$

Control of the L^1-Norm From Proposition 2 with $F(u) = \alpha(t)\frac{u^2}{2}$, we have that the mass of the solution is conserved. Under the additional assumption that the u is nonnegative, we have that the L^1-norm is conserved.

Hence, for any $u_0 \in L^1(\mathbb{R}^N)$,

$$\int |e^{-t}u_0(x) + (H(t,x) * u_0)(x)|dx \le \int |u_0(x)|dx,$$

and as a consequence,

$$\int |(H(t,x) * u_0)(x)|dx \le 2\int |u_0(x)|dx.$$

Choosing $(u_0)_n \in L^1(\mathbb{R}^N)$ such that $(u_0)_n \to \delta_0$ in $\mathcal{S}'(\mathbb{R}^N)$, we obtain, in the limit, that

$$\|H(x,t)\|_{L^1(\mathbb{R}^N)} = \int |H(t,x)|dx \le 2. \tag{20}$$

By interpolating the L^1 and L^∞ decay estimates obtained in (19) and (20) L^p, we obtain the inequality claimed in (18).

□

4 Steady-State Solutions

This section focuses on the Cahn–Hilliard system on a bounded domain, with associated boundary data imposed on "twin" layers corresponding to each kernel K, respectively, J that model the nonlocal interactions. Moreover, we consider the system as $t \to \infty$, i.e., steady-state solutions.

4.1 Well-posedness of Solutions

We begin by showing well-posedness for the linearized steady-state boundary problem corresponding to (13). As mentioned above, the formulation involves a double-layer boundary, each accommodating the nonlocal interactions for kernels K, respectively, J. For an open bounded set Ω with smooth boundary, consider

$$\Gamma_K := \{x \in \mathbb{R}^n \setminus \Omega | K(x, y) \neq 0, \text{ for some } y \in \Omega\}. \tag{21}$$

Surrounding the collar Γ_K, we have the additional layer Γ_J given by

$$\Gamma_J := \{x \in \mathbb{R}^n \setminus (\Omega \cup \Gamma_K) | J(x, y) \neq 0, \text{ for some } y \in \Omega \cup \Gamma_K\}. \tag{22}$$

For simplicity, we assume that $K(x, y) = 0$ if $|x - y| \geq \delta$ and $J(x, y) = 0$ if $|x - y| \geq \varepsilon$. A representation of the domain and its double-collar boundary is provided in Fig. 1.

The boundary value problem for the doubly nonlocal Cahn–Hilliard system with homogeneous Dirichlet-type boundary conditions is given by

$$\begin{cases} \mathcal{L}_K[\mu] = f & x \in \Omega \\ \mu = 0, & x \in \Gamma_K \\ \mu = -\mathcal{L}_J[u] + F'(u) & x \in \Omega \cup \Gamma_K \\ u = 0 & x \in \Gamma_J. \end{cases} \tag{23}$$

For the linearized problem, we will consider

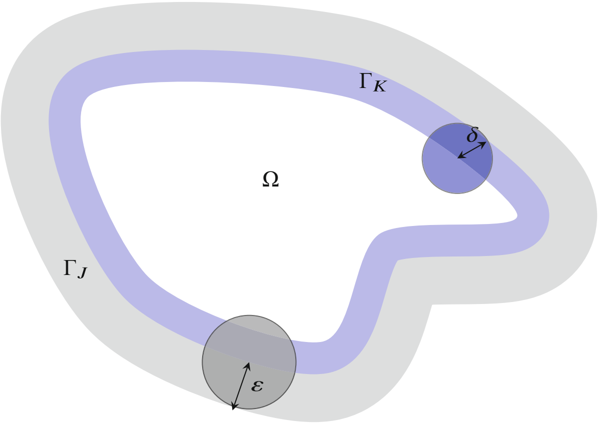


Fig. 1 The domain Ω with the induced nonlocal boundary layers Γ_K (with balls of horizons δ) and Γ_J with balls of horizons ε centered at points inside the domain Ω and on the boundary $\partial\Omega$

$$F'(u) = a(x)u + b(x). \tag{24}$$

Theorem 2 *Given $J, K \in L^1(\mathbb{R}^n)$ satisfying (A1), and the coefficient functions from* (24) *$a, b \in L^2(\Omega)$, satisfying $a \geq 0$, the system* (23) *has a unique solution.*

Proof The system is written as an iteration of two nonlocal problems with domains and boundaries conveniently chosen. First, the existence and uniqueness of a solution $\mu \in L^2(\Omega \cup \Gamma)$ for the system

$$\begin{cases} \mathcal{L}_K[\mu] = f & \text{in } \Omega \\ \mu = 0 & \text{on } \Gamma_K \end{cases} \tag{25}$$

follows easily from well-posedness results established in [1, 15, 22]. Next, consider the system

$$\begin{cases} -\mathcal{L}_J[u] + F'(u) = \mu & \text{in } \Omega \cup \Gamma_K \\ u = 0 & \text{on } \Gamma_J, \end{cases} \tag{26}$$

where $F'(u)$ is given by (24) and Γ_J is the collar of $\Omega \cup \Gamma_K$. The energy functional associated with this system satisfies the assumptions of the existence and uniqueness Theorem 4.2 in [15]. Alternatively, one can easily show continuity and coercivity for the bilinear form

$$B(u, v) := \int_{(\Omega\cup\Gamma_K\cup\Gamma_J)^2} \Big(\mathcal{G}_{\sqrt{J}}[u](x, y)\mathcal{G}_{\sqrt{J}}[v](x, y) + a(x)u(x)v(x)\Big)\, dxdy \tag{27}$$

for $v \in V := \Big\{v \in L^2(\Omega \cup \Gamma_K) : v = 0 \text{ on } \Gamma_J\Big\}$. □

Remark 2 Nonlocal Neumann-type (zero-flux) boundary conditions may be imposed in systems (25) and (26), which would similarly yield existence and uniqueness (up to a constant) of solutions.

Remark 3 The assumptions in Theorem 4.2 in [15] can be verified also by *nonlinear convex* profiles for the function F. However, physical considerations for the Cahn–Hilliard system impose that F is a double-well potential.

4.2 Regularity of Steady-State Solutions in the Nonlinear Settings

We have shown well-posedness of the steady-state solution to the linear doubly nonlocal Cahn–Hilliard equation. We now turn to the regularity of the solution.

Theorem 3 (Regularity of Solutions) *Let $K, J \in C^k(\mathbb{R})$ for $k \in [1, \infty]$, satisfying $\|K\|_{L^1(\mathbb{R})} = \|J\|_{L^1(\mathbb{R})} = 1$ be given. For given $f \in C^k(\mathbb{R})$ and $F : \mathbb{R} \to \mathbb{R}$, let $\mu \in L^1(\mathbb{R})$ be a solution of*

$$\mathcal{L}_K[\mu](x) = \int_{\mathbb{R}} (\mu(y) - \mu(x))K(x - y)dy = f(x) \tag{28}$$

and $u \in L^1(\mathbb{R})$ be a solution to

$$-\mathcal{L}_J[u] + F'(u) = \mu. \tag{29}$$

Additionally, assume that

(H1) F is differentiable and $F' \in C^k(\mathbb{R})$.
(H2) The function $v \mapsto u + F'(u)$ is invertible, so there exists g such that

$$g(u + F'(u)) = u. \tag{30}$$

Additionally, assume that $g \in C^k(\mathbb{R})$.

Then, $u \in C^k(\mathbb{R})$.

Proof Note that since μ is a solution to (28), then the convolution structure of the nonlocal Laplacian gives that μ satisfies

$$\mu * K - \mu = f,$$

and hence $\mu = \mu * K + f$. Due to the fact that $f, K \in C^k(\mathbb{R})$, we find that $\mu \in C^k(\mathbb{R})$. The equality in (29) implies that

$$u + F'(u) = \mu + u * J,$$

so $u = g(\mu + u * J)$. The C^k regularity of g in assumption (30) together with the fact that $J \in C^k$ yields the conclusion $u \in C^k(\mathbb{R})$. □

An immediate corollary for the above theorem applies to the linear steady-state Cahn–Hilliard system.

Corollary 1 *With f, J, K satisfying the assumptions of the Theorem 3 and $F'(u) = \alpha(x)u$ such that $\dfrac{1}{1 + \alpha(x)} \in C^k(\mathbb{R})$, we have $u \in C^k(\mathbb{R})$.*

Proof The proof is immediate as F clearly satisfies assumption (H1). □

4.3 Higher Integrability of Steady-State Solutions

The convolution structure of the nonlocal Laplacian can be employed in a similar manner to show that solutions to the doubly nonlocal Cahn–Hilliard steady-state problem enjoy higher integrability as well. This result is an extension of the results of [15] for doubly nonlocal systems.

Theorem 4 *For $1 < q \leq \infty$, let $K, J \in L^q(\mathbb{R})$, satisfying $\|K\|_{L^1(\mathbb{R})} = \|J\|_{L^1(\mathbb{R})} = 1$. For $f \in L^q(\mathbb{R})$ and $F : \mathbb{R} \to \mathbb{R}$, let $\mu, u \in L^1(\mathbb{R})$ be solutions of (28) and (29). Assume that g given by assumption (H2) from Theorem 3 satisfies*

$$g \circ v \in L^q(\mathbb{R})$$

for every $v \in L^q(\mathbb{R})$. Then, $u \in L^q(\mathbb{R})$.

Proof The proof follows the exact steps of Theorem 3 with the only difference that Young's inequality is employed for the integrability of the convolution product (rather than differentiability properties of the convolution). As a particular case, note that boundedness of solutions may be obtained. □

5 Conclusions and Future Directions

The introduction of the doubly nonlocal system of Cahn–Hilliard equations (8) with integrable kernels K and J is motivated by the fact that its formulation allows discontinuous solutions, with possibly only L^2 regularity. The system is considered on the entire space and also on bounded domains where special consideration must be given to the "layering" of boundary conditions (as seen in [25], it is important to correctly formulate the distribution of data on the boundary layers in higher order problems). While here we only considered homogeneous boundary conditions, the system with nonlocal Neumann-type nonhomogeneous data will bring forward additional issues (indeed, different nonlocal Neumann-type boundary conditions have been investigated in many works recently [10, 11, 18]). Well-posedness for bounded domains was established for the linearized system, and however, for a nonlinear system, we showed that, under additional assumptions on the interaction kernel, the solutions enjoy increased integrability and regularity properties. These results for the steady-state system were obtained by using the convolution structure of the integral operators. For the evolution system, in the linear setting, we obtained explicit decay estimate rates using the Fourier transform.

The main directions for the future research can be summarized as follows:

- Establish well-posedness of solutions for the doubly nonlocal Cahn–Hilliard system, when the nonlinearity is a polynomial, or of a logarithmic type (see [6], for example), as taken commonly in physical applications. While the investigations for the steady system may pose less difficulty, the evolution

problem, with the selection of the layers (as in the system (23)) appears highly nontrivial.
- Show convergence of solutions for the doubly nonlocal system to its classical counterpart. This project would require convergence results for boundary value problems with Neumann-type boundary conditions, in the nonlinear case, which are not available even for steady-state problems of the second order.
- Study asymptotic behavior ($t \to \infty$) of solutions for the doubly nonlocal Cahn–Hilliard system with nonlinear potentials, on bounded, and unbounded domains. While a series of results (see the monograph [26]) seems to indicate that nonlocal evolution problems and classical counterparts seem to enjoy similar behavior at infinity in the linear setting, less is known for nonlinear problems.

Acknowledgments The first and second authors were supported by the award NSF – DMS 1716790.

References

1. B. Aksoylu, T. Mengesha, Results on nonlocal boundary value problems. Numer. Funct. Anal. Optim. **31**(12), 1301–1317 (2010)
2. V.E. Badalassi, H.D. Ceniceros, S. Banerjee, Computation of multiphase systems with phase field models. J. Comput. Phys. **190**(2), 371–397 (2003)
3. P.W. Bates, J. Han, The Dirichlet boundary problem for a nonlocal Cahn–Hilliard equation. J. Math. Anal. Appl. **311**(1), 289–312 (2005)
4. A.L. Bertozzi, S. Esedoglu, A. Gillette, Inpainting of binary images using the Cahn–Hilliard equation. IEEE Trans. Image Process. **16**(1), 285–291 (2007)
5. F. Bobaru, M. Duangpanya, A peridynamic formulation for transient heat conduction in bodies with evolving discontinuities. J. Comput. Phys. **231**(7), 2764–2785 (2012)
6. O. Burkovska, M. Gunzburger, On a nonlocal Cahn-Hilliard model permitting sharp interfaces. Preprint. arXiv:2004.14379 (2020)
7. J.W. Cahn, On spinodal decomposition. Acta Metall. **9**(9), 795–801 (1961)
8. J.W. Cahn, J.E. Hilliard, Free energy of a nonuniform system. i. interfacial free energy. J. Chem. Phys. **28**(2), 258–267 (1958)
9. R. Choksi, M.A. Peletier, J.F. Williams, On the phase diagram for microphase separation of diblock copolymers: An approach via a nonlocal Cahn–Hilliard functional. SIAM J. Appl. Math. **69**(6), 1712–1738 (2009)
10. C. Cortazar, M. Elgueta, J.D. Rossi, N. Wolanski, How to approximate the heat equation with Neumann boundary conditions by nonlocal diffusion problems. Arch. Rational Mech. Anal. **187**(1), 137–156 (2008)
11. M. D'Elia, X. Tian, Y. Yu, A physically consistent, flexible, and efficient strategy to convert local boundary conditions into nonlocal volume constraints. SIAM J. Sci. Comput. **42**(4), A1935–A1949 (2020)
12. A. DeMasi, E. Presutti, *Mathematical Methods for Hydrodynamic Limits* (Springer, 2006)
13. Q. Du, M. Gunzburger, R.B. Lehoucq, K. Zhou, A nonlocal vector calculus, nonlocal volume-constrained problems, and nonlocal balance laws. Math. Models Methods Appl. Sci. **23**(03), 493–540 (2013)
14. M.H. Farshbaf-Shaker, C. Heinemann, A phase field approach for optimal boundary control of damage processes in two-dimensional viscoelastic media. Math. Models Methods Appl. Sci. **25**(14), 2749–2793 (2015)

15. M.D. Foss, P. Radu, C. Wright, Existence and regularity of minimizers for nonlocal energy functionals. Differential Integral Equations **31**(11/12), 807–832 (2018)
16. C.G. Gal, On the strong-to-strong interaction case for doubly nonlocal Cahn-Hilliard equations. Discrete Continuous Dynam. Syst. A **37**(1), 131 (2017)
17. C.G. Gal, Doubly nonlocal Cahn–Hilliard equations, in *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, vol. 35 (Elsevier, 2018), pp. 357–392
18. C.G. Gal, M. Warma, Nonlocal transmission problems with fractional diffusion and boundary conditions on non-smooth interfaces. Commun. Partial Differential Equations **42**(4), 579–625 (2017)
19. G. Giacomin, J.L. Lebowitz, Phase segregation dynamics in particle systems with long range interactions. I. macroscopic limits. J. Stat. Phys. **87**(1–2), 37–61 (1997)
20. G. Gilboa, S. Osher, Nonlocal operators with applications to image processing. Multiscale Model. Simul. **7**(3), 1005–1028 (2008)
21. D. Hilhorst, J. Kampmann, T.N. Nguyen, K.G. Van Der Zee, Formal asymptotic limit of a diffuse-interface tumor-growth model. Math. Models Methods Appl. Sci. **25**(06), 1011–1043 (2015)
22. B. Hinds, P. Radu, Dirichlet's principle and wellposedness of solutions for a nonlocal p-Laplacian system. Appl. Math. Comput. **219**(4), 1411–1419 (2012)
23. D. Jeong, S. Lee, Y. Choi, J. Kim, Energy-minimizing wavelengths of equilibrium states for diblock copolymers in the hex-cylinder phase. Curr. Appl. Phys. **15**(7), 799–804 (2015)
24. J. Kim, S. Lee, Y. Choi, S.M. Lee, D. Jeong, Basic principles and practical applications of the Cahn–Hilliard equation. Math. Probl. Eng. **2016** (2016)
25. P. Radu, D. Toundykov, J. Trageser, A nonlocal biharmonic operator and its connection with the classical analogue. Arch. Rational Mech. Anal. **223**(2), 845–880 (2017)
26. J.D. Rossi, C.B. Schönlieb, Nonlocal higher order evolution equations. Appl. Anal. **89**(6), 949–960 (2010)
27. D. Ruelle, *Statistical Mechanics: Rigorous Results* (World Scientific, 1999)
28. S.A. Silling, Reformulation of elasticity theory for discontinuities and long-range forces. J. Mech. Phys. Solids **48**(1), 175–209 (2000)
29. H. Sun, D. Uminsky, A.L. Bertozzi, Stability and clustering of self-similar solutions of aggregation equations. J. Math. Phys. **53**(11), 115610 (2012)
30. J.D. van der Waals, The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density. J. Stat. Phys. **20**(2), 200–244 (1979)
31. A.A. Wheeler, W.J. Boettinger, G.B. McFadden, Phase-field model for isothermal phase transitions in binary alloys. Phys. Rev. A **45**(10), 7424 (1992)
32. S.M. Wise, J.S. Lowengrub, H.B. Frieboes, V. Cristini, Three-dimensional multispecies nonlinear tumor growth—I: model and numerical method. J. Theor. Biol. **253**(3), 524–543 (2008)
33. M.A. Zaeem, H. El Kadiri, M.F. Horstemeyer, M. Khafizov, Z. Utegulov, Effects of internal stresses and intermediate phases on the coarsening of coherent precipitates: A phase-field study. Curr. Appl. Phys. **12**(2), 570–580 (2012)
34. S. Zhou, M.Y. Wang, Multimaterial structural topology optimization with a generalized Cahn–Hilliard model of multiphase transition. Struct. Multidiscip. Optim. **33**(2), 89 (2007)

3D Image-Based Stochastic Micro-structure Modelling of Foams for Simulating Elasticity

Anne Jung, Claudia Redenbach, Katja Schladitz, and Sarah Staub

1 Introduction

3D image data, predominantly generated by computed tomography, have been analyzed for more than 30 years now. In medical applications, the emphasis has been on segmentation and visualization. Algorithmic foundations had been settled in the 1990es, see [37] for a comprehensive overview.

Quantitative analysis of 3D images of micro-structures started in the 1990s, too, with porous rock samples [6, 59] and trabecular bone [18]. Prediction of macroscopic, in particular mechanical, properties based on the observed micro-structure has been a goal right from the beginning [64].

The seminal paper [29] provided Ohser's algorithm for efficiently estimating the intrinsic volume densities based on 3D binary images, linking fundamental characteristics of random closed sets [8, 15, 55] to those observable in image data.

Our contributions to predicting macroscopic materials properties by numerical simulation in realizations of 3D stochastic geometry models started with acoustic adsorption [52] and had been concerned with rigid foams early on [17, 48]. After establishing the mathematical [30] and algorithmic [31] bases for modelling cellular

A. Jung
Universität des Saarlandes, Applied Mechanics – Foams and Metamaterials, Saarbrücken, Germany
e-mail: anne.jung@mx.uni-saarland.de

C. Redenbach
Mathematics Department, Technische Universität Kaiserslautern, Kaiserslautern, Germany
e-mail: redenbach@mathematik.uni-kl.de

K. Schladitz (✉) · S. Staub
Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Kaiserslautern, Germany
e-mail: katja.schladitz@itwm.fraunhofer.de; sarah.staub@itwm.fraunhofer.de

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_11

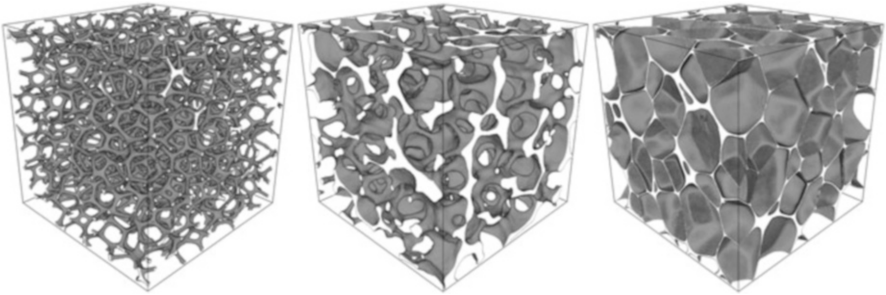


Fig. 1 Visualizations of 500^3 voxel sub-volumes from reconstructed CT images of rigid foams. From left to right: open polymer, partially closed ceramic, and closed polymer foams. All CT scans taken at ITWM with voxel sizes 49 μm for the open polymer foam, 34 μm for the ceramic foam, and 3 μm for the closed polymer foam. Samples provided by Vesuvius (ceramic) and Evonik (closed polymer foam)

material structures, elastic properties were studied [50]. Here, we will concentrate on characterization and modelling of rigid foams, too. Some typical examples of these materials are shown in Fig. 1.

Stochastic modelling allows for detailed studies of structure property relations [11, 34, 65]. Geometric features closely tied in practice can be modified individually. The use of stochastic instead of deterministic geometry models naturally captures microscopic heterogeneity as well as macroscopic homogeneity assuming invariances of the underlying distribution law. Moreover, the size of the representative volume element (RVE) can be determined statistically [19, 26]. In this work, we investigate RVE of constant size corresponding to the size of the foam samples characterized experimentally. Alternatively, more but smaller volume elements in the sense of stochastic homogenization as devised, e.g., in [1], could be used. In the case of rigid foams, the size of these volume elements is surely bounded by one foam cell. The theory of homogeneous random closed sets provides the corresponding mathematical concept by the typical cell of a tessellation [55]. Exploring this alternative approach and comparing the two are subject of future research. Fitting stochastic geometry models to the observed real structure is tedious as analytic relations are available only for models not suitable for real materials. Starting at the real micro-structure ensures however that observed trends and correlations have a practical meaning.

The concept of statistically similar representative volume elements (SSRVEs) [3] is similar to our approach in the sense of fitting the synthetic structure to the real one by minimizing the weighted sum of the squared differences of several geometric characteristics estimated for both real and synthetic structure. It differs however fundamentally in aiming at deterministic, geometrically significantly simplified structures right from the beginning. Simple structures consisting, e.g., of just a couple of spheres are used to generate the SSRVE that is subsequently periodically continued.

Here, we focus on numerical computation of effective elastic properties of an aluminum alloy foam by homogenization. The foam sample's micro-structure is observed in 3D image data obtained by micro-computed tomography (μCT). We use both the segmented image data and a random Laguerre tessellation model fit to the observed foam structure based on the estimated geometric characteristics. We provide the basics on random closed sets, their characteristics, in particular the densities of the intrinsic volumes, and estimating them based on 3D image data in Sect. 2. The image processing workflow for dividing the pore space of open foams into individual image objects is described in the same section. Section 3 is dedicated to random Laguerre tessellations and fitting them to the observed structure.

Simulation of elastic properties in voxel representations of the real foam or of model realizations is described in Sect. 4. A computational homogenization scheme [62] is used to transition from the microscopic scale of the foam structure to the macroscopic scale of the effective properties. A comprehensive overview on this class of homogenization schemes is given in [12]. A microscopic boundary value problem is formulated by imposing admissible boundary conditions and subsequently solved numerically, by a finite element method [10] or a fast Fourier transform (FFT) based solution of the Lippmann–Schwinger (LS) equations of elasticity [42]. Efficiency in terms of computational effort and memory use is critical as the images consist of several hundreds of voxels in each direction. The LS-FFT homogenization applied here is described in Sect. 4.

Finally, in Sect. 5, we apply the methods from Sects. 2–4 to the real aluminum foam sample. A random Laguerre tessellation is fit to the foam structure based on the characteristics estimated from the segmented CT image. Four synthetic foams are derived from its edge system by applying two cross-sectional shapes of the struts and relaxing the foam or not. Elastic properties derived by computation in the segmented μCT image and the structural modulus measured experimentally are compared in Sect. 5.3. The effective stiffnesses of the synthetic foams are numerically predicted in, too.

2 3D Image Analysis for Foams

We introduce the general concept of random closed sets, define basic characteristics for them, and describe how to estimate these characteristics based on 3D volume images as generated by computed tomography.

2.1 Random Closed Sets and Their Characteristics

Material micro-structures are often macroscopically homogeneous in some sense but locally heterogeneous. A commonly used mathematical model for such structures is random closed sets, random variables whose realizations are closed subsets of $\mathbb{R}^3$. Schneider and Weil [55] attribute the concept to Matheron [39] and Kendall [27], with the first detailed description being [40].

Denote by $\mathcal{F}$ the system of closed subsets of $\mathbb{R}^3$ and by $\mathfrak{F}$ the hit-or-miss σ-algebra generated by the sets $\{F \in \mathcal{F} : F \cap A \neq \emptyset\}$ for all compact $A \subset \mathbb{R}^3$. The pair $(\mathcal{F}, \mathfrak{F})$ is then a measurable space, and we can define random closed sets (RACS) as random variables with values in this space. More precisely, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A random closed set Ξ is a measurable mapping

$$\Xi : (\Omega, \mathcal{A}, \mathbb{P}) \mapsto (\mathcal{F}, \mathfrak{F}).$$

We call Ξ stationary (or macroscopically homogeneous) and isotropic if its distribution is invariant with respect to translations and rotations, respectively.

Point Processes One class of RACS models is locally finite random closed sets: Denote by C the system of compact subsets of $\mathbb{R}^3$. A set $A \subseteq \mathbb{R}^3$ is locally finite if $\#(A \cap C) < \infty$ for any $C \in C$. Here, $\#B$ refers to the number of elements of the set B. A (simple) point process Φ is a random variable with values in the system of locally finite sets and can be read as a sequence of random points in $\mathbb{R}^3$ as well as a random counting measure. The measure Λ defined by $\Lambda(B) = \mathbb{E}(\#(\Phi \cap B))$ is called the intensity measure of Φ. For stationary point processes, it is a multiple of the Lebesgue measure $\Lambda(B) = \lambda V(B)$ with the intensity $\lambda > 0$ giving the mean number of points per unit volume.

The Poisson Point Process The Poisson point process representing complete spatial randomness plays a particular role as it is analytically tractable and the basis for constructing other point process as well as a variety of random closed set models. A Poisson point process Φ in $\mathbb{R}^3$ with intensity measure Λ is a point process with the number of points $\#(\Phi \cap B)$ in a Borel set B being Poisson distributed with parameter $\Lambda(B)$ and the numbers of points $\#(\Phi \cap B_1), \ldots, \#(\Phi \cap B_n)$ in pairwise disjoint Borel sets $B_1, \ldots, B_n$ being independent random variables.

As a consequence, a stationary Poisson point process with intensity λ in a compact observation window $W \subset \mathbb{R}^3$ can easily be simulated by first drawing the number of points Poisson distributed with parameter $\lambda V(W)$. Then, the points are placed independently identically uniformly distributed in W. There is no interaction of the points. This is why the Poisson point process is used as reference for complete spatial randomness.

The random points can additionally be decorated by marks representing, for instance, a species, age, or size. Moreover, the definitions of point process and Poisson point process can directly be transferred any locally compact space with countable base.

Intrinsic Volumes The intrinsic volumes are basic characteristics of RACS. Let $\mathcal{K}$ denote the system of compact and convex sets (convex bodies) in $\mathbb{R}^3$. Dilating $K \in \mathcal{K}$ with a ball of radius r yields the so-called parallel set $K \oplus b(0, r) := \{x + y : x \in K, y \in b(0, r)\}$. Steiner's formula expresses the volume of this set as a polynomial in r:

$$V(K \oplus b(0,r)) = \sum_{k=0}^{3} \kappa_k V_{3-k}(K) r^k, \quad r \geq 0,$$

where κ_k is the volume of the k-dimensional unit ball and the coefficients $V_0, \ldots, V_3$ are the four intrinsic volumes in $\mathbb{R}^3$. These are—up to constant factors—the volume $V = V_3$, the surface area $S = 2V_2$, the integral of mean curvature $M = \pi V_1$, and the Euler number $\chi = V_0$, see, e.g., [54, p. 210]. For $K \in \mathcal{K}$, the integral of mean curvature is up to a constant the mean width $M = 2\pi \bar{b}$—the distance of two parallel planes enclosing K, averaged w.r.t. rotation.

The intrinsic volumes form a basis of geometric characteristics in the sense of Hadwiger's theorem [55, Theorem 14.4.6]: every rigid motion invariant, additive, and continuous functional on $\mathcal{K}$ is a linear combination of the intrinsic volumes. Additivity yields a straightforward extension of the intrinsic volumes to finite unions of convex bodies, so-called polyconvex sets.

Densities of the Intrinsic Volumes The intrinsic volumes can be used to characterize single cells of a foam. For characterizing the complete solid component, we need another generalization. It applies for RACS Ξ whose intersections with convex, compact observation windows $W \in \mathcal{K}$ are almost surely polyconvex. In this case, the intrinsic volumes of $\Xi \cap W$ are well defined, too. Denote by $\mathbb{E}V_k(\Xi \cap W)$ the expected k-volumes of this intersection with respect to the distribution of Ξ.

The *densities of the intrinsic volumes* are the limits of these expectations for growing observation window:

$$V_{V,k}(\Xi) = \lim_{r \to \infty} \frac{\mathbb{E}V_k(\Xi \cap rW)}{V(rW)}, \qquad k = 0, \ldots, 3. \tag{1}$$

The limit exists if $\mathbb{E}2^{N(\Xi \cap [0,1]^3)} < \infty$ [55, Theorem 9.2.1], where $N(X)$ is the smallest $m \in \mathbb{N}$ such that $X = K_1 \cup \ldots \cup K_m$ with $K_1, \ldots, K_m \in \mathcal{K}$.

Of particular interest for foams are the (solid) volume fraction $V_V = V_{V,3}$, with $p = 1 - V_V$ being the porosity, and the specific surface area $S_V = 2V_{V,2}$. Additionally, the density of the Euler characteristic $\chi_V = V_{V,0}$ is related to the mean number of nodes of the foam skeleton per unit volume [53].

2.2 Image Analysis

We quickly define the basic concepts image and adjacency system, discuss the segmentation tasks posed by our problem, and finally describe how to estimate the characteristics needed for model fitting later on.

Images A typical way of observing a realization X of a suitable RACS Ξ is via discretization in a 3D image. For simplicity, let $\mathbb{L} = s\mathbb{Z}^3$ be a three-dimensional cubic lattice with lattice spacing $s > 0$. Denote by $W \subset \mathbb{R}^3$ a cuboidal observation window. By an image, we understand a function

$$f : \mathbb{L} \cap W \longrightarrow V,$$

where V is the set of real or complex numbers $\mathbb{R}$, $\mathbb{C}$, or $V = \{0, \dots, 2^n - 1\}$ with $n = 1, 8, 16$, or 32.

Black-and-white images f with $V = \{0, 1\}$ are often called binary image. In this case, the image foreground can be identified with the intersection $X \cap \mathbb{L} \cap W$ of a set $X \subset \mathbb{R}^3$ observed at the lattice points in the observation window W. The function f is then the indicator function of X, restricted to the observable points $f = \mathbb{I}_X|_{\mathbb{L}\cap W}$. That is, $f(x) = 1$ if $x \in X \cap \mathbb{L} \cap W$ and $f(x) = 0$ otherwise. Usually, in image processing, both $x \in \mathbb{L} \cap W$ and the pair $(x, f(x))$ are called pixel (or voxel).

Mathematically, we interpret an image as a set of lattice points equipped with gray values, not as a set of small cubes. We nevertheless address the lattice spacing s as pixel or voxel size, too.

By construction, the pixels of a 3D image are discrete. To represent concepts such as the connected components of a RACS in an image, a notion of connectivity of pixels must be introduced. To define the discrete connectivity in 3D unambiguously, we follow [45] in using adjacency systems consisting of j-dimensional faces with $j = 0, \dots, 3$. The *discretization* $X \sqcap \mathbb{F}$ of a compact subset $X \subset \mathbb{R}^3$ w.r.t. a given adjacency system $\mathbb{F}$ is defined as the union of the elements F of $\mathbb{F}$ whose vertices $\mathcal{F}^0(F)$ are contained in X

$$X \sqcap \mathbb{F} = \bigcup \{F \in \mathbb{F} : \mathcal{F}^0(F) \subseteq X\}.$$

It can be interpreted as an approximation of X by a polyhedral set built using the bricks provided by the adjacency system. For background on adjacency systems and consistent connectivity for foreground and background of an image, see [43–46, 51].

Segmentation Segmentation subsumes a wide variety of methods assigning pixel values to classes, usually coded by integer values. Here, we have to solve two segmentation tasks: first, we have to binarize the image f to get hold of the solid foam structure. Second, we exploit moments of the distributions of cell characteristics in the model fitting step. Hence, we have to identify foam cells or pores as image objects. This second segmentation task is much more demanding, as the cells of an open foam are connected. As a consequence, the background or pore space has to be divided to "reconstruct" the cell structure.

Binarization Very roughly, computed tomography generates 3D images with voxel gray values related to the mean X-ray absorption of the small sub-volume they represent. Solid material structure thus appears brighter than air. For binarizing the image, we look for a transformation

$$T(f) : \mathbb{L} \cap W \longrightarrow \{0, 1\}$$

such that $T(f) \equiv 1$ corresponds to the solid foam material X observed in $\mathbb{L} \cap W$ as described above. The simplest way to achieve a binarization is just to use a global gray value threshold $\theta \in \mathbb{R}$

$$T(f)(x) = \begin{cases} 1 & f(x) \geq \theta \\ 0 & \text{else.} \end{cases}$$

This simple binarization method can of course be applied only, if the image is free of global gray value fluctuations. Prior to binarization, we apply a median filter with a $3 \times 3 \times 3$ filter mask to reduce noise and smooth the foam's surface slightly. Subsequently, the global gray value threshold θ can be chosen by Otsu's method [47].

Cell Reconstruction Now, we separate the multiply connected pore system of the foam into individual image objects by a combination of two strong morphological tools, the Euclidean distance transform and the watershed transform.

Let $X \subseteq \mathbb{R}^3$ be the set under consideration, here the pore space of the foam. The Euclidean distance transform maps each point in $\mathbb{R}^3$ to its shortest distance to the complementary set $\mathbb{R}^3 \setminus X$,

$$\text{EDT} : \mathbb{R}^3 \mapsto [0, \infty) : x \mapsto \min\{||x - y|| : y \in \mathbb{R}^3 \setminus X\}.$$

This results in $\text{EDT}(y) = 0$ for all $y \in \mathbb{R}^3 \setminus X$ and local maxima in the centers of spherical regions in X. Inverting the EDT image $\text{EDT}(X \cap \mathbb{L} \cap W)$ turns these local maxima into minima: $f(x) = \max\{\text{EDT}(X \cap \mathbb{L} \cap W)\} - \text{EDT}(x)$. The exact Euclidean distance transforms can be calculated very efficiently, i.e., in linear time, exploiting the Voronoi paradigm [41].

Now the watershed transform assigns a connected region to each local minimum. This transform can be interpreted as the flooding of the topographic surface $\{(x, f(x)) : x \in \mathbb{L} \cap W\}$: water rises uniformly with growing gray value f from all local minima. Watersheds are formed by all pixels where water basins filled from different sources meet. Finally, the image is segmented into regions and the system of watersheds dividing them [66].

In practice, this morphological separation strategy suffers from superfluous local minima in the inverted EDT (iEDT) image due to discretization, imperfect binarization, and cell shapes deviating from spherical. The watershed transform assigns each such minimum a region, inevitably resulting in a strong over-segmentation. A remedy for this is to prevent small regions from arising at all by altering the flooding accordingly (pre-flooded watershed [63]). The foam structure under consideration here is rather regular. Predescribing a minimal cell volume is thus easily applicable. See Fig. 2 for an illustration using 2D sections through the 3D images.

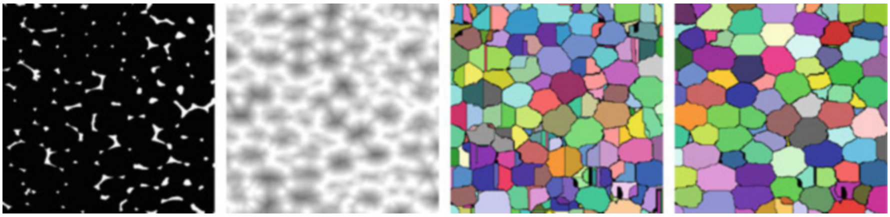


Fig. 2 Morphological cell reconstruction, illustrated using 2D slices of the 3D image. Left: Binarized—solid foam structure appearing white pore space black. Center left: inverted Euclidean distance map on the pore space, small values dark, high white. Center right: pore system generated by the watershed transform. Right: pore system generated by the pre-flooded watershed transform

> It is remarkable that the heavyweights in this segmentation procedure—watershed transform and EDT—as devised by Vincent and Soille [66] and Maurer and Raghavan [41] work "as is" in arbitrary dimensions. Often, image processing methods are mathematically easily formulated for arbitrary dimensions and applied to 2D images. Practical application in higher dimensions is nevertheless often impossible due to computational complexity or ambiguities, e.g., arising from the existence of two subdimensions or connectivity issues.

Estimating the Intrinsic Volumes Based on Image Data The intrinsic volumes and their densities can be estimated based on 3D image data very efficiently by Ohser's algorithm [29]. This algorithm codes the $2 \times 2 \times 2$ voxel configurations in a binary image using a convolution with a cubic filter mask assigning weights $2^k,\ k = 0, \ldots, 7$, to the vertices of the cube. The gray value histogram of the resulting 8-bit gray value image contains all information needed to derive the intrinsic volumes by multiplication with a suitable weights vector. See [44] for details.

3 Random Laguerre Tessellations and Fitting Them

Laguerre tessellations are a generalization of the well-known Voronoi tessellation model. When generated by random sphere packings, they reflect the topology of rigid foams very well and enable particularly good control over the volume distribution of the resulting cells. We recall the concept and a fitting strategy.

3.1 Laguerre Tessellations Generated by Random Sphere Packings

Let T be a set of bounded convex three-dimensional subsets of $\mathbb{R}^3$, the cells of T. The system T is called a tessellation of $\mathbb{R}^3$ if T is space-filling, i.e., $\bigcup_{C\in T} C = \mathbb{R}^3$, and if the interiors of different cells do not intersect. Convexity of the cells and T being space-filling force the cells to be three-dimensional polytopes (3-polytopes) [55, Lemma 10.1.1].

Write

$$F(x) = \bigcap_{C\in T, x\in C} C, \quad x \in \mathbb{R}^3$$

for the intersection of all cells of T containing the point x. Then, $F(x)$ is a non-empty finite intersection of 3-polytopes and hence a k-polytope, $k \in \{0, \dots, 3\}$. The set

$$\Delta^k(T) := \{F(x) : \dim F(x) = k, \quad x \in \mathbb{R}^3\}, \quad k = 0, \dots, 3,$$

is the set of k-faces of the tessellation T. We denote by $\mathcal{F}^k(C)$ the set of all k-faces, $k = 0, \dots, 3$, of a 3-polytope C. For a set $\mathbb{F}$ of convex polytopes, write $\mathcal{F}^k(\mathbb{F}) = \bigcup\{\mathcal{F}^k(F) : F \in \mathbb{F}\}$.

The tessellation T is called face to face if the faces of the tessellation coincide with the faces of the cells. If additionally exactly four cells meet in a vertex and three cells meet in an edge, the tessellation is called normal. A random tessellation is face to face or normal if its realizations almost surely have these properties.

Due to Plateau's laws, soap froths are normal, too [67]. Normal tessellations are thus a natural choice for modelling foam structures. The best known and most used tessellation model is the Voronoi tessellation that can be defined in arbitrary dimension n based on a locally finite set $\varphi \subset \mathbb{R}^n$ of generators. The Voronoi tessellation of φ consists of the cells

$$C\big(x\big) := \{z \in \mathbb{R}^n \ : \ ||x - z|| \le ||y - z||, \ \text{for all } y \in \varphi\}, \ x \in \varphi.$$

The size of a Voronoi cell thus only depends on the distance of its generator to the neighboring generators.

To gain more flexibility regarding cell shapes and in particular sizes, we use Laguerre tessellations. In this weighted generalization of the Voronoi tessellation, each generating point $x \in \mathbb{R}^n$ is assigned a positive weight $r > 0$ that can be interpreted as radius of a sphere with center x. The Laguerre cell of $(x, r) \in \varphi$ is defined as

$$C\big((x, r), \varphi\big) := \{z \in \mathbb{R}^3 \ : \ ||x - z||^2 - r^2 \le ||y - z||^2 - s^2, \ \text{ for all } (y, s) \in \varphi\}.$$

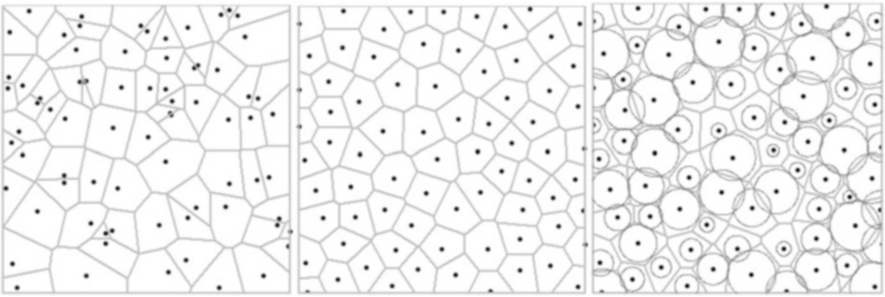


Fig. 3 Left and middle: realizations of Voronoi tessellations generated by a Poisson point process and a regular point process. Right: Laguerre tessellation generated by the same point pattern as in the middle

The Laguerre tessellation $L(\varphi)$ is the set of the non-empty Laguerre cells of φ. The Voronoi tessellation corresponds to the special case of the Laguerre tessellation generated by a system of spheres of constant radius.

Laguerre tessellations are the most general model satisfying our assumptions: each normal tessellation of $\mathbb{R}^3$ with convex cells is a Laguerre tessellation [2, 32].

In Laguerre tessellations, empty cells as well as cells not containing their generators can occur. Laguerre tessellations generated by nonoverlapping spheres are however free of such irregularities, see Fig. 3 for examples in $\mathbb{R}^2$.

Random systems of closely packed spheres like those generated by the force-biased algorithm [4, 5] are well suited to generate Laguerre tessellations with prescribed cell volume distribution [49]. The cell shape is however rather restricted. In particular, the edge length distributions in Laguerre tessellations generated by dense packings of spheres are known to differ significantly from those observed in real foams, see [65]. This problem can be alleviated by relaxing the structure using the Surface Evolver [7].

Voronoi or Laguerre tessellations generated by stationary and isotropic (marked) point processes are also isotropic. In contrast, real foam structures often show anisotropies as their cells are elongated in particular directions due to the foam generation process. A simple way of incorporating this anisotropy into the models is by appropriate scaling of the cell systems.

3.2 *Fitting a Tessellation Model*

Two strategies of fitting Laguerre tessellation models can be found in the literature. Laguerre approximation aims at finding a Laguerre tessellation that represents the observed cell system best in the sense that discrepancy between the observed cells and the cells of the approximation is minimized [33, 60]. Alternatively, a parametric tessellation model can be fit. That is, the observed cell system is approximated in a stochastic sense. The model is supposed to fit distributions of cell characteristics such as the volume, surface area, or the number of facets [31, 49]. The former approach yields an exact representation of the observed structure in the class of Laguerre tessellations. Here, we concentrate on the latter approach, as it allows for generating an arbitrary number of model realizations of basically arbitrary size.

The deviation of the model realization from the real observed foam structure is measured using the relative distance measure

$$\rho(\hat{m}, m) = \sqrt{\sum_{i=1}^{n} \left(\frac{m_i - \hat{m}_i}{\hat{m}_i}\right)^2}, \tag{2}$$

where $\hat{m} = (\hat{m}_1, \ldots, \hat{m}_n)$ and $m = (m_1, \ldots, m_n)$ are moments of the distributions of geometric characteristics of the cells of the original foam and the model, respectively. For rigid foams whose structure is observed in 3D images, the means and standard deviations of the volume $V = V_3$, the surface area $S = 2V_2$, the mean width $\bar{b} = \frac{1}{2}V_1$, and the number of facets F_C of the cells have proven to be particularly well suited [31, 49]. The choice is based on mean value relations for normal random tessellations, see [8, Section 9.4].

For minimizing the distance measure (2) on the parameter space of a tessellation model, one would ideally use analytic formulas relating the model parameters and the required moments of cell characteristics. Unfortunately, only Laguerre tessellations generated by a homogeneous Poisson process are analytically tractable [32]. For Laguerre tessellations generated by random sphere packings, model fitting therefore has to rely on Monte Carlo simulations of model realizations.

The cell volumes in cellular materials are usually assumed to be lognormal or gamma distributed. Moreover, Laguerre tessellations of dense packings of spheres with lognormal and gamma distributed volumes are very regular and therefore well suited to model rigid foams. In [49], model realizations for various packing fractions κ and coefficients of variation c of the volume distribution were generated. Subsequently, polynomials in c were fitted to the estimated geometric characteristics for each value of κ. Using these results, the minimization of $\rho(\hat{m}, m)$ in (2) reduces to minimizing a polynomial, thus allowing for quick and easy model fit.

4 Numerical Simulation of Elastic Properties

The following subsections outline the simulation of the effective mechanical stiffness of inhomogeneous micro-structures such as foams. Therefore, the averaged stresses and strains are defined as volume averages over their microscopic counterparts. Furthermore, the fundamentals of the solution of the microscopic boundary value problem in terms of an LS-FFT are outlined briefly.

4.1 Effective Properties of Micro-Structured Materials

We apply a computational homogenization scheme to determine the effective stiffness of the foam numerically. In this scheme, the microscopic geometry of the structure is captured either by a CT image of the foam or by a realization of a stochastic geometry model. Furthermore, the mechanical properties of the micro-constituents, here Young's modulus and Poisson's ratio, are required.

The micro-structure is captured in volume element $\Omega \subseteq \mathbb{R}^3$ with volume $V(\Omega)$ and boundary $\partial\Omega$. The effective stiffness of this structure $\mathbb{C}^*$ connects the averaged stresses $\langle\sigma\rangle$ and strains $\langle\varepsilon\rangle$ via

$$\langle\sigma\rangle = \mathbb{C}^* : \langle\varepsilon\rangle. \tag{3}$$

The : in Eq. (3) denotes the double inner product between the fourth-order tensor $\mathbb{C}^*$ and the second-order tensor $\langle\varepsilon\rangle$ and corresponds to the mapping of the strains to the stresses (compare Hooke's law in one-dimensional elasticity $\sigma = E\,\varepsilon$). The averaged stresses and strains are defined as volume averages of their microscopic counterparts as

$$\langle\sigma\rangle := \frac{1}{V(\Omega)} \int\limits_{\Omega} \sigma(x)\,\mathrm{d}x \quad \text{and} \quad \langle\varepsilon\rangle := \frac{1}{V(\Omega)} \int\limits_{\Omega} \varepsilon(x)\,\mathrm{d}x. \tag{4}$$

For the numerical computation of the effective foam properties, a displacement, which corresponds to a constant strain in a homogeneous reference material, is applied to the boundary of the volume element. Due to the possible anisotropy of the foam, six load cases are computed—three tension, three shear as sketched in Fig. 4.

After solving the six microscopic boundary value problems, we compute the complete effective stiffness tensor $\mathbb{C}^*$.

4.2 Lippmann–Schwinger Fast Fourier Transform-Based Solver

In the following, the details on the solution of the boundary value problem are outlined. Our microscopic simulation is based on the solution of the Lippmann–Schwinger (LS) equations for elasticity, see [36] and [68]. The required Green's operator is explicitly known in the Fourier space. Thus, these equations are solved efficiently by application of a fast Fourier transform (FFT). This LS-FFT scheme is implemented in ITWM's micro-structure solver FeelMath, see [22] and [24], also available as module ElastoDict [9] in the commercial software GeoDict [13]. The LS-FFT solver enables precise computation of microscopic stresses and strains directly in µCT images or other voxel-based three-dimensional structures. It is

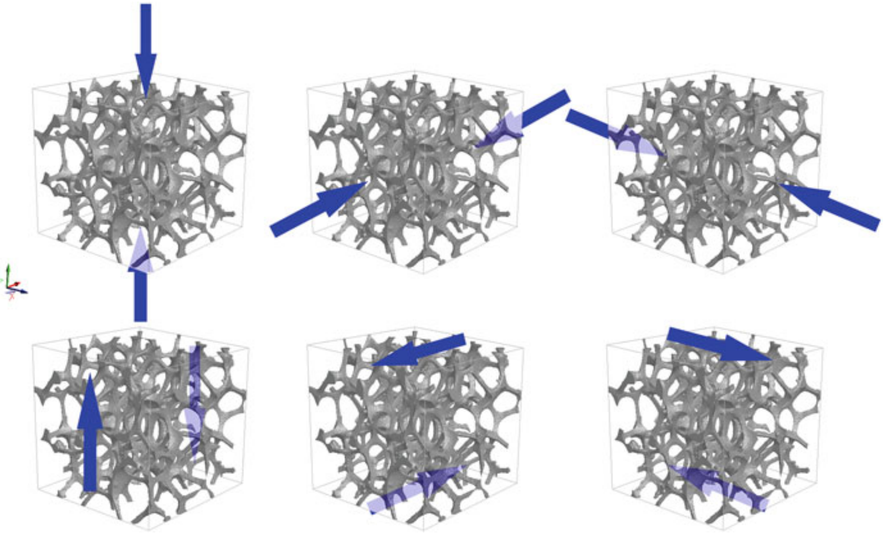


Fig. 4 Application of boundary conditions for homogenization. Blue arrows indicate the forces applied

applicable to porous structures and therefore suitable for the open aluminum foam considered here.

In the solver, the equilibrium equation of the Cauchy stress σ

$$\operatorname{div} \sigma(x) = 0, \quad x \in \Omega, \tag{5}$$

is considered in the micro-domain Ω. The kinematics for the strains ε depending on the displacements u and the fluctuations v read

$$\left.\begin{aligned} \varepsilon(u)(x) &= \langle \varepsilon \rangle + \varepsilon(v)(x) \\ \varepsilon(v)(x) &= \tfrac{1}{2}\left(\operatorname{grad} v(x) + \operatorname{grad}^t v(x)\right) \end{aligned}\right\} \quad x \in \Omega \tag{6}$$

in terms of the applied macroscopic strain ε. At the boundary of the micro-structure $\partial\Omega$ (anti-)periodic boundary conditions are applied via

$$\left.\begin{array}{ll} v(x) & \# \\ \sigma(x) \cdot n(x) & -\# \end{array}\right\} \quad x \in \partial\Omega. \tag{7}$$

The symbol # refers to periodicity, i.e., the fluctuations at opposite faces of the boundary are equal, whereas $-\#$ denotes antiperiodicity, i.e., the tractions $\sigma \cdot n$ at opposite faces point into opposite directions but have the same magnitude. The set of underlying equations is completed by a constitutive equation for the microscopic constituents. For the aluminum, we restrict ourselves to the linear elastic case. Thus, the microscopic stresses and strains are connected by an elasticity tensor $\mathbb{C}$ via

$$\sigma(x) = \mathbb{C}(x) : \varepsilon(x), \tag{8}$$

which only depends on Young's modulus E and Poisson's ratio ν of the aluminum alloy. In the pores, the stiffness is set to zero. In general, the presented approach is however suitable to capture more complex material behavior like inelasticity and rate dependency, see [25, 56, 61].

Next, we focus on reformulating the periodic boundary value problem into an integral expression of the Lippmann–Schwinger type. To this end, we introduce a constant homogeneous reference stiffness tensor $\mathbb{C}^0$ instead of the stiffnesses of the aluminum and the pores. This homogeneous stiffness tensor is applied to define the polarization tensor τ as

$$\tau(x) = \sigma(x) - \mathbb{C}^0 : \varepsilon(x). \tag{9}$$

With the help of Green's operator Γ^0 associated with the reference stiffness, the solution of the equilibrium equation (3) reads

$$\varepsilon(x) = \langle\varepsilon\rangle - \left(\Gamma^0 * \tau\right)(x). \tag{10}$$

The convolution operator $*$ is defined by

$$\left(\Gamma^0 * \tau\right)(x) = \int\limits_{\Omega} \Gamma^0(x - y) : \tau(y)\,\mathrm{d}y. \tag{11}$$

Combining the constitutive law (6), the definition of the polarization stress (9), and the solution (10) yields the LS equation as

$$\varepsilon = \varepsilon(x) + \Gamma^0(x) * \left((C(x) - \mathbb{C}^0) : \varepsilon(x)\right) = (I + B_\varepsilon(x))\,\varepsilon(x). \tag{12}$$

Green's operator Γ^0 does not depend on the fluctuations and thus only depends on the homogeneous reference stiffness $\mathbb{C}^0$, see [28].

The LS equation (12) can be solved iteratively using the Neumann series expansion or by using the conjugate gradient method. As Green's operator is explicitly known in Fourier space, the Fourier transform is applied. A more detailed description of the algorithm can be found in [61].

Note that the discretization by Fourier polynomials as presented in [42] leads to convergence problems for porous structures due to the infinite stiffness contrast of the microscopic constituents. Therefore, we use a finite difference discretization based on a staggered grid, which converges also for highly porous materials, see [57]. A comprehensive overview of FFT-based homogenization methods is given in [56].

5 Application Example

Now, all methods described above are applied to a real-world open aluminum foam sample.

5.1 Material

We consider one of the open-cell aluminum alloy foam samples investigated in [21]. The sample made by CellTec Materials GmbH, Dresden, Germany consists of $AlSi_7Mg_{0.3}$ and has nominal pore size 10 ppi (pores per inch) and mean density 0.156 g/cm^3 corresponding to a porosity of 94.2%.

A cubic sample of edge length 40 mm is spatially imaged by μCT at voxel size 29.44 μm. See Fig. 5 for a volume rendering.

Larger samples of size 40 mm × 40 mm × 80 mm were tested mechanically. Uniaxial compression and tensile tests were performed, where the samples had to be infiltrated by a resin to allow for clamping for the tensile tests. The details are described in [20]. In [21], the foam structure was image analytically separated into vertices and struts. Mechanical behavior of struts was investigated for five classes divided according to orientation.

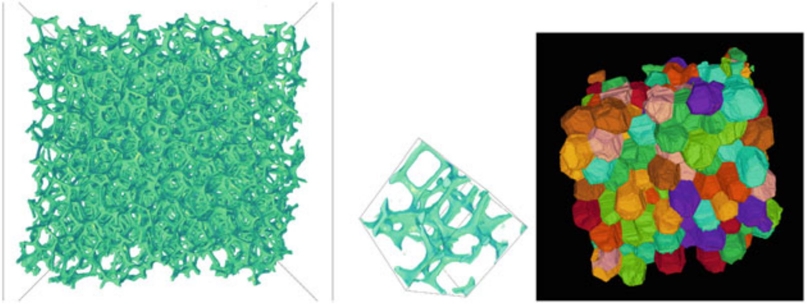


Fig. 5 Left: volume rendering of the reconstructed CT image. The CT image taken at ITWM has originally a voxel size of 29.9 μm and the sample is contained in a cube of edge length 1500 voxels corresponding to 4.5 cm. Center: sub-volume of edge length 0.9 cm with blob like production leftover. Right: system of reconstructed cells

Table 1 Estimated mean values and standard deviations of the cell characteristics of the aluminum foam and the best fit models for lognormally distributed volumes of the generating spheres

		Scaled data		Isotropic model			
		Mean	Std	Mean	Deviation	Std	Deviation
v	[mm^3]	79.738	21.972	79.739	±0.00%	26.598	±21.06%
s	[mm^2]	104.468	22.241	98.608	−5.61%	19.832	−10.83%
d	[mm]	5.811	0.740	5.806	−0.08%	0.568	−23.22%
F_C		13.828	2.256	14.06	+1.69%	2.173	−3.71%

5.2 Image Analysis and Model Fit

The CT image of the foam was binarized by a global gray value threshold chosen according to Otsu's method. Subsequently, the cells were reconstructed by the watershed on iEDT approach described in Sect. 2.2. Due to contamination of the foam structure (see Fig. 5, middle), some cells were erroneously split into two parts. These errors were corrected manually.

Estimation of the mean cell diameters in the three coordinate directions reveals that cells are elongated by a factor $s = 1.2526$ in y-direction. The model fitting procedure detailed in Sect. 3.2 is formulated for isotropic structures. Hence, the structure is scaled by $1/s$ prior to the fitting. The cell characteristics estimated from the isotropic structure are listed in Table 1. Cells intersecting the boundary of the image are not observed completely. Hence, they are not included in the statistics. As larger cells are more likely to intersect the image boundary, ignoring boundary cells will result in a sampling bias. This is corrected by weighting the cells according to the inverse of their probability of being observed (Miles–Lantuejoul correction, [58, pp. 246]). In total, 336 cells with a total volume of 26246.11mm^3 were included in the statistics.

The model fitting procedure returns a packing fraction of $\kappa = 60\%$ and a coefficient of variation $c = 0.414$ as optimal parameters when assuming a lognormal distribution of sphere volumes. Using these parameters, realizations of the tessellation models containing 820 cells are simulated in the unit cube. After rescaling the y-axis by the factor s, the tessellation edges are discretized into volume images. The voxel spacing for the discretization is chosen identical to the CT image, i.e., 29.44 µm. The images are cropped to a size of 1359^3 voxels to obtain the original sample edge length of 40 mm.

Dilation of the edge system of the tessellations finally yields synthetic foam structures. Strut thickness is chosen such that the volume fraction is close to the value $V_V = 5.755\%$ estimated from the binarized CT image of the real foam. We consider two cross-sectional shapes of the struts—the simplest case of perfectly circular strut cross sections of constant radius and concave triangular struts. Additionally, to investigate the effect of relaxation, the tessellations are relaxed by using the Surface Evolver and also discretized with the two choices of strut cross-sectional shape. That way, each realization of the Laguerre tessellation

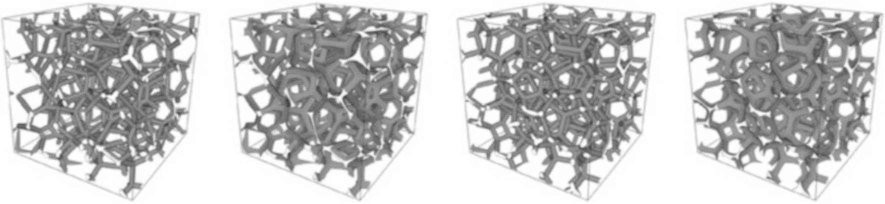


Fig. 6 Volume renderings of four synthetic foams derived from the same Laguerre tessellation model realization. From left to right: circular and triangular strut cross sections and their relaxed versions. Sub-volumes of 600^3 voxels

model yields four synthetic foam structures. All four synthetic foams are visualized in Fig. 6.

5.3 Prediction of Mechanical Properties

We now predict the mechanical properties of the aluminum foam based on the binarized CT image and on realizations of the four synthetic foams.

Validation of Simulation Before actually predicting, we validate the numerical computation of the effective quantities with ElastoDict [9] by comparing with experimental data. To this end, a compression load case with a prescribed strain is simulated directly on the binarized CT image of the foam. The resulting effective stresses and the corresponding stiffness in the loading direction are calculated and compared to those determined by a compression test on the foam as described in [20].

First, we summarize the solver settings: the considered foam structure is not periodic. Hence, symmetric boundary conditions with free deformations in tangential directions are chosen. These symmetric boundary conditions and their compatibility with the periodic boundary conditions from Sect. 4.2 are described in detail in [14]. Thanks to the compatibility, this type of boundary condition is often referred to as periodicity compatible mixed uniform boundary conditions (PCMUBC). Generating a periodic structure by means of mirroring and effective implementation of PCMUBC in the context of FFT are connected in [14], too.

We apply composite voxels to reduce computation time as suggested in [23]. In this approach, the image is downsampled and the stiffness of the resulting mixed voxels or composite voxels is determined with the appropriate material of the rotated laminate, i.e., not only the local volume fraction is taken into account. Additionally, this mixing rule incorporates directional information on the material interface. Applying this strategy, we cut 1314 × 1335 × 1285 voxels down to 328 × 333 × 321 voxels and consequently reduce CPU time and memory consumption significantly while keeping the loss of accuracy acceptable.

Table 2 Experimentally identified Young's moduli for the aluminum alloy [16]

Pore no.	Young's modulus [MPa]
1	4112
2	3832
3	3768
4	4152
5	3978

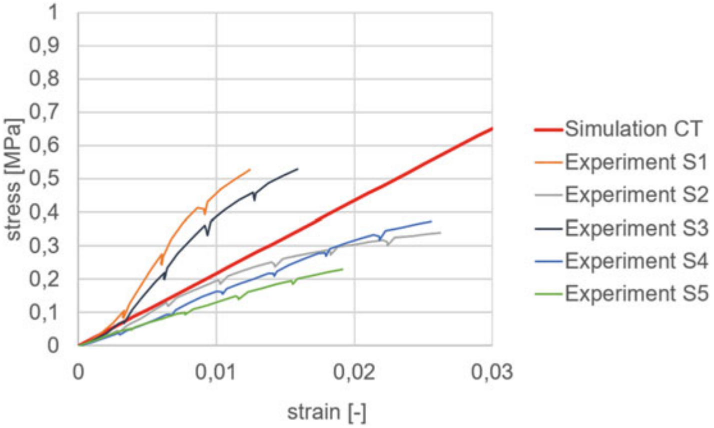


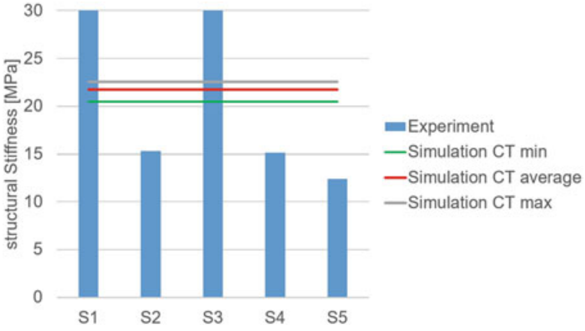
Fig. 7 Comparison of simulation and experiment in longitudinal direction y in terms of stress–strain diagram, Young's modulus of the aluminum $E_{Al} = 4000\,\text{MPa}$

Second, we care for the material parameters: We restrict ourselves to the linear elastic case here. Therefore only Young's modulus E_{Al} and Poisson's ratio ν_{Al} of the aluminum at the micro-scale are required. Poisson's ratio is taken from literature as $\nu_{Al} = 0.33$. Young's modulus of the aluminum is chosen based on the experimental data from [16] measured in five different pores. The values are summarized in Table 2. In the simulations, the average of these moduli (4000 MPa) and the minimal and maximal values are considered. Attention should be paid to the fact that these values differ significantly from those usually reported for aluminum alloys ($\approx 70 \cdot 10^3$ MPa). These deviations are due to micro-porosity of the struts and inclusions resulting from the manufacturing process [38].

Figure 7 displays the experimental stress–strain curves of the five foam samples under compression in the linear loading regime, see [20] for the complete loading and unloading curves and the effective stress–strain curves simulated directly in the complete binarized CT image. Simulated and experimental data match very well. Young's modulus of the aluminum alloy is chosen as the average of the five samples summarized in Table 2.

Figure 8 shows how the microscopic Young's modulus of the aluminum alloy influences the simulation result in terms of the effective modulus, i.e., the slope of the effective stress–strain curve. As input for the simulation, the minimal, maximal,

Fig. 8 Comparison of simulation and experiment in longitudinal direction y in terms of structural stiffness for varying Young's moduli of the aluminum alloy



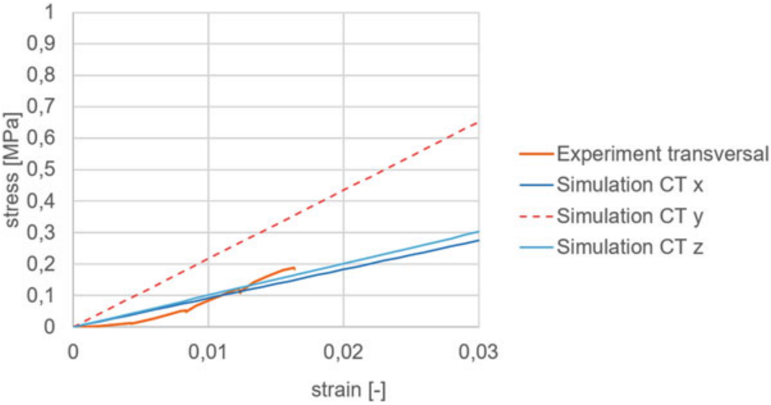


Fig. 9 Comparison of simulation and experiment in transversal direction, Young's modulus of the aluminum $E_{Al} = 4000$ MPa

and average values from Table 2 are considered. These simulation results are compared to the structural stiffness obtained by the compression experiments.

The structural stiffness corresponds to the slope of the experimental stress–strain curves in Fig. 7. Clearly, variations of the microscopic Young's modulus in the considered range have only minor influence on the effective modulus. Therefore, we take only the average value of 4000 MPa into account in the following.

Validation of the simulation is completed by a comparison of the effective stresses simulated in the transversal directions (here x and z) and a compression experiment in the corresponding directions in Fig. 9. The simulated effective stresses in x- and z-directions are almost the same and they are much smaller than those in y-direction. Thus, the foam is much stiffer in the longitudinal y-direction. Figure 9 also shows that simulation and experiment also fit very well in the transversal directions.

In the next section, the validated simulation is applied to the synthetic foams derived from realizations of the Laguerre tessellation model fit in Sect. 5.2.

Effective Mechanical Properties of the Synthetic Foams In order to predict the properties of the synthetic foams, five realizations of the Laguerre tessellation model yielding altogether 20 synthetic foams with circular or concave triangular struts, in relaxed state or not, are considered. All load cases displayed in Fig. 4 are applied to each synthetic foam and the full effective stiffness tensor is computed. The effective Young's modulus in each direction (E_x, E_y and E_Z) is approximated orthotropically, i.e., it is assumed that the inverse effective stiffness tensor reads

$$\mathbb{C}^{-1} = \begin{pmatrix} \frac{1}{E_x} & -\frac{\nu_{yx}}{E_y} & -\frac{\nu_{zx}}{E_z} & 0 & 0 & 0 \\ -\frac{\nu_{xy}}{E_x} & \frac{1}{E_y} & -\frac{\nu_{zy}}{E_z} & 0 & 0 & 0 \\ -\frac{\nu_{xz}}{E_x} & -\frac{\nu_{yz}}{E_y} & \frac{1}{E_z} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{G_{yz}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{G_{zx}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{G_{xy}} \end{pmatrix}, \qquad (13)$$

where ν_{xy}, ν_{xz}, ... denote Poisson's ratios and G_{xy}, G_{zx}, G_{xy} the shear moduli in the corresponding directions.

We compare our simulation results to those obtained by direct simulations in the CT image. For the longitudinal (y) direction, the variations within one synthetic foam type do not influence the effective stiffness significantly, see Fig. 10. For the two synthetic foam models with circular struts, the stiffness is underestimated. On the other hand, concave triangular struts lead to overestimation of the stiffness. The true strut cross section observed in the CT image of the real foam is in between both models—it is clearly triangular, but less concave than the model. The relaxed version of the synthetic foam with triangular struts is nevertheless able to reproduce the stiffness behavior in longitudinal direction of the real foam very well as the relaxation procedure reduces the effective stiffness for both strut geometries.

Analysis of the mechanical properties of the synthetic foams is completed by the stiffness in the transversal directions (x and z) reported in Fig. 11. The anisotropic behavior of the real foam is captured by all synthetic foam models. The stiffness in the transversal directions is smaller than in longitudinal direction, agreeing with the experimental results as well as with those obtained by direct simulation in the binarized CT image. Not surprisingly, the two relaxed synthetic foams

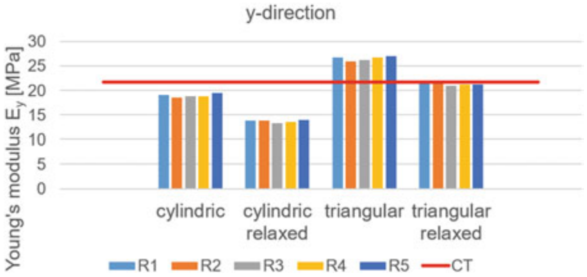


Fig. 10 Comparison of effective stiffness of synthetic and real foams in longitudinal direction. CT is the real foam's micro-structure as given by the binarized CT image. R1–R5 are the five realizations of the fitted Laguerre tessellation model

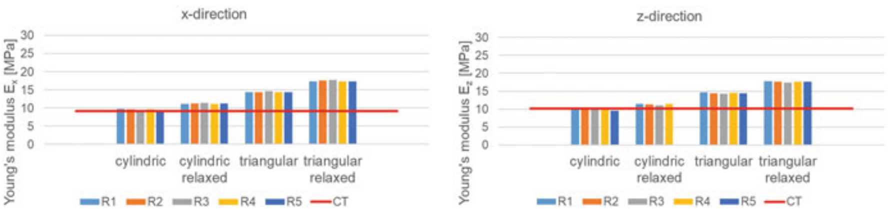


Fig. 11 Comparison of effective stiffness of synthetic (derived from R1 to R5) and real (CT) foams in transversal directions

behave less anisotropic than their unrelaxed counterparts. Relaxation increases the effective stiffness in transversal direction, while it decreases the effective stiffness in longitudinal direction. The unrelaxed synthetic foam with the circular struts matches the mechanical behavior of the foam in transversal direction best.

In summary, all four synthetic foam models reproduce the mechanical compression behavior of the real aluminum foam well, including the anisotropy. Taking into account all three directions, the effective stiffness of the real foam is best reproduced by the unrelaxed synthetic foam with circular strut cross sections, although the stiffness in longitudinal direction is slightly underestimated.

6 Conclusion

In this contribution, we describe completely the well-established nevertheless rarely fully expanded workflow for simulating mechanical properties of a material based on stochastic geometry models of its micro-structure. Our use case is an open metal foam. Consequently, a random tessellation model is fit to it.

Stochastic geometry models do not only capture naturally the microscopic heterogeneity of materials structures. Additionally, fitting them to the observed structure allows to generate many realizations as large as needed. Moreover, the effects of selected micro-structural geometric features can be investigated independently as demonstrated here for cross-sectional shape of the struts and relaxation. Model fitting and careful validation of the simulations further open the opportunity to determine the size of the representative volume and to devise optimized micro-structures.

Validation or calibration of the simulations by experiments is indispensable as the respective properties of the bulk material needed as input are hard to get. The values for the material, e.g., in the struts can differ significantly from tabular values due to effects of the production process on scales finer than the micro-structure.

The elastic properties of the aluminum foam could be reproduced well by all four synthetic foams considered here. More sophisticated geometries might be needed when moving on to plastic properties. For instance, the local thickness of the struts could be fit [21, 35]. Moreover, real cellular structures can feature much more complex anisotropies due to production processes being based on polymer foams. In the latter, the foaming direction usually stands out. The structure might however be additionally distorted, e.g., by the movement of a conveyor belt during foaming. Adequate characterization methods and modelling are subject of the current research.

Acknowledgments We thank Franz Schreiber for the CT scans, Felix Arnold for assisting in the image processing of the CT image, and Sonja Föhst for providing the code for model simulation.

References

1. S. Armstrong, T. Kuusi, J.C. Mourrat, *Quantitative Stochastic Homogenization and Large-Scale Regularity* (Springer, Berlin, 2019)
2. F. Aurenhammer, A criterion for the affine equivalence of cell complexes in $\mathbb{R}^d$ and convex polyhedra in $\mathbb{R}^{d+1}$. Discrete Comput. Geometry **2**, 49–64 (1987)
3. D. Balzani, L. Scheunemann, D. Brands, J. Schröder, Construction of two- and three-dimensional statistically similar RVEs for coupled micro-macro simulations. Comput. Mech. **54**(5), 1269–1284 (2014). https://doi.org/10.1007/s00466-014-1057-6
4. A. Bezrukov, D. Stoyan, M. Bargieł, Spatial statistics for simulated packings of spheres. Image Anal. Stereol. **20**, 203–206 (2001)
5. A. Bezrukov, M. Bargieł, D. Stoyan, Statistical analysis of simulated random packings of spheres. Part. Part. Systems Char. **19**, 111–118 (2002)
6. B. Biswal, C. Manwart, R. Hilfer, Three-dimensional local porosity analysis of porous media. Physica A Statist. Mech. Appl. **255**(3), 221–241 (1998). https://doi.org/10.1016/S0378-4371(98)00111-3 . https://www.sciencedirect.com/science/article/pii/S0378437198001113
7. K.A. Brakke, The surface evolver. Exp. Math. **1**(2), 141–165 (1992)
8. S. Chiu, D. Stoyan, W.S. Kendall, J. Mecke, *Stochastic Geometry and Its Applications*, 3rd edn. (Wiley, Chichester, 2013)
9. Elastodict. https://www.geodict.de/Modules/Dicts/ElastoDict.php
10. F. Feyel, J.L. Chaboche, FE2 multiscale approach for modelling the elastoviscoplastic behaviour of long fibre SiC/Ti composite materials. Comput. Methods Appl. Mech. Eng. **183**, 309–330 (2000)
11. S. Föhst, S. Osterroth, F. Arnold, C. Redenbach, Influence of geometry modifications on the permeability of open-cell foams. AIChE Journal. **68**(2), e17446 (2022)
12. M. Geers, V. Kouznetsova, W. Brekelmans, Multi-scale computational homogenization: trends and challenges. J. Comput. Appl. Math. **234**, 2175–2182 (2010)
13. Geodict. https://www.geodict.de/
14. H. Grimm-Strele, M. Kabel, Fast fourier transform based homogenization with mixed uniform boundary conditions. Int. J. Num. Meth. Eng. **122**, 7241–7265 (2021). Published online
15. H. Hadwiger, *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie* (Springer, Berlin, 1957)
16. S. Heinze, T. Bleistein, A. Düster, S. Diebels, A. Jung, Experimental and numerical investigation of single pores for identification of effective metal foams properties. Zeitschrift für angewandte Mathematik **98**, 682–695 (2018)
17. L. Helfen, H. Stanzick, J. Ohser, K. Schladitz, P. Rejmánková-Pernot, J. Banhart, T. Baumbach, Investigation of the foaming process of metals by synchrotron-radiation imaging, in ed. by B.M. Norbert Meyendorf George, Y. Baaklini, *Proceedings SPIE 5045: Testing, Reliability, and Application of Micro- and Nano-Material Systems*, vol. 5045 (2003), pp. 254–265
18. T. Hildebrand, P. Rüegsegger, A new method for the model independent assessment of thickness in three-dimensional images. J. Microsc. **185**, 67–75 (1997)

19. D. Jeulin, *Morphological Models of Random Structures* (Springer International Publishing, Cham, 2021). https://doi.org/10.1007/978-3-030-75452-5
20. A. Jung, S. Diebels, Microstructural characterisation and experimental determination of a multiaxial yield surface for open-cell aluminium foams. Mater. Design **131**, 252–264 (2017)
21. A. Jung, T. Bleistein, M. Reis, X. Cheng, C. Redenbach, S. Diebels, Multiscale microsphere modelling of open-cell metal foams enriched by statistical analysis of geometric parameters. Mech. Mater. **142**, 103295 (2020)
22. M. Kabel, H. Andrä, Fast numerical computation of precise bounds of effective elastic moduli. ITWM Berichte **224**, 1–16 (2013)
23. M. Kabel, D. Merkert, M. Schneider, Use of composite voxels in FFT-based homogenization. Comp. Meth. Appl. Mech. Eng. **294**, 168–188 (2015)
24. M. Kabel, S. Fliegener, M. Schneider, Mixed boundary conditions for fft-based homogenization at finite strains. Comput. Mech. **67**, 193–210 (2016)
25. M. Kabel, A. Fink, M. Schneider, The composite voxel technique for inelastic problems. Comp. Meth. Appl. Mech. Eng. **322**, 396–418 (2017)
26. T. Kanit, S. Forest, I. Galliet, V. Mounoury, D. Jeulin, Determination of the size of the representative volume element for random composites: statistical and numerical approach. Int. J. Solids Struct. **40**(13), 3647–3679 (2003). https://doi.org/10.1016/S0020-7683(03)00143-4
27. D. Kendall, Foundations of a theory of random sets, in *Stochastic Geometry. A Tribute to the Memory of Rollo Davidson* (Wiley, Hoboken, 1974)
28. E. Kröner, Bounds for effective elastic moduli of disordered materials. J. Mech. Phys. Solids **25**, 127–155 (1977)
29. C. Lang, J. Ohser, R. Hilfer, On the analysis of spatial binary images. J. Microsc. **203**, 303–313 (2001)
30. C. Lautensack, Random Laguerre Tessellations. Ph.D. Thesis, Universität Karlsruhe, Verlag Lautensack, Weiler bei Bingen, 2007
31. C. Lautensack, Fitting three-dimensional Laguerre tessellations to foam structures. J. Appl. Statist. **35**(9), 985–995 (2008)
32. C. Lautensack, S. Zuyev, Random Laguerre tessellations. Adv. Appl. Probab. **40**(3), 630–650 (2008)
33. A. Liebscher, Laguerre approximation of random foams. Philos. Mag. **95**(25), 2777–2792 (2015). https://doi.org/10.1080/14786435.2015.1078511.
34. A. Liebscher, C. Redenbach, 3D image analysis and stochastic modelling of open foams. Int. J. Mat. Res. **103**(2), 155–161 (2012)
35. A. Liebscher, C. Redenbach, Statistical analysis of the local strut thickness of open cell foams. Image Analy. Stereol. **32**(1), 1–12 (2013). https://doi.org/10.5566/ias.v32.p1-12. https://www.ias-iss.org/ojs/IAS/article/view/944
36. B. Lippmann, J. Schwinger, Variational principles for scattering processes. Phys. Rev. **79**, 469–480 (1950)
37. G. Lohmann, *Volumetric Image Analysis* (Wiley-Teubner, Chichester, Leipzig, 1998)
38. J. Luksch, T. Bleistein, K. Koenig, J. Adrien, E. Maire, A. Jung, Microstructural damage behaviour of al foams. Acta Materialia **208**, 116739 (2021)
39. G. Matheron, Ensembles fermeés aléatoires, ensembles semi-markoviens et polyèdres poissoniens. Adv. Appl. Probab. **4**, 508–541 (1972)
40. G. Matheron, *Random Sets and Integral Geometry* (Wiley, New York, 1975)
41. C.R. Maurer, V. Raghavan, A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. IEEE Trans. Pattern Analy. Mach. Intell. **25**(2), 265–270 (2003)
42. H. Moulinec, P. Suquet, A numerical method for computing the overall response of nonlinear composites with complex microstructure. Comp. Meth. Appl. Mech. Eng. **157**, 69–94 (1998)
43. W. Nagel, J. Ohser, K. Pischang, An integral-geometric approach for the Euler-Poincaré characteristic of spatial images. J. Microsc. **198**, 54–62 (2000)
44. J. Ohser, K. Schladitz, *3D Images of Materials Structures: Processing and Analysis* (Wiley VCH, Weinheim, 2009)

45. J. Ohser, W. Nagel, K. Schladitz, The Euler number of discretized sets—on the choice of adjacency in homogeneous lattices, in ed. by K.R. Mecke, D. Stoyan, *Morphology of Condensed Matter, LNP*, vol. 600 (Springer, Berlin, 2002), pp. 275–298
46. J. Ohser, W. Nagel, K. Schladitz, The Euler number of discretised sets – surprising results in three dimensions. Image Anal. Stereol. **22**, 11–19 (2003)
47. N. Otsu, A threshold selection method from gray level histograms. IEEE Trans. Syst. Man Cybern. **9**, 62–66 (1979)
48. A. Rack, L. Helfen, T. Baumbach, S. Kirste, J. Banhart, K. Schladitz, J. Ohser, Analysis of spatial cross-correlations in multi-constituent volume data. J. Microsc. **232**(2), 282–292 (2008)
49. C. Redenbach, Microstructure models for cellular materials. Comput. Mat. Sci. **44**(4), 1397–1407 (2009)
50. C. Redenbach, I. Shklyar, H. Andrä, Laguerre tessellations for elastic stiffness simulations of closed foams with strongly varying cell sizes. Int. J. Eng. Sci. **50**(1), 70–78 (2012). https://doi.org/10.1016/j.ijengsci.2011.09.002. https://www.sciencedirect.com/science/article/pii/S0020722511001807
51. K. Schladitz, J. Ohser, W. Nagel, Measurement of intrinsic volumes of sets observed on lattices, in ed. by A. Kuba, L.G. Nyul, K. Palagyi, *13th International Conference on Discrete Geometry for Computer Imagery, LNCS*, vol. 4245. (DGCI/Springer, Szeged/Berlin, 2006), pp. 247–258
52. K. Schladitz, S. Peters, D. Reinel-Bitzer, A. Wiegmann, J. Ohser, Design of acoustic trim based on geometric modeling and flow simulation for non-woven. Comput. Mater. Sci. **38**(1), 56–66 (2006)
53. K. Schladitz, C. Redenbach, T. Sych, M. Godehardt, Model based estimation of geometric characteristics of open foams. Methodol. Comput. Appl. Probab. **14**, 1011–1032 (2012)
54. R. Schneider, *Convex Bodies: The Brunn–Minkowski Theory*. No. 44 in Encyclopedia of Mathematics and Its Application (Cambridge University Press, Cambridge, 1993)
55. R. Schneider, W. Weil, *Stochastic and Integral Geometry*. Probability and Its Applications (Springer, Heidelberg, 2008)
56. M. Schneider, A review of nonlinear FFT-based computational homogenization method. Acta Mechanica **232**, 2051–2100 (2021)
57. M. Schneider, F. Ospald, M. Kabel, Computational homogenization of elasticity on a staggered grid. Int. J. Num. Methods Eng. **105**, 693–720 (2016)
58. J. Serra, *Image Analysis and Mathematical Morphology*, vol. 1 (Academic, London, 1982)
59. P. Spanne, J.F. Thovert, C.J. Jacquin, W.B. Lindquist, K.W. Jones, P.M. Adler, Synchrotron computed microtomography of porous media: topology and transports. Phys. Rev. Lett. **73**, 2001–2004 (1994). https://doi.org/10.1103/PhysRevLett.73.2001. https://link.aps.org/doi/10.1103/PhysRevLett.73.2001
60. A. Spettl, T. Brereton, Q. Duan, T. Werz, C.E. Krill III, D.P. Kroese, V. Schmidt, Fitting laguerre tessellation approximations to tomographic image data. Philos. Mag. **96**(2), 166–189 (2016). https://doi.org/10.1080/14786435.2015.1125540
61. S. Staub, H. Andrä, M. Kabel, Fast fft based solver for rate-dependent deformations of composites and nonwovens. J. Solids Struct. **154**, 33–42 (2018)
62. P. Suquet, Elements of homogenization for inelastic solid mechanics, in *Homogenization Techniques for Composite Materials* (Springer, Berlin, 1985), pp. 193–278
63. F.B. Tek, A.G. Dempster, I. Kale, Blood cell segmentation using minimum area watershed and circle Radon transformations, in C. Ronse, L. Najman, E. Decencière, *Proceedings of the 7th International Symposium on Mathematical Morphology*. Computational Imaging and Vision, vol. 30 (Springer, Dordrecht, 2005), pp. 441–454
64. D. Ulrich, T. Hildebrand, B. van Rietbergen, R.Müller, P. Rüegsegger, The quality of trabecular bone evaluated with micro-computed tomography, FEA and mechanical testing, in G. Lowet, P. Rüegsegger, H. Weinans, *Bone Research in Biomechanics, Studies in Health Technology and Informatics* (IOS Press, Netherlands, 1997), pp. 97–112
65. I. Vecchio, C. Redenbach, K. Schladitz, A. Kraynik, Improved models of solid foams based on soap froth. Comput. Mater. Sci. **120**, 60–69 (2016)

66. L. Vincent, P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulation. IEEE Trans. Pattern Analy. Mach. Intell. **13**(6), 583–598 (1991)
67. D. Weaire, S. Hutzler, *The Physics of Foams* (Oxford University Press, Oxford, 1999)
68. R. Zeller, P. Dederichs, Elastic constants of polycristals. Phys. Status Solidi B **55**, 831–842 (1973)

Machine Learning for Failure Analysis: A Mathematical Modelling Perspective

Judith Pérez-Velázquez, Meltem Gölgeli, and Carlos Alfonso Ruiz Guido

1 Introduction

An object's usability, affordability, and reliability are very much determined by the materials and processes employed to create it. The discipline of failure analysis is heavily rooted in materials science and involves understanding and determining the cause of a failure. From the mathematical viewpoint, we can roughly separate the approaches used by those that make **diagnosis**, i.e., methods that allow to assess the extent of deviation or degradation from an expected normal operating condition, for example, by training a neural network with images portraying an ample range of degradation states of a material such as the profiles of tyres and those aimed to make **prognosis**, which possess prediction capabilities. For example, using the current and past machine condition to predict how much time is left before a failure occurs.

Failure is usually investigated in terms of failure criteria, which separate "failed" states from "unfailed" states. There are two ways to define this separation:

Threshold-dependent The definition of failure is simply defined that the failure occurs when the fault reaches a predetermined level.

J. Pérez-Velázquez (✉)
School of Computation, Information and Techology (CIT), Technische Universität München, Garching, Germany
e-mail: cerit@cit.tum.de

M. Gölgeli
TOBB University of Economics and Technology, Ankara, Turkey
e-mail: mgolgeli@etu.edu.tr

C. A. Ruiz Guido
Colegio de Matemáticas Bourbaki, Mexico City, Mexico
e-mail: alfonso@escuela-bourbaki.com

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_12

Data-dependent In this case, different definitions of failure can be given: an event that the machine is operating at an unsatisfactory level; a functional failure when the machine cannot perform its intended function at all; a breakdown when the machine stops operating, etc.

Formally, failure is a gradual or sudden loss of the ability to operate. In the context of material sciences, failure can be accredited to one of four reasons: Design, when a material is modified to such an extent that it no longer serves its original designed purpose; Manufacturing, the way certain materials are processed and pieced together can also generate failure; Service, the interaction with the environment such as excessive wear, overloading, can lead to failures in service; Material, material-related problems may lead to failure such as poor micro-structure, material defects and contamination from foreign particles.

Probability theory has traditionally been used in failure analysis [6, 11]. Standard approaches used here commonly employ current and past machine condition to predict how much time is left before a failure occurs. The data era has brought about approaches that employ data-driven methods to not only detect failure but also make predictions about the reliability of devices. For example, using neural networks to evaluate the role of solder material and the joint thickness on the reliability of electronic devices [16].

These data-driven approaches such as artificial intelligence (AI) and its sub-area machine learning (ML) are increasingly being used in all aspects of failure analysis, from predictive maintenance to plant facilities monitoring, including remaining useful life forecasting, etc.

In this chapter, we present machine learning approaches in failure analysis. From the mathematical viewpoint, these approaches constitute a wide spectrum of models, ranging from survival random forests for reliability analysis to convolutional neural networks for defect classification, including hybrid approaches that allow to include physical information to describe aging and degradation. As we will see, reliability in this context describes the ability of a device to function under stated conditions for a specified period of time, and time-dependent models are used here.

This chapter is intended to serve two purposes. First is a description and illustration of the assumptions and basic models of ML in failure analysis. Second is to present a variety of applications in this context. The idea is to give an overview to mathematicians of the ML approaches used in this sub-area of material science.

This chapter is structured as follows. In Sect. 2, we describe the approach used in survival analysis, failure prognosis/prediction, also called time-to-event analysis, in our context, “time to failure.” It consists of a set of statistical analyses that takes a series of observations and attempts to estimate the time it takes for an event of interest to occur. Classical tools of survival analysis have been recently powered by the use of ML. In Sect. 2, we define ML methods, and finally in Sect. 3, we present some use cases. Note that the overall aim of the chapter is to present the basic assumptions and definitions of ML and illustrate how these can be applied in failure analysis. The details of these approaches are covered in many papers and textbooks

[2, 4, 7, 10]; therefore, here we restrict our attention to present the fundamentals that underlie all of these procedures.

2 Survival Analysis

The classical survival analysis in material science consists of dynamic statistical methods to analyse the time to a certain event. The incomplete observations due to time limitations or loss of some data points are known as censoring, and the ability of these methods to include censoring data is the main difference of survival analysis compared to standard regression methods. As explained in Sect. 1, undesired changes of a certain state (oxidation, refraction, crash, etc.) of a material are examples of an event that we call "failure." Thus, the lifetime distributions are described by the survival or hazard function. We present the general terminology of survival analysis based on [15].

Definition 2.1 *Let us define the probability density function $f(x)$ of having a failure at time x:*

- *The cumulative distribution function $P(T \leq t) = \int\limits_0^t f(x)dx = F(t)$ represents the survival time of a material until time T, where T is a continuous random variable that refers the time to failure.*
- *The survival function $S(t) = P(T > t) = 1 - F(t)$ demonstrates the probability of a material surviving beyond a certain time t.*
- *The hazard function $h(t) = \dfrac{f(t)}{S(t)}$ represents the conditional failure rate that indicates the likelihood of the event occurring at time t given that no event has occurred before time t.*
- *The cumulative hazard function $H(t) = \int\limits_T^\infty h(x)dx$ represents the instantaneous rate of failure at a certain time t.*

□

Generally, the modelling approach of the classical survival analysis is classified into three subgroups: parametric models, non-parametric models, and semi-parametric models. We present some common examples of these approaches below.

Example 2.2

- *Parametric approach:*
  - *The Weibull distribution has a survival function $S(t) = exp(-\lambda t^{\alpha})$, where $t \geq 0$ and $\lambda > 0$ is a scale parameter and $\alpha > 0$ is a shape parameter and hazard function $h(t) = \alpha\lambda t^{\alpha-1}$.*
- *Non-parametric approach:*

- *The Kaplan–Meier estimator is defined by* $\hat{S}(t) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j}$*, where* d_j *is the number of system components that have an event at time* t_j*, where* $j = 1, \ldots, k$*;* m_j *is the number of system components censored in the interval* $[t_j, t_j + 1)$*; and* $n_j = (m_j + d_j) + \cdots + (m_k + d_k)$ *is the number of system components at risk just prior to* t_j*.*
- *The Nelson–Aalen estimator is a method to estimate and plot the cumulative hazard function* $H_{NA}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$*, where* d_i *is the number of system components that have an event at time* t_i *and* n_i *is the total components at risk at time* t_i*.*

- *Semi-parametric approach:*
  - *Cox regression is expressed by the hazard rate* $h(t, x) = h_0(t) \exp(xB)$*, where* $h_0(t)$ *is a baseline function,* z *is the variable, and* B *is the hazard coefficient for the variable. The hazard ratio between the two groups in a factor can be estimated by using* $H_R(t, x_1, x_2) = \exp(B(x_1 - x_2))$*.*

□

Example 2.3 *Öztürk et al. presented in [15] an application of survival analysis on wind turbine reliability taking into account data such as previous failures and the history of scheduled maintenance. The probability of failure of a wind turbine at a certain time is defined by the Kaplan–Meier estimator, and the cumulative hazard function is estimated by a Nelson–Aalen estimator. Then, a comparison for the survival of separate groups of wind turbines is given by applying statistical tests such as a log-rank test. As a result, the survival of frequently failing wind turbine components between the geared-drive and direct-drive wind turbines is compared.*□

3 Machine Learning

Machine learning is a set of algorithms that use databases in order to build mathematical models and solve tasks such as classification, regression, clusterization, anomaly detection, etc. Since this text is focused on tasks relevant to failure analysis, we will only focus on the aforementioned problems; however, we invite the reader to explore the diverse applications of machine learning algorithms. For the readers who are interested in knowing more on machine learning and its applications, we recommend you to read [2, 7, 10, 16].

Historically, there is a difference between discriminative and generative methods in machine learning; in the former, the prior is the set of characteristics from the independent variable, and in the latter, the prior uses the dependent variable. In this text, we will follow this classification in Sects. 3.1 and 3.2, respectively.

3.1 Discriminative Machine Learning

In this section, we will define and exemplify a discriminative machine learning problem.

Definition 3.1 *A discriminative problem consists of a tuple* (d, X, Y, S)*, where:*

- $d \in \mathbb{N}$ *is a positive integer that from now on will be referred to as the dimension of our problem.*
- *A set* X *where our observations live,* $X = \mathbb{R}, \{0, 2\}^d, \mathbb{Z}$.
- *A set* Y *where the true labels of our data live,* $Y = \{-1, +1\}, \{c_1, c_2, \ldots, c_K\}, \mathbb{R}, \mathbb{N}$.
- S *is a family of random variables* $S_1, S_2, S_3, \ldots$ *over the set:* $X^d \times Y$*, the random variables with marginal distributions from* S_i *over each coordinate* j *of* X^d *will be named* $X_{i,j}$ *and* Y_i *over* Y.

Depending on the choice of Y*, the problem we aim to solve could be a binary classification problem, anomaly detection, multi-class classification, regression, or agnostic clustering.*

When we add to the tuple of a machine learning problem a family of functions $F : \{f : X^d \to Y\}$*, the tuple* (d, X, Y, S, F) *will be called in this text a machine learning model.*

Remark It is common in machine learning to assume that the random variables S_i in S are independent and identically distributed, and this is not the most realistic assumption.

Let us start with some examples of machine learning models from a mathematical point of view.

Example 3.3 *A linear regression model will assume that* $X = \mathbb{R}, Y = \mathbb{R}$ *y* $F_{Lin} = \{f_\beta(x) = \langle \beta, x\rangle + \beta_0 : \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}$*. For this problem, we will suppose that the random variables* X_i, Y_i *are such that there exists some* $\beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}$*, and it is satisfied that* $\langle \beta, X_i\rangle + \beta_0 + \epsilon_i = Y_i$*, where* ϵ_i *is a set of Gaussian random variables, which are independent and identically distributed.* □

Remark In the previous section, we talked about a censored regression problem that in some cases will not provide us with credible information about the variable Y, meaning that some of the observations y were not generated by Y; in this case, we are talking about a censored survival analysis problem. In the case of survival analysis *à la Cox* (see Example 2.2), the family of functions is of the form: $h(t, x) = h_0(t)\exp(xB)$.

Example 3.5 *A polynomial regression model of degree* g *is similar to the previously described problem except that, instead of considering* $F = F_{Lin}$*, our set of functions will be the set of polynomials of degree* g.

Example 3.6 *A linear classification model assumes that* $X = \mathbb{R}, Y = \{-1, +1\}$ *and* $F_{LinC} = \{f_\beta(x) = sign(\langle \beta, x\rangle + \beta_0) : \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}$*. In this case, we* *suppose that the random variables: X_i, Y_i are such that there exist $\beta \in \mathbb{R}^d$, $\beta_0 \in \mathbb{R}$ and these satisfy:*

$$\mathbb{P}\left(\left(\langle \beta, X_i \rangle + \beta_0\right) Y_i > 0\right) = 1.$$

In this case, Y_i are Bernoulli random variables. □

Example 3.7 *A support vector machine (SVM) model with margin γ will assume that $X = \mathbb{R}, Y = \{-1, +1\}$ y $F_{LinC} = \{f_\beta(x) = sign(\langle \beta, x \rangle + \beta_0) : \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}$. In this case, we suppose that the random variables X_i, Y_i are such that there exists some $\beta \in \mathbb{R}^d$, $\beta_0 \in \mathbb{R}$, and some $\gamma > 0$ such that*

$$\left(\langle \beta, X_i \rangle + \beta_0\right) Y_i > 0$$

and

$$\mathbb{P}\left(Y_i \left(\langle \beta, X_i \rangle + \beta_0\right) \geq \gamma ||| X_i ||_2 = 1\right) = 1.$$

Example 3.8 *A binary decision tree model with depth $r \leq d$ decision trees where we assume that $X = \{0, 2\}$, $Y = \{-1, +1\}$ and F_{DS} a set of all possible functions $f : \{0, 2\}^d \rightarrow \{-1, +1\}$ that correspond to a binary decision tree whose branches are labelled with $-1, +1$. It is also possible to use decision trees for regression problems when $X = Y = \mathbb{R}$.* □

Example 3.9 *A neural network model with feed-forward and a single activation function (see [17] for more details) will assume that F is a family of functions described as follows:*

- *Let us fix some acyclic directed graph $G = (V, E)$.*
- *A function called weight function over the set of edges: $w : E \rightarrow \mathbb{R}$.*
- *A function called the activation function $\rho : \mathbb{R} \rightarrow \mathbb{R}$, for example, the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$.*
- *A disjoint partition of $V = V_1 \cup \ldots \cup V_T$, where each node in V_{t-1} is connected to some in V_t, and s will be the number of layers in the network.*
- *V_1 is a distinguished set of vertices of size $d + 1$ and V_T is a single node.*
- *A function in F, $f : X^d \rightarrow Y$ is a neural network if it can be calculated using the information below always going from V_{t-1} to V_t.*

□

Example 3.10 *An agnostic clusterization model will assume that $X = \mathbb{R}, Y = \mathbb{N}$ and $F_{Clus} = \{f(x) \in \mathbb{N}\}$.* □

Example 3.11 *A clusterization model with K clusters will assume that $X = \mathbb{R}, Y = \{c_1, c_2, \ldots, c_K\}$ and $F_{Clus,K} = \{f(x) \in \{c_1, c_2, \ldots, c_K\}\}$. For this problem, the random variables X_i, Y_i are such that there exist some $f \in F_{Clus,K}$ satisfying $\mathbb{P}(f(X_i) = Y_i) = 1$.* □

Example 3.12 *A time series model ARMA(p,q) assumes that* $d = 1$, $X = \mathbb{N}$, $Y = \mathbb{R}$, *and* $F_{ARMA(p,q)}$ *is such that for any* $t \in \mathbb{N}$ *the random variables* X_i, Y_i *over* $\mathbb{N}, \mathbb{R}$ *satisfy*

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t, \epsilon_t \sim WN\left(\mu, \sigma^2\right), \phi_j, \theta_j \in \mathbb{R}.$$

Example 3.13 *A anomaly detection model will assume that* $X = \mathbb{R}$, $Y = \{-1, +1\}$ *and* $F_{Anom} = \{f(x) \in \{-1, +1\}\}$. □

3.1.1 The Algorithms of Machine Learning

In the previous section, we talked about a machine learning problem from a mathematical point of view; however, in practice, of course we will not have access to the stochastic process S; instead, we will have access to a database in the following sense:

Definition 3.14 *Given a stochastic process* S *over* $X^d \times Y$, *a supervised dataset of size* N *is a set* $S_{Emp} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ *of experiments of* N *random variables* $S_1, S_2, \ldots, S_N$ *inside* S. *We call* y_i *the label of the example* x_i, *and not all datasets are of this type.* □

Using these databases, the discriminative machine learning algorithms are trying to build a function $f_{S_{emp}} : X \to Y$. It is worth mentioning that in practice some problems will not be supervised, and still it is possible to build a model inside F from the unsupervised dataset $\{x_1, x_2, \ldots, x_N\}$.

Definition 3.15 *Given a discriminative machine learning model* (d, X, Y, S, F), *an algorithm is a function* A *such that for all* N *and any dataset* S_{Emp} *with size* N *generated by* S, *it assigns a function* $f_{S_{Emp}} \in F$. □

In order to give some examples of algorithms, we first need to give some examples of the so-called loss functions.

Example 3.16 *In the context of a linear regression, the loss function of an* $f \in F_{Lin}$ *is the following:*

$$l_{S_{emp}}(f) = \frac{1}{N} \sum_{i \leq N} (y_i - \langle \beta, x_i \rangle - \beta_0)^2 .$$

Example 3.17 *An example of a loss function in the case of a classification model is the following:*

$$l_{S_{emp}}(f) = |\{i : f(x_i) \neq y_i\}|/N.$$

Unfortunately, this function is far from being as smooth as the previous one; this could be a problem because classical optimization methods might not work. □

Definition 3.18 *Let (d, X, Y, S, F) a binary classification problem for example as in Example 3.6 or in Example 3.8. If S_{emp} is a dataset of size N whose examples are indexed by i, for a given loss function l, we define the empirical risk minimization (family of) algorithm(s) as follows:*

$$A_{ERM}\left(S_{emp}\right) = \underset{f \in F}{argmin}\left(l_{S_{emp}}(f)\right).$$

Remark When the loss function is the one given in Example 3.16, we recover the classical least square error.

3.1.2 Evaluating a Machine Learning Model

So far we have defined what problem and algorithm are in machine learning, but we are still missing how an algorithm is evaluated. Notice that in the previous section we defined the empirical error, which is associated with a single dataset; the so-called true error of the model should depend on the stochastic process S; for the *connoisseur*, this difference is close to the classical split into train and test.

In this section, we will only focus on a linear regression or binary classification problem. It is worth mentioning that for some of the described cases such as agnostic clustering or anomaly detection, the evaluation process might be an extremely difficult problem, see [12].

Definition 3.20 *Let (d, X, Y, S, F) be a discriminative machine learning model either when we are in Example 3.3 or in Example 3.6.*

- *If we are in Example 3.3, for each $f : \mathbb{R}^d \to \mathbb{R}$ in F_{Lin} and each $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, we define $l(f, (x, y)) = (y - f(x))^2$.*
- *Let S_i be one of the random variables in S; if $f : \mathbb{R}^d \to \mathbb{R}$ is a function in F_{Lin}, we define the true error of f over S_i as $L_{S_i}(f) = \mathbb{E}[l(f, S_i)]$.*
- *If we are in Example 3.6), for each $f \in F_{LinC}$ and each $(x, y) \in \mathbb{R}^d \times \{-1, +1\}$, we define $l(f, (x, y)) = 0$ if $f(x) = y$ and $l(f, (x, y)) = 1$ if $f(x) \neq y$.*
- *Let S_i be one of the random variables in S, $f \in F_{LinC}$; we define the true error of f over S_i as follows: $L_{S_i}(f) = \mathbb{E}[l(f, S_i)]$.*

Remark Notice that the previous definitions are difficult to calculate since they depend on the random variables S. In practice, there exists several statistics that could be computed for a dataset and a machine learning model; some notable examples are:

- R^2 for linear regression
- Confusion matrix for classification problems
- Silhouette function for clustering

3.1.3 Under-fitting and Over-fitting

In this section, we are going to talk about two undesirable mathematical properties an algorithm might have: over-fitting and under-fitting.

Let us start with the first one; in order to motivate the formal definition, we give an intuitive definition of one of the most delicate problems in machine learning: the infamous over-fitting.

Definition 3.22 (Informal Definition of Over-fitting) *Given a discriminative machine learning model, an algorithm A over-fits when it systematically obeys the noise of the joint distribution* $(S_1, S_2, \ldots)$. □

We have not defined what "systematically obeying the noise" means. In order to give a formal definition of over-fitting, we introduce one possible mathematical definition of it.

Definition 3.23 *Given a discriminative machine learning model* (d, X, Y, S, F) *and a supervised dataset* S_{emp}*, we define a validation set of size* N' *as a set of experiments of the random variables in S*

$$S_V = \{(x'_1, y'_1), \ldots, (x'_{N'}, y'_{N'})\}$$

that is statistically independent from the original dataset S_{emp}. □

Definition 3.24 *Given a discriminative machine learning model* (d, X, Y, S, F) *together with a dataset* S_{emp} *and a validation set* S_V*, for some function* $f \in F$*, we define:*

- *The over-fitting measure of* f *is* $\left(l_{S_V}(f) - l_{S_{emp}}(f)\right)$*. When this quantity is large, we will say that the model over-fits.*
- *The under-fitting measure of* f *is* $l_{S_{emp}}$*. When this quantity is large, we will say that the model under-fits.*

□

3.2 Generative Machine Learning

Unlike discriminative machine learning model where the solutions are functions $f : X \to Y$, generative machine learning models will assume that the solutions are distributions over $X^d \times Y$.

Definition 3.25 *A generative machine learning model consists of* (d, X, Y, S, D)*, where:*

- $d \in \mathbb{N}$ *is a positive integer that from now on will be referred to as the dimension of our problem.*

- *A set X where our observations live, $X = \mathbb{R}, \{0, 2\}^d, \mathbb{Z}$.*
- *A set Y where the true labels of our data live, $Y = \{-1, +1\}, \{c_1, c_2, \ldots, c_K\}, \mathbb{R}, \mathbb{N}$.*
- *S is a family of random variables $S_1, S_2, S_3, \ldots$ over the set: $X^d \times Y$, the random variables with marginal distributions from S_i over each coordinate j of X^d will be named $X_{i,j}$ and Y_i over Y.*
- *$D = \{D_i : \mathbb{P}_{D_i}(X \times Y)\}$ is a family of distributions over $X^d \times Y$.*

Example 3.26 *A Naïve Bayes model assumes that $X = \{0, 2\}$, $Y = \{-1, +1\}$, and for each distribution $\mathbb{P}_{D_i} \in D$, the marginal distribution $\mathbb{P}_{D_i,\{0,2\}^d}$ over $\{0, 2\}^d$ is equal to the product of its d-marginal distributions, i.e.,*

$$\mathbb{P}_{D_i,\{0,2\}^d} = \prod_{j \le d} \mathbb{P}_{D_i,j}.$$

Example 3.27 *A Bayesian network model assumes there is some fixed graph $G = (V, E)$ with $|V| = d$, and the dependencies on the random variables $\{X_1, \ldots, X_d\}$ are modelled with the graph G.*

4 Use Cases

In this section, we include some use cases of machine learning models to the problem of failure analysis. It is worth mentioning that in many of the examples not only one model has been used.

4.1 Regression Models

Regression models come often into place when estimating the time to failure of a machine or its remaining useful life (RUL). There are several approaches to achieve this, namely one can use similarity methods that require run-to-failure data, survival methods that require lifetime data related to events such as part replacement and part failure and trend-based methods that require a known failure threshold.

4.1.1 Random Forest Regression

Random forest regression is an ensemble algorithm that makes predictions based on the average prediction of an ensemble of decision trees. See decision tree in Definition 3.8.

Example 4.1 *Biswas et al. [3] used a regression approach by estimating the remaining time to failure of disordered samples from plastic materials. This problem is difficult because this kind of data exhibits a time correlation that is classically used to enhance the predictions.* □

4.1.2 Survival Analysis

Example 4.2 *Zhang et al. [19] predicted RUL employing the NASA data repository. They proposed a time-dependent survival neural network that additively estimates a latent failure risk and performs multiple binary classifications to generate prognostics of RUL-specific probability.* □

4.1.3 Random Survival Forests

Random survival forests were introduced in [8]. They constitute survival models for predicting the probability of failure over time. This is a method for analysing right-censored survival data. Censored data is often found in survival tests. As compared to classical random forest, survival splitting rules for growing survival trees are introduced, as well as a missing data algorithm for dealing with missing (censored) data.

Example 4.3 *Weeraddana et al. [18] employed a random survival forest to predict the failure likelihood of water main breaking using historical failure records, descriptors of pipes, and other environmental factors.* □

4.1.4 Neural Networks

Example 4.4 *Samavatian et al. [16] used a co-relational neural network (NN) for reliability assessment of solder joints. They computed useful lifetime based on the materials properties, device configuration, and thermal cycling variations.* □

4.2 Classification Models

These models come into place in context such as failure classification, such as distinguishing between damage or fracture modes.

4.2.1 Support Vector Machines

Example 4.5 *In [14], using multi-class determination by support vector machine (SVM) of failure modes in a laser-irradiation device, it was shown that SVM could be successfully used to determine damage/fracture modes.* □

4.2.2 Neural Networks

Example 4.6 *Günnemann and Pfeffer [5] use convolutional neural networks on temporal input signals data from tests over combustion engines. The authors subsequently combined temporal data with additional static features; the results are very interesting because they deal with success with highly imbalanced data.* □

4.3 Anomaly Detection

A general definition of anomaly detection is given in Example 3.13 where x_i may refer to some subset of the data, e.g., an individual data point, a group of data points, a subsequence of a time series, or a region of an image.

Anomaly detection is a common approach for fault detection. Within the taxonomy of error, fault, and failure, an anomaly can be considered as a potential error, where an error is caused by a fault and may in turn cause a failure.

Example 4.7 *Jabbar et al. [9] used anomaly detection of manufacturing electronic cards employing variational autoencoders (VAEs), a deep Bayesian network. See Sect. 4.4 or Example 3.27.* □

4.4 Generative Models

4.4.1 Naïve Bayes

Example 4.8 *Addin et al. [1] introduce a Naïve Bayes model that simulates the damage detection in quasi-isotopic laminated composite materials.* □

4.4.2 Bayesian Networks

Example 4.9 *Medjaher et al. [13] describe a failure prognostic using Bayesian models for modelling complex systems with non-homogeneous sources of data and dealing with uncertainty by estimating the remaining useful life (RUL) before a failure.* □

5 Conclusions

In this chapter, we presented basic concepts of ML-enabled failure analysis. As a complement, we presented a selection of examples of the application of these models. By addressing mathematicians who have previously worked in material sciences, the chapter aims to create the basis for them to explore further the capabilities of ML.

A number of recommendations can be made: there are a number of available datasets that can be used to try specific applications; hybrid approaches that combine traditional methods with ML can improve model accuracy and enable new applications; the use of ML methods in failure analysis is likely to increase further, but this requires tailored methods in terms of efficiency and interpretability, and the mathematical community can make here a great contribution. As a final remark, we believe that machine learning has enhanced the set of tools in failure analysis and will continue to do so. This does, however, not mean that ML will fully replace other approaches. The use cases presented show that ML can indeed effectively predict failures or abnormalities in a wide range of applications.

References

1. O. Addin, S.M. Sapuan, E. Mahdi, M. Othman, A Naïve-Bayes classifier for damage detection in engineering materials. Mater. Des. **28**(8), 2379–2386 (2007)
2. C.A. Azencott, Introduction au Machine Learning, Dunod (2019).
3. S. Biswas, D. Fernandez Castellanos, M. Zaiser, Prediction of creep failure time using machine learning. Sci. Rep. **10**(1), 16910 (2020)
4. J. Feldman, R. Rojas, *Neural Networks: A Systematic Introduction* (Springer, Berlin, 2013)
5. N. Günnemann, J. Pfeffer, Predicting defective engines using convolutional neural networks on temporal vibration signals, in *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, ed. by L. Torgo, B. Krawczyk, P. Branco, N. Moniz, volume 74 of Proceedings of Machine Learning Research, ECML-PKDD, Skopje, Macedonia, 22 Sep 2017. PMLR (2017), pp. 92–102
6. C. Huber, N. Limnios, M. Mesbah, M.S. Nikulin, *Mathematical Methods in Survival Analysis, Reliability and Quality of Life*. ISTE (Wiley, 2013)
7. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (The MIT Press, 2016)
8. H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, Random survival forests. Ann. Appl. Stat. **2**(3), 841–860 (2008)
9. E. Jabbar, P. Besse, J.M. Loubes, C. Merle, Conditional anomaly detection for quality and productivity improvement of electronics manufacturing systems, in *Machine Learning, Optimization, and Data Science*, ed. by G. Nicosia, P. Pardalos, R. Umeton, G. Giuffrida, V. Sciacca (Springer International Publishing, Cham, 2019), pp. 711–724
10. G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R* (Springer, 2013)
11. A. Kaufmann, D. Grouchko, R. Cruon, *Mathematical Models for the Study of the Reliability of Systems*. ISSN (Elsevier Science, 1977)
12. J. Kleinberg, An impossibility theorem for clustering, in *Advances in Neural Information Processing Systems*, ed. by S. Becker, S. Thrun, K. Obermayer, vol. 15 (MIT Press, 2003)

13. K. Medjaher, J.Y. Moya, N. Zerhouni, Failure prognostic by using dynamic Bayesian networks. IFAC Proc. Vol. **42**(5), 257–262 (2009)
14. T. Okabe, Y. Otsuka, Proposal of a validation method of failure mode analyses based on the stress-strength model with a support vector machine. Reliab. Eng. Syst. Saf. **205**, 107247 (2021)
15. S. Ozturk, V. Fthenakis, S. Faulstich, Assessing the factors impacting on the reliability of wind turbines via survival analysis – a case study. Energies **11**(11), 3034 (2018)
16. V. Samavatian, M. Fotuhi-Firuzabad, M. Samavatian, P. Dehghanian, F. Blaabjerg, Correlation-driven machine learning for accelerated reliability assessment of solder joints in electronics. Sci. Rep. **10**(1), 14821 (2020)
17. S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning* (Cambridge University Press, 2019)
18. D. Weeraddana, S. MallawaArachchi, T. Warnakula, Z. Li, Y. Wang, Long-term pipeline failure prediction using nonparametric survival analysis, in *Machine Learning and Knowledge Discovery in Databases: Applied Data Science Track*, ed. by Y. Dong, D. Mladenić, C. Saunders (Springer International Publishing, Cham, 2021), pp. 139–156
19. J. Zhang, S. Wang, L. Chen, G. Guo, R. Chen, A. Vanasse, *Time-dependent survival neural network for remaining useful life prediction, in Advances in Knowledge Discovery and Data Mining. PAKDD 2019. Lecture Notes in Computer Science*, ed. by Q. Yang, Z. H. Zhou, Z. Gong, M. L. Zhang, S. J. Huang, vol. 11439, (Springer, Cham, 2019) https://doi.org/10.1007/978-3-030-16148-4_34

Invertibility of Orlicz–Sobolev Maps

Giovanni Scilla and Bianca Stroffolini

MSC2020 Classifications: 49J95, 46E30, 74B20

1 Introduction

We are interested in the invertibility of maps within the logarithmic scale of Orlicz–Sobolev spaces in view of the variational models in elasticity. In order to state our result, we will review the existing literature, focusing on the cavitation and fracture models.

Let Ω be a bounded open set of $\mathbb{R}^n$, and consider a map $\mathbf{u}\colon \Omega \to \mathbb{R}^n$. In nonlinear elasticity, the map $\mathbf{u}$ represents the deformation of a body that occupies the set Ω in the reference configuration. The first example of stored energy functional for compressible materials is

$$E(\mathbf{u}) = \int_\Omega W(D\mathbf{u}(\mathbf{x}))\,\mathrm{d}\mathbf{x}\,, \tag{1}$$

where $W(\boldsymbol{\xi}) = |\boldsymbol{\xi}|^p + g(\det D\boldsymbol{\xi})$, $p > 1$. The function g is assumed to be convex and accounts for changes in volume, that is, it blows up as $\det D\mathbf{u} \to +\infty$ (expand the solid) and as $\det D\mathbf{u} \to 0^+$ (compress it). The lower semicontinuity of such functionals together with suitable coercivity conditions was addressed in the papers by Ball and Murat [4] and Marcellini [34]. In the general case, namely, when $W =$

G. Scilla
Dipartimento di Scienze di Base ed Applicate per l'Ingegneria (SBAI), Sapienza Università di Roma, Roma, Italy
e-mail: giovanni.scilla@uniroma1.it

B. Stroffolini (✉)
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università di Napoli Federico II, Napoli, Italy
e-mail: bstroffo@unina.it

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_13

$W(\boldsymbol{\xi}, \det D\boldsymbol{\xi}, \operatorname{cof} D\boldsymbol{\xi})$ is polyconvex, and under suitable coercivity conditions, the existence of minimizers is well understood, see [2, 16].

From the physical point of view, the interpenetration of matter prescribes that two points cannot be mapped in the same one. Mathematically speaking, this problem can be formulated as the global invertibility of $\mathbf{u}$. A first global invertibility result was proven by Ball in the Sobolev space $W^{1,p}$, $p > n$ such that $\det D\mathbf{u} > 0$ and $\mathbf{u}$ coincides on $\partial\Omega$ with an invertible map $\mathbf{u}_0$, see [3].

In the case $p < n$, there are deformations in $W^{1,p}$ which present singularities and, in particular, are not continuous. One of such type of singularities is that of *cavitation*, that is, the formation of voids in solids. In this respect, Ball considered the minimization problem of a polyconvex energy among the restricted class of deformations that are radially symmetric; i.e.,

$$\mathbf{u}(\mathbf{x}) = \frac{r(|\mathbf{x}|)}{|\mathbf{x}|}\mathbf{x}$$

when Ω is the unit ball of $\mathbb{R}^n$. The prescribed boundary condition was the radial stretching $\mathbf{u}(\mathbf{x}) = \lambda\mathbf{x}$. He showed that there is a threshold $\lambda_c > 1$, such that the linear map is the unique minimizer for $1 \leq \lambda \leq \lambda_c$, whereas for $\lambda > \lambda_c$ there is a unique singular radial transformation with $r(0) > 0$. This means that it would be energetically favorable for the minimizer of the elastic energy to exhibit a cavitation.

Marcellini [34] revisited this example using the relaxation: the energy corresponding to a singular radial deformation must be defined through the lower semicontinuous envelope taken among all regular transformation $u_k(\mathbf{0}) = \mathbf{0}$ and with respect to the weak convergence. He derived a representation formula for the relaxed energy with an additional term proportional to the n-dimensional measure of the cavity, see also [12].

A further model was studied by Müller and Spector [36]. They were analyzing a counterexample that consists of a sequence of deformations that create more and more cavities. Consequently, they pointed out that such behavior could be prevented by including an extra term in the energy that penalizes the creation of new surface. More precisely, their elastic energy consists of a bulk term plus a constant multiple of the perimeter of the geometric image of the deformation, see Definition 6. In proving the weak continuity of the determinants, it is required to know that not only deformations of the sequence have to be one-to-one almost everywhere but also their limit. They constructed a counterexample where the weak limit of one-to-one almost everywhere maps with $\det D\mathbf{u}_j > 0$ satisfied $\det D\mathbf{u} > 0$, but it was not one-to-one almost everywhere. To overcome this difficulty, they introduced a new invertibility condition "INV", formulated in terms of degree for maps in $W^{1,p}(\Omega, \mathbb{R}^n)$, $p > n-1$. This condition prevents the possibility of creating cavities that are subsequently filled with matter from elsewhere, is stable with respect to weak convergence in $W^{1,p}$, $p > n-1$, and implies invertibility almost everywhere. Its formulation relies on the topological degree. Later, Conti and De Lellis [15] were relaxing this condition to maps in $W^{1,n-1} \bigcap L^\infty$ obtaining some partial results.

An alternative theory for cavitation, which also includes fracture, was given by Henao and Mora-Corral. They replaced the perimeter with a surface energy, see Eq. (19) in Sect. 5, and proved that finite surface energy implies SBV-regularity of a suitable defined inverse of $\mathbf{u}$. In addition, they specified a notion of surface created by $\mathbf{u}$ and gave a precise meaning to the idea that $\mathcal{E}(\mathbf{u})$ measures the area of this surface $\Gamma(\mathbf{u})$. Looking deeply into the above counterexample of Müller and Spector, they decompose the surface energy into the $\mathcal{H}^{n-1}$-measure of a visible and invisible part, this latter created from elsewhere in the body. Henao and Mora-Corral extensively studied Lusin (N) condition in connection with local invertibility, [21–24]. A key tool is the use of topological image $\mathrm{im}_{\mathrm{T}}(\mathbf{u}, \Omega)$ of $\mathbf{u}$, see Definition 5, which is defined as the set of points for which $\mathbf{u}$ has nonzero degree, coincide a.e. with the image of $\mathbf{u}$ and the union of cavities created. If $\mathbf{u} \in W^{1,p}$, $p > n-1$, and $\det D\mathbf{u} > 0$, the condition INV is satisfied, and the surface energy is finite, they were able to prove that the set $\mathrm{im}_{\mathrm{T}}(\mathbf{u}, \Omega)$ is open, and in this case, the generalized inverse belongs to $W^{1,1}(\mathrm{im}_{\mathrm{T}}(\mathbf{u}, \Omega), \mathbb{R}^n)$. In the limit case $p = n-1$, instead, their result states an SBV regularity of the inverse and that the jump set does not intersect $\mathrm{im}_{\mathrm{T}}(\mathbf{u}, \Omega)$.

Barchiesi et al. [7] were focused in defining a class of orientation-preserving maps that do not exhibit cavitation. The solely condition of equality between pointwise determinant and distributional determinant, $\det D\mathbf{u} = \mathrm{Det} D\mathbf{u}$, together with $\det D\mathbf{u} > 0$ was shown not to be sufficient for this requirement. In this case, the condition INV was not satisfied. In light of the aforementioned results, the surface energy (being 0) and the topological image were involved. We point out that the definition of degree is given only on selective "good" open subsets of Ω. The theorem states:

Theorem 1 *Let $p > n-1$, and suppose that $\mathbf{u} \in W^{1,p}(\Omega, \mathbb{R}^n)$ satisfies $\det D\mathbf{u} \in L^1_{loc}(\Omega)$. The following conditions are equivalent:*

- $\mathcal{E}(\mathbf{u}) = 0$ *and* $\det D\mathbf{u} > 0$ *a.e.*
- $(\mathrm{adj}\, D\mathbf{u})\mathbf{u} \in L^1_{loc}(\Omega, \mathbb{R}^n)$, $\det D\mathbf{u} \neq 0$ *for a.e.* $x \in \Omega$, $\mathrm{Det} D\mathbf{u} = \det D\mathbf{u}$, *and* $\deg(\mathbf{u}, B, \cdot) \geq 0$ *for all balls B for which* $\deg(\mathbf{u}, B, \cdot)$ *is defined.*

In addition, they were applying their result to prove an existence theorem for minimizers of a variational model where the elastic energy has two terms: one written in Lagrangian coordinates and the mechanical one in Eulerian coordinates.

In a previous paper [25], the second author with Henao proved that many properties of orientation-preserving maps, such as local invertibility and a.e. differentiability, can be pushed to a special class of Orlicz–Sobolev spaces, with an integrability exponent just above the space dimension minus one, in the logarithmic scale. In addition, they were showing that the maps considered in [7] were only weakly monotone. An important tool for relaxing the condition about integrability was the interplay between fine properties of Orlicz–Sobolev maps on manifolds of dimension $n-1$ and the n-absolute continuity introduced by Malý. Indeed, this condition is satisfied by a function $u \in W^{1,1}(\Omega)$ whenever its weak derivatives belong to the Lorentz space $L^{n,1}(\Omega)$, which in turn is formulated via an Orlicz integrability condition, see [13]. The theorem reads as follows:

Theorem 2 *Let* $A(t) = t^{n-1}\log^{\alpha}(e+t)$, $\alpha > n-2$, *and suppose that* $\mathbf{u} \in W^{1,A}(\Omega, \mathbb{R}^n)$ *satisfies* $\det D\mathbf{u} \in L^1_{\mathrm{loc}}(\Omega)$. *The following conditions are equivalent:*

- $\mathcal{E}(\mathbf{u}) = 0$ *and* $\det D\mathbf{u} > 0$ *a.e.*
- $(\operatorname{adj} D\mathbf{u})\mathbf{u} \in L^1_{\mathrm{loc}}(\Omega, \mathbb{R}^n)$, $\det D\mathbf{u} \neq 0$ *for a.e.* $x \in \Omega$, $\operatorname{Det} D\mathbf{u} = \det D\mathbf{u}$, *and* $\deg(\mathbf{u}, B, \cdot) \geq 0$ *for all balls* B *for which* $\deg(\mathbf{u}, B, \cdot)$ *is defined.*

This kind of generalization could have not only a mathematical interest per se but it is also related to questions of integrability of Jacobian determinants and mappings of finite distortion (see, e.g., [20, 27, 28, 41]).

The drawback of this generalization is an existence theorem for models of magnetic elastomers, liquid crystals, and magnetoelasticity, see, e.g., [6, 10, 31]. The existence theorems were proved in the scale of Sobolev spaces with $p > n-1$ in [7] and extended to our Sobolev–Orlicz class in [25]. Both the theorems were provided assuming polyconvexity in the mechanic energy and quadratic growth in the deformed configuration (nematic).

An existence theorem for the magnetoelastic model without any polyconvexity or quasiconvexity assumption was proven for the relaxed functional, the quasiconvex envelope in the same Sobolev–Orlicz class, see [39]. Actually, the quasiconvex envelope is the sum of the two envelopes: the quasiconvex for the mechanical and the tangential quasiconvexification for the nematic term (see [35] for the case t^p, $p > n-1$).

Our Result With this contribution, we are willing to push the previous result even further. To this aim, we got inspired by the paper by Henao et al. [26] where a global invertibility result was presented in the scale of Sobolev spaces. In particular, they were also revisiting the counterexamples of [36, Section 11], by showing that the creation of a cavitation or leakage to the boundary does not occur within their special class of maps: $\overline{\mathcal{A}}_p$. In our work, we extend the global invertibility result [26] to a class $\overline{\mathcal{A}}$ of orientation-preserving Orlicz–Sobolev maps with an integrability just above $n-1$, whose traces on the boundary are also Orlicz–Sobolev and which do not present cavitation in the interior or on the boundary. Namely, we consider deformations belonging to the Orlicz–Sobolev space $W^{1,A}(\Omega; \mathbb{R}^n) \cap W^{1,A}(\partial\Omega; \mathbb{R}^n)$, generated by the N-function $A(t)$ as in Theorem 2. We then apply these results to prove the existence of minimizers within (a suitable subclass of) $\overline{\mathcal{A}}$ for functionals often used as models in nonlinear elasticity, of the form

$$\int_{\Omega} W(\mathbf{x}, \mathbf{u}(\mathbf{x}), D\mathbf{u}(\mathbf{x}))\, \mathrm{d}\mathbf{x}\,,$$

where W is assumed to be polyconvex in the last variable.

We would like to mention the recent article of Krömer [30] where he raised the question of whether a continuous deformation is invertible on the boundary. He was working in the regime of $p \geq n$, and he was able to obtain the existence of homeomorphic minimizers under stronger assumptions. In our case, self-contact at the boundary is allowed, see [14].

Overview of the Chapter This chapter is organized as follows. In Sect. 2, we fix the main notation which will be used throughout the chapter. Section 3 collects some basic definitions and results concerning N-functions and the Orlicz–Sobolev spaces. In particular, in Sect. 3.1, we define traces of Orlicz–Sobolev functions. Then, with Sect. 4, we recall the notions of topological degree for Orlicz–Sobolev maps (Definition 4), of topological image of a set (Definition 5), and the concept of geometric image (Definition 6). The class of admissible deformations $\overline{\mathcal{A}}$ is introduced in Sect. 5, where we prove their fine properties (Sect. 5.3), in particular, boundedness (Proposition 6) and global invertibility (Proposition 4). In the last section, Sect. 6, we exploit the results of Sect. 5 to prove the existence of minimizers in $\overline{\mathcal{A}}$ for a class of functionals in nonlinear elasticity.

2 Notation

In this section, we fix the notation and introduce some definitions used in the chapter.

Throughout the paper, we will assume $n \geq 3$, because our Orlicz class makes sense only for $n > 2$, see Sect. 5. In all the chapter, Ω will be a non-empty open, bounded set of $\mathbb{R}^n$, which represents the body in its reference configuration. There, the coordinates will be denoted by $\mathbf{x}$, while in the deformed configuration by $\mathbf{y}$. Vector-valued and matrix-valued functions will be written in boldface. The closure of a set A is denoted by $\bar{A}$ and its topological boundary by ∂A. Given a square matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, its determinant is denoted by $\det \mathbf{M}$. The adjugate matrix $\operatorname{adj} \mathbf{M} \in \mathbb{R}^{n\times n}$ satisfies $(\det \mathbf{M})\mathbf{I} = \mathbf{M} \operatorname{adj} \mathbf{M}$, where $\mathbf{I}$ denotes the identity matrix. The transpose of $\operatorname{adj} \mathbf{M}$ is the cofactor $\operatorname{cof} \mathbf{M}$. We recall the identity

$$\mathbf{M} \operatorname{adj} \mathbf{M} = \operatorname{cof} \mathbf{M}\mathbf{M}^T = (\det \mathbf{M})\mathbf{I}\,. \tag{2}$$

If $\mathbf{M}$ is invertible, its inverse is denoted by $\mathbf{M}^{-1}$. The inner product of vectors and of matrices will be denoted by $\cdot$ and their associated norms are denoted by $\|\cdot\|$. Given $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, the tensor product $\mathbf{a} \otimes \mathbf{b}$ is the $n \times n$ matrix whose component (i, j) is $a_i\, b_j$. The set $\mathbb{R}^{n\times n}_+$ denotes the subset of matrices in $\mathbb{R}^{n\times n}$ with positive determinant. The set $\mathbb{S}^{n-1}$ denotes the unit sphere in $\mathbb{R}^n$.

The Lebesgue measure in $\mathbb{R}^n$ is denoted by $|\cdot|$ or $\mathcal{L}^n$ and the $(n-1)$-dimensional Hausdorff measure by $\mathcal{H}^{n-1}$. The abbreviation a.e. stands for *almost everywhere* or *almost every*; unless otherwise stated, it refers to $\mathcal{L}^n$. For Φ a Young function, L^Φ denotes the corresponding Orlicz space and $W^{1,\Phi}$, $W^{1,\Phi}_0$ the Orlicz–Sobolev spaces (see Sect. 3 for the precise definitions). The symbols C^1_c and C^∞_c stand for the spaces of C^1 and C^∞ functions, respectively, with compact support. The derivative of a Sobolev–Orlicz or a smooth vector-valued function $\mathbf{u}$ is written $D\mathbf{u}$.

The strong convergence in L^Φ or $W^{1,\Phi}$ and the a.e. convergence are denoted by $\to$, while the symbol for the weak convergence is $\rightharpoonup$ and that for the weak* convergence in L^∞ is $\overset{*}{\rightharpoonup}$.

3 Orlicz–Sobolev Spaces

We recall here few basic definitions and results concerning N-functions and Orlicz–Sobolev spaces. We refer the interested reader to [1, 9, 29, 32] for a detailed treatment of the topic.

An N-function A is a convex function from $[0,\infty)$ to $[0,\infty)$, which vanishes only at 0 and such that

$$\lim_{s\to 0^+}\frac{A(s)}{s}=0\,,\qquad \lim_{s\to\infty}\frac{A(s)}{s}=\infty.$$

If A is an N-function, then we denote by A^* the *Young–Fenchel–Yosida* dual or conjugate transform of A, namely, the N-function defined as

$$A^*(s):=\sup\{sr-A(r)\,:\,0<r<+\infty\}.$$

In this chapter, we restrict our analysis to functions A whose growth at infinity is at least such that

$$\int_{t_0}^{\infty}\left(\frac{t}{A(t)}\right)^{\frac{1}{n-2}}\mathrm{d}t<\infty \tag{3}$$

for some $t_0 \geq 0$. The condition is satisfied, in particular, when $A(t)=t^p$ for $p>n-1$ and when $A(t)=t^{n-1}\log^\alpha(e+t)$ for every $\alpha>n-2$.

An N-function A is said to satisfy the Δ_2-*condition near infinity* if it is finite-valued and there exist constants $\mu>2$ and $t_0>0$ such that

$$A(2t)\leq\mu A(t)\quad\text{for } t\geq t_0. \tag{4}$$

If (4) holds for every $t>0$, we say that A satisfies the Δ_2-*condition globally.*

Remark 1 We notice that our function $A(t)=t^{n-1}\log^\alpha(e+t)$ for every $\alpha>n-2$ verifies the Δ_2 condition together with its conjugate. We will also be dealing with the function $B(t)=t\log^\beta(e+t)$ for a $\beta>0$ (see Sect. 5): this function verifies the Δ_2-condition globally.

Let Ω be a measurable subset of $\mathbb{R}^n$. The Orlicz space $L^A(\Omega)$ built upon an N-function A is the Banach function space of those real-valued measurable functions u on Ω for which the Luxemburg norm

$$\|u\|_{L^A(\Omega)} := \inf\left\{\lambda > 0 : \int_\Omega A\left(\frac{|u(\mathbf{x})|}{\lambda}\right) \mathrm{d}\mathbf{x} \le 1\right\}$$

is finite.

Since A is non-decreasing,

$$\int_\Omega A(|u(\mathbf{x})|)\mathrm{d}\mathbf{x} < \infty \;\Rightarrow\; \|u\|_{L^A(\Omega)} \le 1. \tag{5}$$

If A satisfies the Δ_2-condition at infinity, then

$$u \in L^A(\Omega) \;\Leftrightarrow\; \int_\Omega A(|u(\mathbf{x})|)\mathrm{d}\mathbf{x} < \infty. \tag{6}$$

Proposition 1 (Generalized Hölder Inequality) *Let A be an N-function and A^* its dual. Then, it holds that*

$$\left|\int_\Omega u(\mathbf{x})v(\mathbf{x})\,\mathrm{d}\mathbf{x}\right| \le 2\|u\|_{L^A(\Omega)}\|v\|_{L^{A^*}(\Omega)},$$

for every $u \in L^A(\Omega)$ and $v \in L^{A^*}(\Omega)$.

Note that we may introduce another norm on $L^A(\Omega)$, the *Orlicz norm* or *dual norm*, defined as

$$|u|_A := \sup\left\{\int_\Omega u(\mathbf{x})v(\mathbf{x})\,\mathrm{d}\mathbf{x} :\; v \in L^{A^*}(\Omega),\, \|v\|_{L^{A^*}(\Omega)} \le 1\right\}.$$

The norms $\|\cdot\|_{L^A(\Omega)}$ and $|\cdot|_A$ are equivalent, since it holds that

$$\|u\|_{L^A(\Omega)} \le |u|_A \le 2\|u\|_{L^A(\Omega)}, \quad u \in L^A(\Omega).$$

The Orlicz space $L^A(\Omega, \mathbb{R}^n)$ of vector-valued measurable functions on Ω is defined as $L^A(\Omega, \mathbb{R}^n) = (L^A(\Omega))^n$ and is equipped with the norm $\|\mathbf{u}\|_{L^A(\Omega,\mathbb{R}^n)} = \|\,\|\mathbf{u}\|\,\|_{L^A(\Omega)}$ for $\mathbf{u} \in L^A(\Omega, \mathbb{R}^n)$. The Orlicz space $L^A(\Omega, \mathbb{R}^{n\times n})$ of matrix-valued measurable functions on Ω can be defined analogously.

We denote by $W^{1,A}(\Omega)$ the Orlicz–Sobolev space defined by

$$W^{1,A}(\Omega) := \{u \in L^A(\Omega) : u \text{ is weakly differentiable and } Du \in L^A(\Omega, \mathbb{R}^n)\}.$$

The space $W^{1,A}(\Omega)$, equipped with the norm

$$\|u\|_{W^{1,A}(\Omega)} := \|u\|_{L^A(\Omega)} + \|Du\|_{L^A(\Omega,\mathbb{R}^n)},$$

is a Banach space. The space $W_0^{1,A}(\Omega)$ is the closure of $C_c^\infty(\Omega)$ in the $W^{1,A}$ norm.

The Orlicz space $W^{1,A}(\Omega, \mathbb{R}^n)$ of vector-valued measurable functions on Ω is defined as $W^{1,A}(\Omega, \mathbb{R}^n) = (W^{1,A}(\Omega))^n$ and is equipped with the norm $\|\mathbf{u}\|_{W^{1,A}(\Omega,\mathbb{R}^n)} = \|\mathbf{u}\|_{L^A(\Omega,\mathbb{R}^n)} + \|D\mathbf{u}\|_{L^A(\Omega,\mathbb{R}^{n\times n})}$ for $\mathbf{u} \in W^{1,A}(\Omega, \mathbb{R}^n)$. The analogous spaces for matrix-valued functions are defined in the same way.

We now introduce a notion of ordering for Young functions (see, e.g., [32, Definition 3.5.6]).

Let A and B be Young functions. A is said to *dominate* B, and we write $B \prec A$, if there exists a positive constant c_0 such that

$$B(t) \leq A(c_0 t), \quad \text{for every } t > 0. \tag{7}$$

As customary, if there exists also $t_0 > 0$ such that (7) holds for every $t \geq t_0$, we say that A *dominates* B *near infinity*. If $A \prec B$ and $B \prec A$, the functions A and B are said to be *equivalent*, and we write $A \sim B$.

Let $p \geq 1$. The Orlicz–Sobolev space generated by the N-function $t^p \log^\alpha(e+t)$ is the *Zygmund space* $\mathrm{L}^p\mathrm{Log}^\alpha\mathrm{L}$. The space $\mathrm{L}^1\mathrm{Log}^\alpha\mathrm{L}$ will be denoted by $\mathrm{LLog}^\alpha\mathrm{L}$. If $\alpha < 0$, the equivalent notation $\dfrac{\mathrm{L}^p}{\mathrm{Log}^{-\alpha}\mathrm{L}}$ will be used. Since the conjugate N-function of $t^p \log^\alpha(e+t)$, $t \geq 0$, $p > 1$, is equivalent to $t^{p'} \log^{-\alpha\frac{p'}{p}}(e+t)$, where $p' := \frac{p}{p-1}$ (see, e.g., [29, Theorem 7.2]), we have that the dual space of $\mathrm{L}^{n-1}\mathrm{Log}^\alpha\mathrm{L}$, $\alpha > n-2$, is the Zygmund space $\dfrac{\mathrm{L}^{\frac{n-1}{n-2}}}{\mathrm{Log}^{\frac{\alpha}{n-2}}\mathrm{L}}$. Furthermore, by virtue of [8, Theorem 9.1], it holds that

$$L^q \subseteq \frac{\mathrm{L}^{\frac{n-1}{n-2}}}{\mathrm{Log}^{\frac{\alpha}{n-2}}\mathrm{L}}, \quad \text{for every } q > \frac{n-1}{n-2}. \tag{8}$$

We recall that a family of functions $\mathcal{F}$ has equi-absolutely continuous integrals if for every $\varepsilon > 0$ one can find $\delta > 0$ such that for all $u \in \mathcal{F}$ there holds $\int_E |u(\mathbf{x})| \, \mathrm{d}\mathbf{x} < \varepsilon$ provided $|E| < \delta$. A general criterion for the equi-absolute continuity of the integrals of a family of functions in $L^A(\Omega)$ is given by the following version of *De la Vallée Poussin's Theorem* (see, e.g., [29, Ch. II, §11.1]):

Theorem 3 *Let A be an N-function and $\mathcal{F}$ be a family of functions in $L^A(\Omega)$. If there exists $C > 0$ such that*

$$\int_\Omega A(|u(\mathbf{x})|) \, \mathrm{d}\mathbf{x} \leq C, \quad u \in \mathcal{F},$$

then the family $\mathcal{F}$ has equi-absolutely continuous integrals.

Let $\{\mathbf{v}_j\}$ be a sequence of functions in $L^A(\Omega, \mathbb{R}^n)$, and let $\mathbf{v} \in L^A(\Omega, \mathbb{R}^n)$. If A is Δ_2 near infinity, then

$$\lim_{j\to+\infty} \|\mathbf{v}_j - \mathbf{v}\|_{L^A(\Omega,\mathbb{R}^n)} = 0 \Leftrightarrow \lim_{j\to+\infty} \int_\Omega A(\|\mathbf{v}_j - \mathbf{v}\|)\,\mathrm{d}\mathbf{x} = 0.$$

Note that, if A does not satisfy Δ_2-condition, the implication "$\Leftarrow$" fails. If $A \in \Delta_2$ near infinity, instead, we have

$$\lim_{j\to+\infty} \|\mathbf{v}_j - \mathbf{v}\|_{L^A(\Omega,\mathbb{R}^n)} = 0 \Rightarrow \lim_{j\to+\infty} \int_\Omega A(\|\mathbf{v}_j\|)\,\mathrm{d}\mathbf{x} = \int_\Omega A(\|\mathbf{v}\|)\,\mathrm{d}\mathbf{x}.$$

3.1 Traces

We define Orlicz–Sobolev functions on the (sufficiently smooth) boundary $\partial\Omega$ of Ω following the approach of [32, Section 6].

First, we recall the definition of open set of class $C^{k,\alpha}$. Since the minimum regularity for Ω will be Lipschitz, we are assuming that $k + \alpha \geq 1$.

Definition 1 Let $k \geq 0$ be an integer and $\alpha \in [0, 1]$ be such that $k + \alpha \geq 1$. A bounded open set Ω is said to be of class $C^{k,\alpha}$ if there exist $r > 0$, $b > 0$, $m \in \mathbb{N}$, $a_1, \dots, a_m \in C^{k,\alpha}([0, r]^{n-1})$ and $\mathbf{M}_1, \dots, \mathbf{M}_m$ proper rigid transformations in $\mathbb{R}^n$ such that, setting

$$\begin{aligned}
\Gamma_i &:= \mathbf{M}_i^{-1}(\{(\hat{\mathbf{x}}, x_n) \in (0, r)^{n-1} \times \mathbb{R} :\ x_n = a_i(\hat{\mathbf{x}})\}),\\
U_i^+ &:= \mathbf{M}_i^{-1}(\{(\hat{\mathbf{x}}, x_n) \in (0, r)^{n-1} \times \mathbb{R} :\ a_i(\hat{\mathbf{x}}) < x_n < a_i(\hat{\mathbf{x}}) + b\}),\\
U_i^- &:= \mathbf{M}_i^{-1}(\{(\hat{\mathbf{x}}, x_n) \in (0, r)^{n-1} \times \mathbb{R} :\ a_i(\hat{\mathbf{x}}) - b < x_n < a_i(\hat{\mathbf{x}})\}),
\end{aligned}$$

we have that

$$\partial\Omega = \bigcup_{i=1}^m \Gamma_i, \quad \bigcup_{i=1}^m U_i^+ \subset \Omega \quad \text{and} \quad \bigcup_{i=1}^m U_i^- \subset \mathbb{R}^n \backslash \bar{\Omega}.$$

For each $i = 1, \dots, m$, the set Γ_i is relatively open in $\partial\Omega$ and the sets U_i^+, U_i^- are open. We denote by U_i the open set given by $U_i^+ \cup \Gamma_i \cup U_i^-$ for every $i = 1, \dots, m$. Then, the family $\{U_i\}_{i=1}^m$ is an open cover of $\partial\Omega$. Furthermore, we consider an open set $U_0 \subset\subset \Omega$ such that $\bar{\Omega} \subset \bigcup_{i=0}^m U_i$.

For each $i = 1, \dots, m$, we define $\mathbf{P}_i : [0, r]^{n-1} \times [-b, b] \to \mathbb{R}^n$ by $\mathbf{P}_i(\hat{\mathbf{x}}, x_n) := (\hat{\mathbf{x}}, a_i(\hat{\mathbf{x}}) + x_n)$, and $\mathbf{N}_i := \mathbf{M}_i^{-1} \circ \mathbf{P}_i$. Then, each $\mathbf{N}_i$ is injective, and we have $\mathbf{N}_i((0,r)^{n-1} \times (-b,b)) = U_i$, $\mathbf{N}_i((0,r)^{n-1} \times (0,b)) = U_i^+$, $\mathbf{N}_i((0,r)^{n-1} \times [0,b)) = \Gamma_i \cup U_i^+$, and $\mathbf{N}_i((0,r)^{n-1} \times \{0\}) = \Gamma_i$.

Denoting by $\pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$ the projection onto the first $n-1$ coordinates and by $\eta : \mathbb{R}^{n-1} \to \mathbb{R}^n$ the map $\eta(\hat{\mathbf{x}}) = (\hat{\mathbf{x}}, 0)$, for any function u defined on Γ_i, we consider the map $L_i(u) := \pi(\mathbf{N}_i^{-1}(\Gamma_i)) \to \mathbb{R}$ defined by $L_i(u) := u \circ \mathbf{N}_i \circ \eta$.

We are now in position to give the definition of an Orlicz–Sobolev function defined on the boundary. Assume that Ω is a Lipschitz open set.

Definition 2 We denote by $W^{1,A}(\partial\Omega)$ the set of functions $u : \partial\Omega \to \mathbb{R}$ such that $L_i(u) \in W^{1,A}((0,r)^{n-1})$ for all $i \in \{1,\dots,m\}$, equipped with the norm

$$\|u\|_{W^{1,A}(\partial\Omega)} := \sum_{i=1}^{m} \|L_i(u)\|_{W^{1,A}((0,r)^{n-1})} .$$

An analogous definition can be given, with minor modifications, for the space $W^{1,A}(\Gamma_i)$ of Orlicz–Sobolev functions defined on the subset Γ_i. Moreover, it can be shown that $W^{1,A}(\partial\Omega)$ is a Banach space (see, e.g., [32, Section 6.3.6]), and this property does not depend on the description of the boundary considered in Definition 1.

Let $u \in W^{1,A}(\Omega)$. We denote by $u|_{\partial\Omega}$ the trace of u on $\partial\Omega$, which belongs to $L^A(\partial\Omega)$. With abuse of notation, we will write $u \in W^{1,A}(\partial\Omega)$ when the trace of u belongs to $W^{1,A}(\partial\Omega)$, following the definition given in 2. We then define the intersection space

$$W^{1,A}(\Omega) \cap W^{1,A}(\partial\Omega) := \{u \in W^{1,A}(\Omega) :\ u|_{\partial\Omega} \in W^{1,A}(\partial\Omega)\} .$$

It is equipped in a natural way with the norm of an intersection

$$\|u\|_{W^{1,A}(\Omega)\cap W^{1,A}(\partial\Omega)} := \|u\|_{W^{1,A}(\Omega)} + \|u\|_{W^{1,A}(\partial\Omega)} ,$$

and it can be easily shown that it is a Banach space. For Γ a relatively open subset of $\partial\Omega$, the notation $W^{1,A}(\Gamma)$ and $W^{1,A}(\Omega) \cap W^{1,A}(\Gamma)$ will be used. For vector-valued functions, the symbol $W^{1,A}(\Omega;\mathbb{R}^n) \cap W^{1,A}(\partial\Omega;\mathbb{R}^n)$ will be adopted.

In order to extend an Orlicz–Sobolev function defined on $(0,r)^{n-1} \times \{0\}$ to $(0,r)^{n-1} \times (0,b)$, we will extend putting the same value on the vertical fiber. Namely, first, we will project onto the first $n-1$ coordinates, and then we will compose with the map η that leaves the last coordinate fixed (equal to 0). The proof can be obtained as in [26, Lemma 4.1] dealing with the $W^{1,p}$ case.

Lemma 1 *Let $r, b > 0$, and set $D := (0,r)^{n-1} \times (0,b)$ and $\Gamma := (0,r)^{n-1} \times \{0\}$. Then, the map $E : W^{1,A}(\Gamma) \to W^{1,A}(D)$ defined by $Eu := u \circ \eta \circ \pi$ is linear and bounded. Furthermore,*

$$\frac{\partial(Eu)}{\partial x_i} = \frac{\partial u}{\partial x_i} \circ \eta \circ \pi , \quad \textit{for } i = 1,\dots,n-1 , \quad \textit{and} \quad \frac{\partial(Eu)}{\partial x_n} = 0 . \tag{9}$$

In addition, $(Eu)_{|\Gamma} = u$.

Proof Defining $\widetilde{u} := Eu$, by Fubini's theorem, we have

$$\int_D A(|\widetilde{u}(\mathbf{x})|)\,\mathrm{d}\mathbf{x} = b\int_{(0,r)^{n-1}} A(|u(\hat{\mathbf{x}},0)|)\,\mathrm{d}\hat{\mathbf{x}}, \tag{10}$$

which implies $\widetilde{u} \in L^A(D)$. Now, choosing a test function $\varphi \in C_c^1(D)$, for each $i \in \{1,\ldots,n-1\}$, a simple integration by parts gives

$$\int_D \widetilde{u}(\mathbf{x})\frac{\partial\varphi}{\partial x_i}(\mathbf{x})\,\mathrm{d}\mathbf{x} = -\int_D \frac{\partial u}{\partial x_i}(\eta(\pi(\mathbf{x})))\varphi(\mathbf{x})\,\mathrm{d}\mathbf{x}\,, \tag{11}$$

while for $i = n$

$$\int_D \widetilde{u}(\mathbf{x})\frac{\partial\varphi}{\partial x_n}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{(0,r)^{n-1}} u(\hat{\mathbf{x}},0)\int_0^b \frac{\partial\varphi}{\partial x_n}(\hat{\mathbf{x}},x_n)\,\mathrm{d}x_n\,\mathrm{d}\hat{\mathbf{x}} = 0\,. \tag{12}$$

Thus, (9) holds. Since an analog of (10) holds also for $\frac{\partial\widetilde{u}}{\partial x_i}$, $i = 1,\ldots,n-1$, we conclude that $\widetilde{u} \in W^{1,A}(D)$ and that the map E is linear and bounded. The last assertion $\widetilde{u}_{|\Gamma} = u$ follows from the continuity of the trace operator. □

The previous result is a tool for the proof of the following density result of smooth functions in $W^{1,A}(\Omega) \cap W^{1,A}(\partial\Omega)$.

Proposition 2 *Let $k \geq 0$ be an integer and $\alpha \in [0,1]$ be such that $k+\alpha \geq 1$. Let Ω be an open and bounded set with $C^{k,\alpha}$ boundary. Then, $C^{k,\alpha}(\bar{\Omega})$ is dense in $W^{1,A}(\Omega) \cap W^{1,A}(\partial\Omega)$.*

Proof The strategy of [26, Proposition 4.2], based on the analogous of Lemma 1 and the result of Fonseca and Malý [19, Lemma 2.4], which allows to modify the boundary values of a function without increasing significantly its norm, can be performed in the Orlicz setting with minor modifications. The proof is based on a gluing lemma for functions defined on disjoint subsets, a partition of unity and triangle inequality for the norm. □

As a final remark, we notice that $u \in W^{1,A}(\partial\Omega)$, for an N-function A complying with assumption (3) and Ω of class C^1, admits a continuous representative (see [11, Remark 3.2]) on $n-1$ manifolds. If not stated otherwise, we will always assume that u itself is the continuous representative.

4 Some Definitions and Preliminary Results

This section collects some basic definitions and preliminary results.

Let $\mathbf{u} : \Omega \longrightarrow \mathbb{R}^n$ be a measurable function, and let $\mathbf{x}_0 \in \Omega$. If $\mathbf{u}$ is *approximately differentiable* at $\mathbf{x}_0$, we denote by $\nabla\mathbf{u}(\mathbf{x}_0)$ its approximate differential at $\mathbf{x}_0$. We denote the set of approximate differentiability points of $\mathbf{u}$ by Ω_d. If $\mathbf{u}$ is approximately differentiable a.e., for any $E \subset \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, we define

$$\mathcal{N}_{\mathbf{u},E}(\mathbf{y}) := \mathcal{H}^0(\{\mathbf{x} \in \Omega_d \cap E : \mathbf{u}(\mathbf{x}) = \mathbf{y}\}). \tag{13}$$

The number $\mathcal{N}_{\mathbf{u},\Omega}$ will be denoted by $\mathcal{N}_{\mathbf{u}}$.

Now, we recall the definition of *almost everywhere (a.e.) invertibility* for a vector-valued function.

Definition 3 A function $\mathbf{u} : \Omega \longrightarrow \mathbb{R}^n$ is said to be one-to-one a.e. in a subset $E \subset \Omega$ if there exists a subset $N \subset E$, with $\mathcal{L}^n(N) = 0$, such that $\mathbf{u}_{|_{E\setminus N}}$ is one-to-one.

Since we are assuming Ω to be Lipschitz, for $\mathcal{H}^{n-1}$-a.e. $\mathbf{x} \in \partial\Omega$ the tangent space of $\partial\Omega$ at $\mathbf{x}$, denoted by $T_{\mathbf{x}}\partial\Omega$, and the unit exterior normal $\boldsymbol{\nu}(\mathbf{x})$ to Ω at $\mathbf{x}$ are defined.

Let V be an $(n-1)$-dimensional subspace of $\mathbb{R}^n$, and let $\mathbf{L} : V \to \mathbb{R}^n$ be a linear map. We denote by $\Lambda_{n-1}V$ the space of all alternating $(n-1)$-tensors on V and by $\Lambda_{n-1}\mathbf{L} : \Lambda_{n-1}V \to \mathbb{R}^n$ the transformation defined by

$$(\Lambda_{n-1}\mathbf{L})(\mathbf{a}_1 \wedge \cdots \wedge \mathbf{a}_{n-1}) = \mathbf{L}\mathbf{a}_1 \wedge \cdots \wedge \mathbf{L}\mathbf{a}_{n-1}\,, \quad \mathbf{a}_1, \ldots, \mathbf{a}_{n-1} \in V\,,$$

where $\wedge$ indicates the exterior product between vectors in $\mathbb{R}^n$. It is well known that the one-dimensional space $\Lambda_{n-1}V$ can be identified in a canonical way with the subspace generated by $\mathbf{v}$, $\mathbf{v}$ being any of the two unit normal vectors to V. Thus, the linear transformation $\Lambda_{n-1}\mathbf{L}$ is determined by the value $(\Lambda_{n-1}\mathbf{L})\mathbf{v}$, and the identity

$$(\Lambda_{n-1}\mathbf{L})\mathbf{v} = (\operatorname{cof}\widetilde{\mathbf{L}})\mathbf{v} \tag{14}$$

holds whenever $\widetilde{\mathbf{L}} : \mathbb{R}^n \to \mathbb{R}^n$ is any linear map extending $\mathbf{L}$. Let $\mathbf{u} \in W^{1,A}(\partial\Omega;\mathbb{R}^n)$. Then, the tangential derivative $D\mathbf{u}(\mathbf{x}) : T_{\mathbf{x}}\partial\Omega \to \mathbb{R}^n$ exists for a.e. $\mathbf{x} \in \partial\Omega$ and $|D\mathbf{u}| \in L^A(\partial\Omega)$. As a consequence, $(\Lambda_{n-1}D\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})$ exists for a.e. $\mathbf{x} \in \partial\Omega$, and $(\Lambda_{n-1}D\mathbf{u})\boldsymbol{\nu} \in \mathrm{LLog}^{\frac{\alpha}{n-1}}\mathrm{L}(\partial\Omega,\mathbb{R}^n)$.

4.1 Degree for Orlicz–Sobolev Maps, Topological Image of a Set, and Geometric Image of a Set

In order to introduce the concept of *topological image* (according to Šverák [40] (see also [36])), we need to recall the notion of topological degree for continuous functions (see, e.g., [17, 18]).

In the end of Sect. 3.1, we have recalled that every map $\mathbf{u} \in W^{1,A}(\partial\Omega,\mathbb{R}^n)$, where Ω is an open set of class C^1, and A satisfies (3) and the Δ_2-condition at infinity, admits a continuous representative $\bar{\mathbf{u}} : \partial\Omega \longrightarrow \mathbb{R}^n$. It can be extended to a continuous function $\tilde{\mathbf{u}} : \overline{\Omega} \longrightarrow \mathbb{R}^n$ (see, e.g., [38, Theorem 35.1]), and therefore, the following definition of degree can be given.

Definition 4 The degree $\deg(\bar{\mathbf{u}}, \Omega, \cdot) : \mathbb{R}^n \setminus \bar{\mathbf{u}}(\partial\Omega) \to \mathbb{Z}$ of $\bar{\mathbf{u}}$ on U is defined as the degree $\deg(\tilde{\mathbf{u}}, \Omega, \cdot) : \mathbb{R}^n \setminus \bar{\mathbf{u}}(\partial\Omega) \to \mathbb{Z}$ of $\tilde{\mathbf{u}}$ on Ω.

We will denote by $\deg(\mathbf{u}, \Omega, \cdot)$ the degree of $\mathbf{u} \in W^{1,A}(\partial\Omega, \mathbb{R}^n)$, with a slight abuse of notation, tacitly referring to the degree of its continuous representative.

We are now in a position to define the concept of topological image.

Definition 5 Let A be an N-function satisfying (3), and let $\Omega \subset\subset \mathbb{R}^n$ be a non-empty open set with a C^1 boundary. If $\mathbf{u} \in W^{1,A}(\partial\Omega, \mathbb{R}^n)$, we define $\operatorname{im}_{\mathrm{T}}(\mathbf{u}, \Omega)$, the topological image of Ω under $\mathbf{u}$, as the set of $\mathbf{y} \in \mathbb{R}^n \setminus \mathbf{u}(\partial\Omega)$ such that $\deg(\mathbf{u}, \Omega, \mathbf{y}) \neq 0$.

The continuity of function $\deg(\mathbf{u}, \Omega, \cdot)$ implies that the set $\operatorname{im}_{\mathrm{T}}(\mathbf{u}, \Omega)$ is open and $\partial \operatorname{im}_{\mathrm{T}}(\mathbf{u}, \Omega) \subset \mathbf{u}(\partial\Omega)$. Furthermore, as $\deg(\mathbf{u}, \Omega, \cdot) = 0$ in the unbounded component of $\mathbb{R}^n \setminus \mathbf{u}(\partial\Omega)$ (see, e.g., [17, Sect. 5.1]), it follows that $\operatorname{im}_{\mathrm{T}}(\mathbf{u}, \Omega)$ is bounded.

The following formula for the distributional derivative of the degree of Orlicz–Sobolev functions will be widely used (see [25, Proposition 2.12]).

Proposition 3 *Let A be an N-function satisfying (3) and the Δ_2-condition at infinity. Let $\Omega \subset \mathbb{R}^n$ be an open set of class C^1. Assume that $\mathbf{u}$ is the continuous representative of a function in $W^{1,A}(\Omega, \mathbb{R}^n)$. Then, for all $\mathbf{g} \in C^1(\mathbb{R}^n, \mathbb{R}^n)$,*

$$\int_{\partial\Omega} \mathbf{g}(\mathbf{u}(\mathbf{x})) \cdot ((\Lambda_{n-1} D\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x}))\, \mathrm{d}\mathcal{H}^{n-1}(\mathbf{x}) = \int_{\mathbb{R}^n} \operatorname{div} \mathbf{g}(\mathbf{y})\, \deg(\mathbf{u}, \Omega, \mathbf{y})\, \mathrm{d}\mathbf{y},$$

where $\boldsymbol{\nu}$ is the unit outward normal to Ω.

The following is the notion of *geometric image* of a set adapted to the context of Orlicz spaces (see [25, Section 2.2]).

Definition 6 Let $\mathbf{u} \in W^{1,A}(\Omega, \mathbb{R}^n)$, and assume that $\det D\mathbf{u}(\mathbf{x}) \neq 0$ for a.e. $\mathbf{x} \in \Omega$. Let Ω_0 be the subset of $\mathbf{x} \in \Omega$ where the following are satisfied:

(i) $\mathbf{u}$ is approximately differentiable at $\mathbf{x}$ and $\det \nabla\mathbf{u}(\mathbf{x}) \neq 0$.
(ii) There exist $\mathbf{w} \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ and a compact set $K \subset \Omega$ of density 1 at $\mathbf{x}$ such that $\mathbf{u}_{|K} = \mathbf{w}_{|K}$ and $\nabla\mathbf{u}_{|K} = D\mathbf{w}_{|K}$.

The geometric image of Ω under $\mathbf{u}$ is defined as

$$\operatorname{im}_{\mathrm{G}}(\mathbf{u}, \Omega) := \mathbf{u}(\Omega_0). \tag{15}$$

It turns out that Ω_0 is a set of full measure in Ω (see the remarks after [25, Def. 2.4]).

5 The Class of Admissible Functions

From now on, we fix as N-function A satisfying (3) and the Δ_2-condition at infinity (4) the function $A(t) := t^{n-1}\log^{\alpha}(e+t)$ for $\alpha > n-2$.

We start introducing the following class $\mathcal{A}(\Omega)$.

Definition 7 Let $\mathbf{u} \in W^{1,A}(\Omega, \mathbb{R}^n)$, and assume that $\det D\mathbf{u} \in L^1(\Omega)$. For every $\mathbf{f} \in C_c^1(\Omega \times \mathbb{R}^n, \mathbb{R}^n)$, we define

$$\bar{\mathcal{E}}_{\Omega}(\mathbf{u}, \mathbf{f}) := \int_{\Omega} [\operatorname{cof} D\mathbf{u}(\mathbf{x}) \cdot D\mathbf{f}(\mathbf{x}, \mathbf{u}(\mathbf{x})) + \det D\mathbf{u}(\mathbf{x}) \operatorname{div} \mathbf{f}(\mathbf{x}, \mathbf{u}(\mathbf{x}))] \,\mathrm{d}\mathbf{x}\,. \tag{16}$$

We define $\mathcal{A}(\Omega)$ as the set of $\mathbf{u} \in W^{1,A}(\Omega, \mathbb{R}^n)$ such that $\det D\mathbf{u} \in L^1(\Omega)$ and

$$\bar{\mathcal{E}}_{\Omega}(\mathbf{u}, \mathbf{f}) = 0 \quad \text{for all } \mathbf{f} \in C_c^1(\Omega \times \mathbb{R}^n, \mathbb{R}^n)\,. \tag{17}$$

Remark 2 We notice that if $\mathbf{u} \in W^{1,A}(\Omega, \mathbb{R}^n)$, then $D\mathbf{u} \in L^A(\Omega, \mathbb{R}^{n\times n})$, so $\operatorname{cof} D\mathbf{u} \in \mathrm{LLog}^{\frac{\alpha}{n-1}}\mathrm{L}(\Omega, \mathbb{R}^{n\times n})$. In particular, $\operatorname{cof} D\mathbf{u} \in L^1_{\mathrm{loc}}(\Omega, \mathbb{R}^{n\times n})$. This implies that the energy (16) is finite.

In Eq. (16), $D\mathbf{f}(\mathbf{x}, \mathbf{y})$ denotes the derivative of $\mathbf{f}(\cdot, \mathbf{y})$ evaluated at $\mathbf{x}$, while $\operatorname{div} \mathbf{f}(\mathbf{x}, \mathbf{y})$ is the divergence of $\mathbf{f}(\mathbf{x}, \cdot)$ evaluated at $\mathbf{y}$.

The energy $\bar{\mathcal{E}}_{\Omega}(\mathbf{u})$ was introduced in [21] and measures the new surface in the deformed configuration created by $\mathbf{u}$. For our purposes, we are interested into deformations $\mathbf{u}$ such that $\bar{\mathcal{E}}_{\Omega}(\mathbf{u}) = 0$; i.e., that do not exhibit cavitation.

A global version of condition (17) is the following (20), leading to the introduction of the class $\overline{\mathcal{A}}(\Omega)$.

Definition 8 Given $\mathbf{u} \in W^{1,A}(\Omega, \mathbb{R}^n) \cap W^{1,A}(\partial\Omega, \mathbb{R}^n)$ and $\mathbf{f} \in C_c^1(\bar{\Omega} \times \mathbb{R}^n, \mathbb{R}^n)$, we define

$$\mathcal{F}_{\partial\Omega}(\mathbf{u}, \mathbf{f}) := \int_{\partial\Omega} \mathbf{f}(\mathbf{x}, \mathbf{u}(\mathbf{x})) \cdot ((\Lambda_{n-1} D\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x}))\,\mathrm{d}\mathcal{H}^{n-1}(\mathbf{x})\,.$$

Then, we define $\overline{\mathcal{A}}(\Omega)$ as the class of $\mathbf{u} \in W^{1,A}(\Omega, \mathbb{R}^n) \cap W^{1,A}(\partial\Omega, \mathbb{R}^n)$ with $\det D\mathbf{u} \in L^1(\Omega)$ such that

$$\bar{\mathcal{E}}_{\Omega}(\mathbf{u}, \mathbf{f}) = \mathcal{F}_{\partial\Omega}(\mathbf{u}, \mathbf{f}) \quad \text{for all } \mathbf{f} \in C_c^1(\bar{\Omega} \times \mathbb{R}^n, \mathbb{R}^n)\,. \tag{18}$$

Taking into account the density in $C_c^1(\Omega \times \mathbb{R}^n, \mathbb{R}^n)$ of sums of functions of separate variables (see [33, Corollary 1.6.5]), conditions (17) and (18) can be rephrased, respectively, as follows:

$$\mathcal{E}_{\Omega}(\mathbf{u}, \phi, \mathbf{g}) := \int_{\Omega} \big[\operatorname{cof} D\mathbf{u}(\mathbf{x}) \cdot (\mathbf{g}(\mathbf{u}(\mathbf{x})) \otimes D\phi(\mathbf{x})) + \det D\mathbf{u}(\mathbf{x})\phi(\mathbf{x})\operatorname{div}\mathbf{g}(\mathbf{u}(\mathbf{x}))\big]\,\mathrm{d}\mathbf{x} = 0\,, \tag{19}$$

for all $\phi \in C_c^1(\Omega)$ and $\mathbf{g} \in C_c^1(\mathbb{R}^n, \mathbb{R}^n)$, and

$$\bar{\mathcal{E}}_\Omega(\mathbf{u}, \phi\mathbf{g}) = \mathcal{F}_{\partial\Omega}(\mathbf{u}, \phi\mathbf{g}) \quad \text{for all } \phi \in C_c^1(\bar{\Omega}) \text{ and } \mathbf{g} \in C_c^1(\mathbb{R}^n, \mathbb{R}^n)\,, \tag{20}$$

where $\phi\mathbf{g} \in C_c^1(\bar{\Omega} \times \mathbb{R}^n, \mathbb{R}^n)$ stands for the function $(\phi\mathbf{g})(\mathbf{x}, \mathbf{y}) := \phi(\mathbf{x})\mathbf{g}(\mathbf{y})$.

5.1 Extension Properties

As a first feature of the class just introduced, we notice that every function in $\overline{\mathcal{A}}(\Omega)$ can be extended to an open set $\widetilde{\Omega} \supset \bar{\Omega}$ by a function in $\mathcal{A}(\widetilde{\Omega})$.

In fact, we need to assume that Ω is an *extendable domain*.

Definition 9 An open set Ω is said to be *extendable* if it is bounded and has a Lipschitz boundary, and there exist a set N, with $\partial\Omega \subset N \subset \mathbb{R}^n \backslash \Omega$, a $\delta > 0$ and a bi-Lipschitz homeomorphism $\mathbf{w} : \partial\Omega \times (-\delta, 0] \to N$ onto N such that $\mathbf{w}(\mathbf{x}, 0) = \mathbf{x}$ for all $\mathbf{x} \in \partial\Omega$.

It is easy to see that the assumption of piecewise $C^{1,1}$ implies Ω extendable. In addition, the set $\Omega \cup N$ is open (see the remarks below [26, Definition 6.2]).

We start by stating a technical lemma, whose proof can be easily obtained by Lemma 1 as in [26, Lemma 6.1].

Lemma 2 *Let $r, b > 0$, and set $D := (0, r)^{n-1} \times (0, b)$ and $\Gamma := (0, r)^{n-1} \times \{0\}$. Then, the map $E : W^{1,A}(\Gamma, \mathbb{R}^n) \to W^{1,A}(D, \mathbb{R}^n)$ defined by $E\mathbf{u} := \mathbf{u} \circ \eta \circ \pi$ is linear and bounded. Moreover,* $\det D(E\mathbf{u}) = 0$ *and $(E\mathbf{u})_{|_\Gamma} = \mathbf{u}$. If, in addition, $(\Lambda_{n-1} D\mathbf{u})\mathbf{e}_n \in L^q(\Gamma, \mathbb{R}^n)$ for some $q \geq 1$, then* $\operatorname{cof} D(E\mathbf{u}) \in L^q(D; \mathbb{R}^{n\times n})$ *and*

$$\|\operatorname{cof} D(E\mathbf{u})\|_{L^q(D;\mathbb{R}^{n\times n})} = b^{1/q} \|(\Lambda_{n-1} D\mathbf{u})\mathbf{e}_n\|_{L^q(\Gamma,\mathbb{R}^n)}\,.$$

The main result of extension is contained in the following proposition.

Proposition 4 *Let Ω be an extendable open set. Then, there exist an open set $\widetilde{\Omega} \supset \bar{\Omega}$ and a linear bounded operator $E : W^{1,A}(\Omega; \mathbb{R}^n) \cap W^{1,A}(\partial\Omega; \mathbb{R}^n) \to W^{1,A}(\widetilde{\Omega}; \mathbb{R}^n)$ such that $E\mathbf{u} = \mathbf{u}$ a.e. in Ω,* $\det D(E\mathbf{u}) = 0$ *a.e. in $\widetilde{\Omega} \backslash \Omega$, and the following hold:*

(*i*) *If $(\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu} \in L^q(\partial\Omega, \mathbb{R}^n)$ for some $q \geq 1$, then* $\operatorname{cof} D(E\mathbf{u}) \in L^q(\widetilde{\Omega}\backslash\Omega; \mathbb{R}^{n\times n})$ *and*

$$\|\operatorname{cof} D(E\mathbf{u})\|_{L^q(\widetilde{\Omega}\backslash\Omega;\mathbb{R}^{n\times n})} \leq C\|(\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu}\|_{L^q(\partial\Omega,\mathbb{R}^n)}$$

for some constant $C > 0$ independent of $\mathbf{u}$.

(*ii*) $\mathbf{u} \in \overline{\mathcal{A}}(\Omega)$ *if and only if* $E\mathbf{u} \in \mathcal{A}(\widetilde{\Omega})$.

Proof We may follow the argument of [26, Proposition 6.3]. Therefore, we just provide a sketch of the proof.

Let $\mathbf{w}$ and N be as in Definition 9, and recall the notation of Sect. 3.1. We denote by $\widetilde{\pi} : \partial\Omega \times (-\delta, 0] \to \partial\Omega$ the projection onto $\partial\Omega$, and we set $\widetilde{\Omega} := \Omega \cup N$. We then define the function $\widetilde{\mathbf{u}} : \widetilde{\Omega} \to \mathbb{R}^n$ by

$$\widetilde{\mathbf{u}} := \begin{cases} \mathbf{u} & \text{in } \Omega\,, \\ \mathbf{u}_{|\partial\Omega} \circ \widetilde{\pi} \circ \mathbf{w}^{-1} & \text{in } N\,. \end{cases}$$

The first step is to prove that $\widetilde{\mathbf{u}} \in W^{1,A}(N \backslash \partial\Omega, \mathbb{R}^n)$. Then, one shows that the map $\mathbf{u} \to \widetilde{\mathbf{u}}$ is linear and, since

$$\|\widetilde{\mathbf{u}}\|_{W^{1,A}(N\backslash\partial\Omega,\mathbb{R}^n)} \leq c\|\mathbf{u}\|_{W^{1,A}(\partial\Omega,\mathbb{R}^n)}$$

for some constant $c > 0$, the map $\mathbf{u} \to \widetilde{\mathbf{u}}$ is bounded from $W^{1,A}(\partial\Omega, \mathbb{R}^n)$ to $W^{1,A}(N\backslash\partial\Omega, \mathbb{R}^n)$. Furthermore, from the continuity of the trace operator, the trace of $\widetilde{\mathbf{u}}_{|N\backslash\partial\Omega}$ on $\partial\Omega$ is $\mathbf{u}_{|\partial\Omega}$. Thus, $\widetilde{\mathbf{u}} \in W^{1,A}(\widetilde{\Omega}, \mathbb{R}^n)$ and the map $E : W^{1,A}(\Omega; \mathbb{R}^n) \cap W^{1,A}(\partial\Omega; \mathbb{R}^n) \to W^{1,A}(\widetilde{\Omega}; \mathbb{R}^n)$ defined setting $E\mathbf{u} := \widetilde{\mathbf{u}}$ is linear and bounded. The remaining assertions follow from Lemma 2 and explicit computations that can be performed as in [26, Proposition 6.3]. A key ingredient therein is the stability of compositions with bi-Lipschitz homeomorphisms. □

5.2 Regular Functions in $\overline{\mathcal{A}}(\boldsymbol{\Omega})$

As remarked in the introduction, a key question in the theory of existence in nonlinear elasticity is whether the distributional determinant Det$D\mathbf{u}$ equals the pointwise determinant det$D\mathbf{u}$. This can be rewritten equivalently as

$$\frac{1}{n}\mathrm{Div}[\mathrm{adj}\, D\mathbf{u}(\mathbf{x})\mathbf{u}(\mathbf{x})] = \det D\mathbf{u}(\mathbf{x}) \qquad \mathbf{x} \in \Omega\,, \tag{21}$$

where Div in the left-hand side stands for the distributional divergence and can be generalized to

$$\mathrm{Div}[\mathrm{adj} D\mathbf{u}(\mathbf{x})\mathbf{g}(\mathbf{u}(\mathbf{x}))] = \mathrm{div}\mathbf{g}(\mathbf{u}(\mathbf{x}))\det D\mathbf{u}(\mathbf{x})\,, \quad x \in \Omega\,, \tag{22}$$

for all $\mathbf{g} \in C^1(\mathbb{R}^n, \mathbb{R}^n) \cap W^{1,\infty}(\mathbb{R}^n, \mathbb{R}^n)$. It is easy to check that both the identities (21) and (22) hold true for smooth functions, say $\mathbf{u} \in C^2(\bar{\Omega}, \mathbb{R}^n)$, as a consequence of Piola's identity $\mathrm{Div}\,\mathrm{cof}\, D\mathbf{u} = \mathbf{0}$.

In a weaker setting, by using the definition of distributional divergence, (22) can be formulated as in (19), and in [25, 39], the class $\mathcal{A}$ of those orientation-preserving Orlicz–Sobolev functions satisfying (19) has been introduced and analyzed. The identity (20) underlying the definition of our class $\overline{\mathcal{A}}$ can be obtained from (22) multiplying by $\phi \in C^\infty(\bar\Omega)$ and then integrating by parts.

The aim of this section is then to identify the "regular" functions included in the class $\overline{\mathcal{A}}$. As a first remark, we note that the same argument as in [26, Lemma 7.1] gives $C^1(\bar\Omega;\mathbb{R}^n) \subset \overline{\mathcal{A}}(\Omega)$. The main result is the following proposition, where we prove that a suitable subclass of $W^{1,A}(\Omega;\mathbb{R}^n) \cap W^{1,A}(\partial\Omega;\mathbb{R}^n)$ is contained in $\overline{\mathcal{A}}(\Omega)$. The idea is to combine the extension property of Proposition 4 with the adaptation to the Orlicz–Sobolev setting of the result by Müller et al. [37].

Proposition 5 *Let Ω be an extendable open set. Let $q > \frac{n-1}{n-2}$. Then, the class of maps $\mathbf{u} \in W^{1,A}(\Omega;\mathbb{R}^n) \cap W^{1,A}(\partial\Omega;\mathbb{R}^n)$ such that*

$$\mathrm{cof} D\mathbf{u} \in L^q(\Omega;\mathbb{R}^{n\times n})\,, \quad (\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu} \in L^q(\partial\Omega;\mathbb{R}^n)$$

is a subset of $\overline{\mathcal{A}}(\Omega)$.

Proof Let $\mathbf{u} \in W^{1,A}(\Omega;\mathbb{R}^n) \cap W^{1,A}(\partial\Omega;\mathbb{R}^n)$ be such that $\mathrm{cof} D\mathbf{u} \in L^q(\Omega;\mathbb{R}^{n\times n})$ and $(\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu} \in L^q(\partial\Omega;\mathbb{R}^n)$. Then, taking into account (2), (8), and Hölder's inequality, $\mathrm{cof} D\mathbf{u} \in L^q$, with q as above, implies $\det D\mathbf{u} \in L^1$. Since $\frac{n-1}{n-2} \geq \frac{n}{n-1}$ and L^A can be embedded into L^{n-1}, the proof can be deduced by Müller et al. [37, Theorem 3.2] applied to $\widetilde{\mathbf{u}}$ the extension of $\mathbf{u}$ to an open set $\widetilde\Omega \supset \bar\Omega$ obtained with Proposition 4, which complies with $\widetilde{\mathbf{u}} \in W^{1,A}(\widetilde\Omega, \mathbb{R}^n)$ and cof $D\widetilde{\mathbf{u}} \in L^q(\widetilde\Omega;\mathbb{R}^{n\times n})$. This gives $\widetilde{\mathbf{u}} \in \mathcal{A}(\widetilde\Omega)$, whence, with Proposition 4(ii), we infer $\mathbf{u} \in \overline{\mathcal{A}}(\Omega)$. □

5.3 Some Properties of Orientation-Preserving Functions in $\overline{\mathcal{A}}(\Omega)$: Boundedness and Global Invertibility

In this section, we preliminarily prove that functions $\mathbf{u} \in \overline{\mathcal{A}}(\Omega)$ with $\det D\mathbf{u} \geq 0$ a.e. are bounded, thus proving the "global" counterpart of the local boundedness result [25, Proposition 4.2].

Proposition 6 *If $\mathbf{u} \in \overline{\mathcal{A}}(\Omega)$ with* $\det D\mathbf{u} \geq 0$ *a.e., then* $\deg(\mathbf{u},\Omega,\cdot) = \mathcal{N}_{\mathbf{u}}$ *a.e.,* $\mathrm{im_T}(\mathbf{u},\Omega) = \mathrm{im_G}(\mathbf{u},\Omega)$ *a.e. and* $\mathbf{u} \in L^\infty(\Omega;\mathbb{R}^n)$.

Proof The proof can be obtained as in [26, Proposition 8.4] by using the formula for the distributional derivative of the degree, Proposition 3. We then omit the details. □

As a consequence of Proposition 6, we obtain the following global invertibility result.

Theorem 4 *Let $\mathbf{u}, \mathbf{u}_0 \in \overline{\mathcal{A}}(\Omega)$ be such that* $\mathbf{u}|_{\partial\Omega} = \mathbf{u}_0|_{\partial\Omega}$, $\det D\mathbf{u} > 0$ *a.e.,* $\det D\mathbf{u}_0 \geq 0$ *a.e. and* $\mathbf{u}_0$ *injective a.e. Then,* $\mathbf{u}$ *is injective a.e. and* $\mathrm{im_G}(\mathbf{u},\Omega) = \mathrm{im_G}(\mathbf{u}_0,\Omega)$ *a.e.*

Proof See [26, Theorem 9.1]. □

The a.e. injectivity of $\mathbf{u}$ allows to define a.e. its inverse (see [26, Definition 9.2]).

Definition 10 Let $\mathbf{u} \in \overline{\mathcal{A}}(\Omega)$ be injective a.e. Let Ω_0 be the set of Definition 6, and $\Omega_1 \subset \Omega_0$ such that $\mathcal{L}^n(\Omega \backslash \Omega_1) = 0$ and $\mathbf{u}|_{\Omega_1}$ be injective. The inverse $\mathbf{u}^{-1} : \mathrm{im}_{\mathrm{T}}(\mathbf{u}, \Omega) \to \mathbb{R}^n$ is defined a.e. as $\mathbf{u}^{-1}(\mathbf{y}) = \mathbf{x}$, for every $\mathbf{y} \in \mathbf{u}(\Omega_1)$, where $\mathbf{x} \in \Omega_1$ satisfies $\mathbf{u}(\mathbf{x}) = \mathbf{y}$.

The a.e. inverse is a Sobolev function, as ensured by the following result which represents the global counterpart of [25, Proposition 4.11].

Theorem 5 *Let* $\mathbf{u} \in \overline{\mathcal{A}}(\Omega)$ *be injective a.e. with* $\det D\mathbf{u} > 0$ *a.e. Then,* $\mathbf{u}^{-1} \in W^{1,1}(\mathrm{im}_{\mathrm{T}}(\mathbf{u}, \Omega), \mathbb{R}^n)$ *and* $D\mathbf{u}^{-1}(\mathbf{y}) = D\mathbf{u}(\mathbf{u}^{-1}(\mathbf{y}))^{-1}$ *for a.e.* $\mathbf{y} \in \mathrm{im}_{\mathrm{T}}(\mathbf{u}, \Omega)$.

Proof [26, Theorem 9.3]. □

6 Existence of Minimizers

In this section, we prove the existence of minimizers for functionals of the form

$$I(\mathbf{u}) = \int_{\Omega} W(\mathbf{x}, \mathbf{u}(\mathbf{x}), D\mathbf{u}(\mathbf{x}))\, \mathrm{d}\mathbf{x} \tag{23}$$

on the class $\overline{\mathcal{A}}(\Omega)$, under the assumption that W is *polyconvex* in the last variable (see, e.g., [16]).

The following result establishes the compactness in the class $\overline{\mathcal{A}}(\Omega)$.

Proposition 7 *Let* $\{\mathbf{u}_j\}_{j\in\mathbb{N}} \subset \overline{\mathcal{A}}(\Omega)$ *be a bounded sequence in* $W^{1,A}(\Omega;\mathbb{R}^n) \cap W^{1,A}(\partial\Omega;\mathbb{R}^n)$ *and such that* $\{\det D\mathbf{u}_j\}_{j\in\mathbb{N}}$ *is equi-integrable. Then, there exist a subsequence (not relabeled)* $\{\mathbf{u}_j\}$ *and a function* $\mathbf{u} \in \overline{\mathcal{A}}(\Omega)$ *such that*

$$\mathbf{u}_j \rightharpoonup \mathbf{u} \ \ \textit{in}\ W^{1,A}(\Omega;\mathbb{R}^n) \cap W^{1,A}(\partial\Omega;\mathbb{R}^n) \ \ \textit{and} \ \ \det D\mathbf{u}_j \rightharpoonup \det D\mathbf{u} \ \ \textit{in}\ L^1(\Omega) \tag{24}$$

as $j \to +\infty$.

Proof The argument is quite standard, and we adapt the proof of [26, Proposition 10.2] to our setting.

By assumptions, we can find $\mathbf{u} \in W^{1,A}(\Omega;\mathbb{R}^n)$, $\mathbf{v} \in W^{1,A}(\partial\Omega;\mathbb{R}^n)$, and $w \in L^1(\Omega)$ such that, up to a (not relabeled) subsequence, we have

$$\mathbf{u}_j \rightharpoonup \mathbf{u} \ \text{ in } \ W^{1,A}(\Omega;\mathbb{R}^n)\,, \ \mathbf{u}_j \rightharpoonup \mathbf{v} \ \text{ in } \ W^{1,A}(\partial\Omega;\mathbb{R}^n)\,, \ \det D\mathbf{u}_j \rightharpoonup w \ \text{ in } \ L^1(\Omega)\,.$$

Up to a further subsequence, we may assume that $\mathbf{u}_j \to \mathbf{u}$ a.e., and, taking into account the embedding result of [25, Proposition 2.6] and the weak continuity of the cofactors (see, e.g., [16, Theorem 8.20]), we get $\operatorname{cof} D\mathbf{u}_j \rightharpoonup \operatorname{cof} D\mathbf{u}$ in

$L^1(\Omega; \mathbb{R}^{n\times n})$. The continuity of the traces $\mathbf{u}_{j\,|_{\partial\Omega}} \rightharpoonup \mathbf{u}_{|\partial\Omega}$ in $L^A(\partial\Omega; \mathbb{R}^n)$ implies $\mathbf{v} = \mathbf{u}_{|\partial\Omega}$, while (17) and [21, Theorem 3] give $w = \det D\mathbf{u}$ a.e. The rest of the proof leading to the identity (18) is based on standard computations and on the weak continuity result

$$(\Lambda_{n-1} D\mathbf{u}_j)\boldsymbol{\nu} \rightharpoonup (\Lambda_{n-1} D\mathbf{u})\boldsymbol{\nu} \quad \text{in } \mathrm{LLog}^{\frac{\alpha}{n-1}}\mathrm{L}(\partial\Omega, \mathbb{R}^n),$$

which can be inferred along the lines of [26, Proposition 10.1]. □

Now, we can prove the existence of minimizers for I on a suitable subclass of $\overline{\mathcal{A}}(\Omega)$.

Theorem 6 *Let* $\mathbf{u}_0 \in \overline{\mathcal{A}}(\Omega)$ *be injective a.e. and such that* $\det D\mathbf{u}_0 \geq 0$ *a.e. Let* $W : \Omega \times \mathrm{im_T}(\mathbf{u}_0, \Omega) \times \mathbb{R}^{n\times n}_+ \to \mathbb{R}$ *comply with the following assumptions:*

(i) *W is measurable.*
(ii) *$W(\mathbf{x}, \cdot, \cdot)$ is lower semicontinuous for a.e.* $\mathbf{x} \in \Omega$.
(iii) *$W(\mathbf{x}, \mathbf{y}, \cdot)$ is polyconvex for a.e.* $\mathbf{x} \in \Omega$ *and for every* $\mathbf{y} \in \mathrm{im_T}(\mathbf{u}_0, \Omega)$.
(iiii) *There exist a constant $c > 0$, a function $a \in L^1(\Omega)$, and a Borel function $h : (0, \infty) \to [0, \infty)$ with*

$$\lim_{t\searrow 0} h(t) = \lim_{t\to\infty} \frac{h(t)}{t} = \infty \tag{25}$$

such that

$$W(\mathbf{x}, \mathbf{y}, \mathbf{F}) \geq a(\mathbf{x}) + c\, A(\|\mathbf{F}\|) + h(\det \mathbf{F}) \tag{26}$$

for a.e. $\mathbf{x} \in \Omega$*, every* $\mathbf{y} \in \mathrm{im_T}(\mathbf{u}_0, \Omega)$ *and* $\mathbf{F} \in \mathbb{R}^{n\times n}_+$.

Define

$$\mathcal{A} := \left\{\mathbf{u} \in \overline{\mathcal{A}}(\Omega) : \ \det D\mathbf{u} > 0 \ \textit{a.e.} \quad \textit{and} \ \ \mathbf{u}_{|\partial\Omega} = \mathbf{u}_{0|_{\partial\Omega}}\right\},$$

and assume that $\mathcal{A} \neq \emptyset$ and $I \not\equiv \infty$ on $\mathcal{A}$. Then, I admits a minimizer on $\mathcal{A}$, and any element in $\mathcal{A}$ is injective a.e.

Proof Once a compactness result has been proved, the proof of the existence of minimizers is based on a well-known argument. We follow the scheme in [26, Theorem 10.3].

Under the coerciveness assumption (26), the lower semicontinuity of the functional I can be inferred from [5, Theorem 5.4]. The a.e. injectivity of each function in $\mathcal{A}$ is a consequence of Theorem 4.

Now, let $\{\mathbf{u}_j\}_{j\in\mathbb{N}}$ be a minimizing sequence of I in $\mathcal{A}$. Then, by Proposition 6 and Theorem 4, we get

$$\mathbf{u}_j(\mathbf{x}) \in \mathrm{im_T}(\mathbf{u}_0, \Omega) \qquad \text{for a.e. } \mathbf{x} \in \Omega \text{ and all } j \in \mathbb{N}\,. \tag{27}$$

From assumption *(iiii)* and De la Vallée Poussin's Theorem (Theorem 3), we infer that $\{D\mathbf{u}_j\}_{j\in\mathbb{N}}$ is equibounded in $L^A(\Omega, \mathbb{R}^{n\times n})$ and $\{\det D\mathbf{u}_j\}_{j\in\mathbb{N}}$ is equi-integrable. Furthermore, (27) and the boundedness of topological image imply that $\{\mathbf{u}_j\}_{j\in\mathbb{N}}$ is bounded in L^∞ and then in $W^{1,A}(\Omega, \mathbb{R}^n)$. The compactness result of Proposition 7 provides a function $\mathbf{u} \in \overline{\mathcal{A}}(\Omega)$ such that, up to a subsequence,

$$\mathbf{u}_j \rightharpoonup \mathbf{u} \ \text{ in } \ W^{1,A}(\Omega; \mathbb{R}^n)\,, \quad \det D\mathbf{u}_j \rightharpoonup \det D\mathbf{u} \ \text{ in } \ L^1(\Omega)\,.$$

Now, a simple argument by contradiction based on assumption *(iiii)* shows that $\det D\mathbf{u} > 0$ a.e. Then, since the boundary condition is preserved in the limit, $\mathbf{u}_{|\partial\Omega} = \mathbf{u}_{0|\partial\Omega}$, whence $\mathbf{u} \in \mathcal{A}$. This concludes the proof. □

Acknowledgments The authors are members of Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of INdAM. The research of B. S. was supported by PRIN Project 2017TEXA3H. G.S. was supported by the Italian Ministry of Education, University and Research (http://dx.doi.org/10.13039/501100021856) through the MIUR – PRIN project 2017BTM7SN "Variational methods for stationary and evolution problems with singularities and interfaces". The authors gratefully acknowledge the anonymous referee for a careful reading of the paper and for the interesting remarks leading to improvements of the manuscript.

References

1. R.A. Adams, *Sobolev Spaces* (Academic Press, New York, 1975)
2. J.M. Ball, Convexity conditions and existence theorems in nonlinear elasticity. Arch. Rational Mech. Anal. **63**, 337–403 (1977)
3. J.M. Ball, Global invertibility of Sobolev functions and the interpenetration of matter. Proc. Roy. Soc. Edinburgh Sect. A Math. **88**, 315–328 (1981)
4. J.M. Ball, F. Murat, $W^{1,p}$-quasiconvexity and variational problems for multiple integrals. J. Funct. Anal. **58**(3), 255–253 (1984)
5. J.M. Ball, J.C. Currie, P.J. Olver, Null Lagrangians, weak continuity, and variational problems of arbitrary order. J. Funct. Anal. **41**(2), 135–174 (1981)
6. M. Barchiesi, A. De Simone, Frank energy for nematic elastomers: a nonlinear model. ESAIM Control Optim. Calc. Var. **21**, 277–372 (2015)
7. M. Barchiesi, D. Henao, C. Mora-Corral, Local invertibility in sobolev spaces with applications to nematic elastomers and magnetoelasticity. Arch. Rational Mech. Anal. **224**, 743–816 (2017)
8. C. Bennett, K. Rudnick, *On Lorentz-Zygmund Spaces* (Instytut Matematyczny Polskiej Akademi Nauk, Warszawa, 1980)
9. C. Bennett, R. Sharpley, *Interpolation of Operators*. Pure and Applied Mathematics, vol. 129 (Academic, Boston, 1988)
10. M.C. Calderer, C.A. Garavito Garzón, C. Luo, Liquid crystal elastomers and phase transitions in actin rod networks. SIAM J. Appl. Math. **74**, 649–675 (2014)
11. M. Carozza, A. Cianchi, Continuity properties of weakly monotone Orlicz-Sobolev functions. Adv. Calc. Var. **14**(1), 107–126 (2021)
12. P. Celada, S. Perrotta, Polyconvex energies and cavitation. NoDea **20**, 295–321 (2013)
13. A. Cianchi, Continuity properties of functions from Orlicz-Sobolev spaces and embedding theorems. Ann. Scuola Norm. Sup. Pisa Cl. Sci. **23**(4), 575–608 (1996)
14. P.G. Ciarlet, J. Nečas, Injectivity and self-contact in nonlinear elasticity. Arch. Rat. Mech. Anal. **97**, 171–188 (1987)

15. S. Conti, C. De Lellis, Some remarks on the theory of elasticity for compressible Neohookean materials. Ann. Scuola Norm. Sup. Pisa Cl. Sci. **2**, 521–549 (2003)
16. B. Dacorogna, *Direct Methods in the Calculus of Variations*. Applied Mathematical Sciences, vol. 78, 2nd edn. (Springer, New York, 2008)
17. K. Deimling, *Nonlinear Functional Analysis* (Springer, Berlin, 1985)
18. I. Fonseca, W. Gangbo, *Degree Theory in Analysis and Applications* (Oxford University Press, New York, 1995)
19. I. Fonseca, J. Malý, Relaxation of multiple integrals below the growth exponent. Ann. Inst. H. Poincaré Anal. Non Linéaire **14**(3), 309–338 (1997)
20. P. Hajlasz, J. Malý, Approximation in Sobolev spaces of nonlinear expressions involving the gradient. Ark. Mat. **40**, 245–274 (2002)
21. D. Henao, C. Mora-Corral, Invertibility and weak continuity of the determinant for the modelling of cavitation and fracture in nonlinear elasticity. Arch. Rat. Mech. Anal. **197**, 619–655 (2010)
22. D. Henao, C. Mora-Corral, Fracture surface and regularity of inverses for BV deformations. Arch. Rat. Mech. Anal. **201**, 575–629 (2011)
23. D. Henao, C. Mora-Corral, Lusin's condition and the distributional determinant for deformations with finite energy. Adv. Calc. Var. **5**, 355–409 (2012)
24. D. Henao, C. Mora-Corral, Regularity of inverses of Sobolev deformations with finite surface energy. J. Funct. Anal. **208**, 2356–2378 (2015)
25. D. Henao, B. Stroffolini, On Sobolev-Orlicz nematic elastomers. Nonlinear Anal. **194**, 111513 (2020)
26. D. Henao, C. Mora-Corral, M. Oliva, Global invertibility of Sobolev maps. Adv. Calculus Var. **14**(2), 207–230 (2021)
27. S. Hencl, P. Koskela, *Lectures on Mappings of Finite Distortion*. Lecture Notes in Mathematics, vol. 2096 (Springer, Berlin, 2014)
28. T. Iwaniec, G. Martin, *Geometric Function Theory and Nonlinear Analysis*. Oxford Mathematical Monographs (Oxford University Press, Oxford, 2001)
29. M.A. Krasnosel'skiǐ, Y.B. Rutickiǐ, *Convex Functions and Orlicz Spaces* (P. Noordhoff Ltd., Groningen, 1961)
30. S. Krömer, Global invertibility for orientation-preserving Sobolev maps via invertibility on or near the boundary. Arch. Rat. Mech. Anal. **238**, 1113–1155 (2020)
31. M. Kružík, U. Stefanelli, J. Zeman, Existence results for incompressible magnetoelasticity. Discrete Cont. Dyn. Syst. A **35**, 2615–2623 (2015)
32. A. Kufner, O. John, S. Fucik, *Function Spaces* (Springer, Amsterdam, 1977)
33. J.G. Llavona, *Approximation of Continuously Differentiable Functions*. North-Holland Mathematics Studies, vol. 130 (North-Holland, Amsterdam, 1986)
34. P. Marcellini, The stored-energy for some discontinuous deformations in nonlinear elasticity. Progr. Nonlinear Differ. Equ. Appl. **2**, 767–786 (1989)
35. C. Mora-Corral, M. Oliva, Relaxation of nonlinear elastic energies involving the deformed configuration and applications to nematic elastomers. ESAIM: COCV **25**, 19 (2019)
36. S. Müller, S.J. Spector, An existence theory for nonlinear elasticity that allows for cavitation. Arch. Rational Mech. Anal. **131**, 1–66 (1995)
37. S. Müller, T. Qi, B.S. Yan, On a new class of elastic deformations not allowing for cavitation. Ann. Inst. H. Poincaré Anal. Non Linéaire **11**(2), 217–243 (1994)
38. J.R. Munkres, *Topology: A First Course* (Prentice-Hall, Englewood Cliffs, 1975)
39. G. Scilla, B. Stroffolini, Relaxation of nonlinear elastic energies related to Orlicz–Sobolev nematic elastomers. Atti Accad. Naz. Lincei Rend. Lincei Mat. Appl. **31**(2), 349–389 (2020)
40. V. Šverák, Regularity properties of deformations with finite energy. Arch. Rational Mech. Anal. **100**, 105–127 (1988)
41. S.K. Vodop'yanov, Topological and geometrical properties of mappings with summable Jacobian in Sobolev classes. Siberian J. Math. J. **41**, 19–39 (2000)

Global Existence of Solutions for the One-Dimensional Response of Viscoelastic Solids Within the Context of Strain-Limiting Theory

Yasemin Şengül

1 Introduction

It is now known that using implicit theories the response of materials can be described much more generally. In explicit constitutive modelling, which can be seen as a special case of implicit modelling, the stress is given explicitly as a function of the strain and it is not possible to obtain a nonlinear relationship between the stress and the linearized strain in this case. On the other hand, in implicit constitutive modelling, as a result of expressing the strain as a function of the stress, a nonlinear relation can be achieved after linearization. This idea is due to Rajagopal [24, 25, 27–30, 32] who introduced a new class of elastic materials where the response of materials is given by an implicit constitutive relation between the stress and the deformation gradient. For the strain-limiting elastic materials, the displacement gradients and the strain remain bounded even if the stress tends to infinity. The theory of limiting strain provided by such implicit constitutive relations is able to explain such a phenomena which has actually been observed experimentally before (see [39] and the references therein) such as the behaviour of some composite materials (see, e.g., [16]), biological tissues, and fracture of brittle materials (see, e.g., [11, 12, 35]).

In one-dimensional case, the relationship between the linearized strain $\epsilon(x, t)$ and the displacement function $u(x, t)$ is clearly given by $\epsilon = u_x$. Following Rajagopal [24, 25], one can write the strain-limiting model in this case as

$$\epsilon = h(S),$$

Y. Şengül (✉)
School of Mathematics, Cardiff University, Cardiff, UK
e-mail: SengulTezelY@cardiff.ac.uk

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_14

where h is a nonlinear function, and $S(x,t)$ is the Cauchy stress. Rajagopal [26] extended this approach to rate-type viscoelastic materials and proposed an implicit constitutive equation with an additional term ϵ_t corresponding to the time rate of change of the linearized strain, which is called the strain rate. Erbay and Şengül [8] studied the strain-limiting rate-type viscoelastic model given as

$$\epsilon + \nu\epsilon_t = h(S), \tag{1}$$

where $\nu > 0$ is the viscosity constant, and h is a nonlinear function with $h(0) = 0$. The constitutive relation (1) is a generalization of the Kelvin–Voigt viscoelastic model. As explained in depth in [37], while such a constitutive relation allows for creep, a phenomenon exhibited by viscoelastic solids, it is not capable of exhibiting stress relaxation, a characteristic that it shares in common with the Kelvin–Voigt model. While this is not a shortcoming in that several viscoelastic bodies creep but do not exhibit stress relaxation, the model under consideration cannot be used to describe a large class of polymeric solids that do exhibit stress relaxation.

A very important feature of the constitutive relation (1) is that it is linear in the linearized measure of the strain. This makes it very useful in describing the small strain and small strain rate response of a large class of viscoelastic solids (see e.g., [19, 21, 33, 34]. However, as mentioned in [39] (see also [32]), this constitutive relation is a special case of more general models where mixed terms involving products of the linearized strain and the stress appear.

In [8], using (1) and the equation of motion, Erbay and Şengül obtained the third-order semilinear equation

$$S_{xx} + \nu\, S_{xxt} = h(S)_{tt} \tag{2}$$

and studied traveling wave solutions under the assumption of two constant equilibrium states at infinity. As they stated in [8], due to the fact that the stress S involves both the elastic and dissipative character, it is not true that the first and the second terms on the left of (2) represent elastic and dissipative effects, respectively. Different from classical viscoelastic models, the inertia term is nonlinear as well as the fact that the equation is posed in terms of the stress S rather than the displacement or the deformation. In [10], the Cauchy problem defined by (2) together with some suitable initial conditions is considered and it is proven that for initial data in appropriate function spaces and appropriate forms of h appearing in the constitutive relation, the Cauchy problem (2) is locally well-posed in time.

Studies on limiting strain models have attracted a lot of attention recently. Different from the ones mentioned above, there are many valuable contributions to this theory in a wide range of contexts (see, e.g., [5–7, 13–15, 17, 18]) including numerical studies (see, e.g., [20, 22, 23]). In the one-dimensional setting, Şengül [38] studied Eq. (2) with an arctangent type nonlinearity. Also, Erbay and Şengül [9] introduced a stress-rate type model and showed that it is thermodynamically consistent. In multi-dimensional setting, on the other hand, starting with elastic models [1, 2], very recently, Bulíček, Patel, Şengül, and Süli [3, 4] proved existence of weak solutions with both periodic and Dirichlet boundary data.

In the present work, after reviewing the local-in-time existence results in [10], the existence for the displacement is stated. After that, the energy is investigated and it is shown that it decreases over time. The proof of the main theorem, which is the global existence of solutions, relies on this investigation about the energy and a lemma where a characterization for the blow-up conditions is stated and proven. Also, smallness assumptions made on the deformation gradient and its time derivative are shown to hold under additional conditions. As in the local-in-time existence case, the analysis in this work is done in $\mathbb{R}$. However, this could also be done for a bounded domain under suitable boundary conditions.

The structure of the paper is as follows. In Sect. 2 we review local existence results and rigorously show it for u which was not done in [10]. In Sect. 3 we recall some notation we use throughout the paper, and then, in Sect. 4, we prove that the local solutions are, in fact, global by using a characterization of the blow-up condition as well as the decrease of the energy over time. In Sect. 5 we show that the theory developed is, in fact, compatible with the smallness assumptions of the strain-limiting theory.

2 Preliminaries

In [10], the authors convert the equation for the stress to obtain an equation in a new variable defined as the sum of the strain and the strain rate and write it as a time-dependent heat equation. They use the results related to the variable coefficient heat equation and the techniques from the theory of elliptic operators. The proof of their main theorem includes linearization around a given state, definition of a contractive mapping and the usage of Banach's fixed point theorem. We will not briefly review the results about local-in-time existence of solutions.

A new variable $\omega(x, t)$ which is the sum of the strain and strain rate is defined as $\omega = \epsilon + \nu\epsilon_t$, and it is assumed that the function $h(\cdot)$ satisfies $h'(z) > 0$ for any $z \in \mathbb{R}$. As a result (1) is written in the form

$$S = g(\omega), \tag{3}$$

where g is the inverse function of h satisfying $g(0) = 0$. As a consequence of the invertibility g is a sufficiently smooth and $g'(z) > 0$ for any $z \in \mathbb{R}$, which means the stress S is a smooth function of ω and it is strictly increasing. As it is mentioned by Rajagopal in [32], it is best not to invert the expression for the linearized strain $\boldsymbol{\epsilon}$ as a function of $\mathbf{T}$, and that one should solve the equation of motion and the constitutive relation simultaneously in order to avoid misinterpretation of the procedure. However, in the current work we are doing the analysis in one-dimensional space for which it is relatively easier to return to the original variable and hence avoid misinterpretation.

In terms of ω, (2) becomes

$$\omega_{tt} = g(\omega)_{xx} + \nu g(\omega)_{xxt}. \tag{4}$$

As mentioned in [10], it is important to note that due to ω involving both the strain and the strain rate by definition, the first and second terms on the right-hand side of (4) do not represent purely elastic and purely dissipative effects, respectively. Moreover, one can write (4) in the divergence form using the function $\eta(x,t) = \int_{-\infty}^{x} \omega(y,t)dy$. Substituting these expressions into (4), one obtains the Cauchy problem

$$\eta_{tt} = g(\eta_x)_x + \nu g(\eta_x)_{xt}, \qquad x \in \mathbb{R}, \quad t > 0 \tag{5}$$

$$\eta(x,0) = \eta_0(x), \quad \eta_t(x,0) = \eta_1(x), \qquad x \in \mathbb{R}. \tag{6}$$

The analysis in [10] is based on the assumptions $\eta_0(\pm\infty) = 0$ and $\eta_1(\pm\infty) = 0$.

Letting $\eta_t = \phi$, (5) can be rewritten as the system

$$\eta_t = \phi_x, \tag{7}$$

$$\phi_t + A(t)\phi = G(t) + F(t), \tag{8}$$

where

$$A(t) = 1 - \nu D_x(g'(\eta_x)D_x), \quad G(t) = g(\eta_x)_x, \quad F(t) = \eta_t. \tag{9}$$

For the nonlinearity g, the following properties are known, which are also going to be used for the estimates in the proof of global existence of solutions.

Lemma 1 *Let $g \in C^\infty(\mathbb{R})$ with $g(0) = 0$. If $z \in H^s(\mathbb{R}) \cap L^\infty(\mathbb{R})$, $s \geq 0$, then*

$$\|g(z)\|_{H^s \cap L^\infty} \leq K_1 \|z\|_{H^s \cap L^\infty},$$

where K_1 depends only on $\|z\|_{L^\infty}$.

Lemma 2 *Let $g \in C^\infty(\mathbb{R})$. If $z_1, z_2 \in H^s(\mathbb{R}) \cap L^\infty(\mathbb{R})$, $s \geq 0$, then*

$$\|g(z_1) - g(z_2)\|_{H^s \cap L^\infty} \leq K_2 \|z_1 - z_2\|_{H^s \cap L^\infty},$$

where K_2 depends on $\|z_1\|_{H^s \cap L^\infty}$ and $\|z_2\|_{H^s \cap L^\infty}$.

In [10], the following result is obtained which is fundamental for the proof of the global-existence result in this work.

Lemma 3 *Let $s > 5/2$, $T > 0$. Assume that A is defined as in (9) and that $g'(z) > 0$ holds for all $z \in \mathbb{R}$. Also assume that $G \in C^1([0,T]; H^{s-2})$ and $F \in C([0,T]; H^s)$. Then there exists a solution $(\phi, \eta) \in \big(C([0,T]; H^s), C^1([0,T];$*

$H^s)\big)$ to (7) and (8) with $\eta(0) = \eta_0$ and $\phi(0) = \phi_0 = \eta_1$ and satisfying the following estimates

$$\begin{aligned}\|\eta(t)\|_s \le \|\eta_0\|_{H^s} &+ CT\Big(\|\eta_1\|_s + \|G(0)\|_{s-2}\Big)\\ &+ C\,T\int_0^t \Big(\|G(\tau)\|_{s-2} + \|G_\tau(\tau)\|_{s-2} + \|F(\tau)\|_s\Big)d\tau \qquad (10)\end{aligned}$$

$$\begin{aligned}\|\eta_t(t)\|_s \le C\Bigg(&\|\eta_1\|_s + \|G(0)\|_{s-2}\\ &+ \int_0^t \Big(\|G(\tau)\|_{s-2} + \|G_\tau(\tau)\|_{s-2} + \|F(\tau)\|_s\Big)d\tau\Bigg) \qquad (11)\end{aligned}$$

for $0 \le t \le T$, where C is a constant.

Finally in [10], a fixed point scheme is constructed in order to prove local well-posedness for the Cauchy problem (5) and (6) in the Banach space defined by

$$\begin{aligned}X^s([0,T]) = \Big\{z \in C^1([0,T]; H^s(\mathbb{R}))\Big|&z(0) = \eta_0, z_t(0) = \eta_1,\\ &\|z(t)\|_s \le \bar{\delta}, \|z_t(t)\|_s \le M, t \in [0,T]\Big\}\end{aligned}$$

and endowed with the norm

$$\|z\|_{X^s([0,T])} = \sup_{t\in[0,T]}\Big(\|z(t)\|_s + \|z_t(t)\|_s\Big). \qquad (12)$$

The following two results are proven in [10] giving the local-in-time existence of solutions for the Cauchy problem (5) and (6).

Theorem 1 *Let $s > 5/2$. Assume $g \in C^{r+1}$ with $r = [s] + 1$. Assume also that $\eta_0 \in H^s$ with $\|\eta_0\|_s \le \frac{\bar{\delta}}{2(1+T_0K_1(\delta))}$ for some $T_0 > 0$, where $\|\eta_x\|_\infty \le \delta$ and $K_1(\delta)$ is as in Lemma 1, and $\bar{\delta}$ is as in the definition of $X^s([0,T])$. Then, there exists a sufficiently small time $T > 0$ with $T \le T_0$, such that the Cauchy problem (5) and (6) admits a unique solution $\eta \in X^s([0,T])$.*

Note that since $H^s \subset C^k$ if $s > \frac{1}{2} + k$, the solutions are classical and by choosing s large they can be made as smooth as required. Recalling $\omega = \eta_x$ one can also state a similar result for the local existence of ω, which is not necessary for the purpose of the current work.

2.1 Local Existence for the Displacement

From the definitions $\epsilon = u_x$, $\omega = \epsilon + \nu\epsilon_t$ and $\omega = \eta_x$, we obtain $\eta = u + \nu u_t$. Now, given η we can rewrite this as

$$u_t + \frac{1}{\nu}u = \frac{1}{\nu}\eta,$$

which can be viewed as an ordinary differential equation in u. Solving it, one obtains

$$u(x,t) = u(x,0)e^{-\frac{t}{\nu}} + \frac{1}{\nu}\int_0^t e^{-\frac{1}{\nu}(t-\tau)}\eta(x,\tau)d\tau. \tag{13}$$

Since $\eta, \eta_t \in H^s$ for $s > 5/2$, we immediately obtain $u, u_t \in H^s$ for $s > 5/2$. As a result Theorem 1 is valid for the displacement u. Moreover,

$$u(x,t) = u(x,0) + \int_0^t u_t(x,\tau)d\tau. \tag{14}$$

Together with the definition $\eta = u + \nu u_t$, this gives the following local existence result for the displacement.

Theorem 2 *Let $s > 5/2$. Assume $h \in C^{r+1}$ with $r = [s] + 1$, and $h'(z) > 0$ for all $z \in \mathbb{R}$. Assume also that $u_0, u_1 \in H^s(\mathbb{R})$. Then, there exists a sufficiently small time $T > 0$ such that the system*

$$\begin{aligned} &u_{tt} = S_x, \\ &\epsilon + \nu\epsilon_t = h(S), \\ &u(x,0) = u_0(x), \quad u_t(x,0) = u_1(x), \end{aligned} \tag{15}$$

admits a unique solution $u \in X^s([0,T])$.

Remark 1 Even though system (15) seems to have more than one unknowns, using g, which is the inverse of h, and the fact that in one space dimension $\epsilon = u_x$, one can write

$$u_{tt} = g(u_x + \nu u_{xt})_x,$$

which clearly has the only unknown u.

3 Some Conventions

We use the standard notations for Lebesgue and Sobolev spaces as well as the spaces of continuous functions. With $C^k(\mathbb{R})$, $k = 0, 1, 2, \ldots$ we represent the set of functions on $\mathbb{R}$ that are k-times continuously differentiable. Also $C([0,T]; H^s)$

denotes the space of all H^s-valued functions z on the interval $[0, T]$ of real numbers such that z is strongly continuous on $[0, T]$. Similarly, $C^1([0, T]; H^s)$ denotes the space of all continuously differentiable H^s-valued functions. Here, $H^s = H^s(\mathbb{R})$ denotes the L^2-based Sobolev space of order s on $\mathbb{R}$ with the usual norm $\|z\|_s = \left(\int_{\mathbb{R}}(1+\xi^2)^s|\widehat{z}(\xi)|^2 d\xi\right)^{1/2}$, where the symbol $\widehat{\ }$ denotes the Fourier transform. Similarly, we denote the norm in the space $L^\infty = L^\infty(\mathbb{R})$ as $\|z\|_{L^\infty} = \operatorname*{ess\,sup}_{x\in\mathbb{R}} |z(x)|$. Finally, generic positive constants are denoted by C.

4 Global Existence

In this section, we investigate the energy of the system. We recall that the relation between the Helmholtz free energy ψ and the Gibbs free energy G is given by $\psi = G - S\frac{\partial G}{\partial S}$.

4.1 Energy Decay

In one space dimension, the equation of motion for a homogeneous, viscoelastic, infinite medium is given by

$$u_{tt} = S_x. \tag{16}$$

We assume that the density is constant and the stress tends to zero at infinity: $S \to 0$ as $x \to \pm\infty$. Below we show that the energy decays over time for the strain rate type model (1).

Proposition 1 *For any $t \in [0, T)$, the energy defined as*

$$E(t) = \int_{\mathbb{R}} \left(\frac{1}{2}|u_t|^2 + \psi(S)\right) dx \tag{17}$$

decreases over time.

Proof Multiplying (16) by u_t and integrating over the space variable, we get

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} |u_t|^2\, dx = \int_{\mathbb{R}} S_x u_t\, dx. \tag{18}$$

One integration by parts on the right-hand side of this equation yields

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} |u_t|^2\, dx = -\int_{\mathbb{R}} S\epsilon_t\, dx. \tag{19}$$

Rewriting the equation of motion (16) in the form $\epsilon_{tt} = S_{xx}$ in terms of the strain and using the constitutive relation (1), one can indeed verify that $\epsilon_t = (h(S))_t - \nu S_{xx}$ and consequently $S\epsilon_t = S(h(S))_t - \nu S S_{xx}$. Assuming the existence of a potential function $H(S)$ with $h(S) = -dH(S)/dS$ we get $S\epsilon_t = (H - SdH(S)/dS)_t - \nu S S_{xx}$. This can be written as $S\epsilon_t = \psi_t - \nu S S_{xx}$ in terms of the Helmholtz free energy $\psi(S)$ if we require that the potential function $H(S)$ is the Gibbs free energy. Substitution of the above relation obtained for $S\epsilon_t$ into (19) yields

$$\frac{d}{dt}\int_{\mathbb{R}}\left(\frac{1}{2}|u_t|^2 + \psi(S)\right)dx = -\nu\int_{\mathbb{R}} S_x^2\,dx, \tag{20}$$

where one integration by parts has been performed on the right-hand side. For $\nu > 0$, this result shows that the total energy is decreasing over time. □

Remark 2 The right-hand side in relation (20) is a term due to dissipation. Therefore, it is an energy-dissipation balance and the sum of the kinetic energy and the elastic energy is a decreasing function of time.

Any solution $u \in C^1([0,T]; H^s)$ of system (15) can be extended to a maximal time interval of existence $[0, T_{\max})$ where finite $T_{\max}$ is characterized by the blow-up condition

$$\limsup_{t\to T_{\max}^-}\big(\|u(t)\|_s + \|u_t(t)\|_s\big) < \infty. \tag{21}$$

This means that the solution is global, meaning that $T_{\max}^- = \infty$ if and only if

$$\text{for all } T < \infty, \text{ we have } \limsup_{t\to T^-}\big(\|u(t)\|_s + \|u_t(t)\|_s\big) < \infty. \tag{22}$$

Theorem 3 *Assume that $s > 5/2$ and $u_0, u_1 \in H^s(\mathbb{R})$. Then any local-in-time solution $u \in C^1([0,T]; H^s)$ to system (15) exists globally if and only if*

$$\limsup_{t\to T^-}\big(\|u_x(t)\|_\infty + \|u_{xt}(t)\|_\infty\big) < \infty. \tag{23}$$

Proof By the Sobolev embedding theorem, we know that $H^s \subset L^\infty$ for $s > 1/2$. Therefore, since $s > 5/2$, $\|u(t)\|_s < \infty$ implies $\|u_x(t)\|_\infty < \infty$ by

$$\|u_x(t)\|_\infty \le C\|u_x(t)\|_{s-2} \le C\|u_x(t)\|_{s-1} \le C\|u(t)\|_s.$$

Similarly for $\|u_{xt}(t)\|_\infty$. This means (23) holds. Conversely, assume that the solutions exists for $t \in [0,T)$ and (23) holds. Since $\eta = u + \nu u_t$, (23) implies that $\|\eta_x(t)\|_\infty$ is bounded. This means we can use Lemma 1 and Lemma 2 with $K_1(M)$ and $K_2(M)$ in the following estimates

$$\begin{aligned}\|\eta(t)\|_s &\le \|\eta_0\|_s + CT\Big(\|\eta_1\|_s + \|G(0)\|_{s-2}\Big)\\ &\qquad + C\,T\int_0^t \Big(\|G(\tau)\|_{s-2} + \|G_\tau(\tau)\|_{s-2} + \|F(\tau)\|_s\Big)d\tau\\ \|\eta_t(t)\|_s &\le C\Bigg(\|\eta_1\|_s + \|G(0)\|_{s-2}\\ &\qquad + \int_0^t \Big(\|G(\tau)\|_{s-2} + \|G_\tau(\tau)\|_{s-2} + \|F(\tau)\|_s\Big)d\tau\Bigg)\end{aligned}$$

for $0 \le t \le T$, where C is a constant, which we know by Lemma 3. As a result we obtain

$$\begin{aligned}\|G(t)\|_{s-2} &= \|g(\eta_x)_x\|_{s-2} = \|g(\eta_x)\|_{s-1} \le K_1(M)\|\eta_x\|_{s-1} = K_1(M)\|\eta\|_s\\ \|G_t(t)\|_{s-2} &= \|g'(\eta_x)\eta_{xt}\|_{s-1} \le C\big(\|g'(\eta_x)\|_\infty\|\eta_{xt}\|_{s-1} + \|g'(\eta_x)\|_{s-1}\|\eta_{xt}\|_\infty\big)\\ &\le C\big(K_1(M)M\|\eta_t\|_s + K_1(M)\|\eta\|_s\|\eta_{xt}\|_\infty\big).\end{aligned}$$

For $\|\eta_{xt}\|_\infty$, since $\eta = u + \nu u_t$ we have $\|\eta_{xt}\|_\infty \le \|u_{xt}\|_\infty + \nu\|u_{xtt}\|_\infty$. Moreover, by (23), $\|u_{xt}\|_\infty$ and $\|u_{xtt}\|_\infty$ are bounded. Therefore,

$$\|G_t(t)\|_{s-2} \le CK_1(M)M(\|\eta_t\|_s + \|\eta\|_s).$$

This implies

$$\begin{aligned}\|\eta(t)\|_s + \|\eta_t(t)\|_s &\le \|\eta_0\|_s\big(1 + CTK_1(M) + CK_1(M)\big)\\ &\quad + \|\eta_1\|_s\big(CT + C\big)\\ &\quad + CT\int_0^t \big(K_1(M)\|\eta\|_s + CK_1(M)M(\|\eta\|_s + \|\eta_t\|_s) + \|\eta_t\|_s\big)d\tau.\end{aligned}$$

By Grönwall's lemma we obtain $\|\eta\|_s + \|\eta_t\|_s$ is bounded. Then, by (14) we immediately obtain (22). □

Now, we prove a result based on an assumption on the Helmholtz energy ψ. This result plays an important role in the proof of the main result, which is Theorem 4.

Lemma 4 *Assume that $s > 5/2$, $H(z) \ge -Cz^2$ and $g(z) \le Cz$. Then, the following estimate holds.*

$$-\int_{\mathbb{R}} \psi(S)dx \le C\|u(t)\|_{s-1}^2 + K,$$

where $K > 0$ is a finite constant.

Proof By the definition of ψ as $H(S) + Sh(S)$, since $h'(z) > 0$ for any z, we have that $\psi(S) \geq H(S)$. From the constitutive relation (1) we know that $S = h^{-1}(u_x + \nu u_{xt}) = g(u_x + \nu u_{xt})$. Then, also by the assumption $H(z) \geq -Cz^2$, we obtain

$$-\int_{\mathbb{R}} \psi(S)dx \leq -\int_{\mathbb{R}} H(S)dx \leq C\|S\|_2^2 = C\|g(u_x + \nu u_{xt})\|_2^2.$$

By the assumption on the growth of g, we can conclude that

$$-\int_{\mathbb{R}} \psi(S)dx \leq C\|u + \nu u_t\|_1^2 \leq C\|u + \nu u_t\|_{s-1}^2 \leq C\|u(t)\|_{s-1}^2 + C\nu\|u_t(t)\|_{s-1}^2.$$

However, we know that $u_t \in H^s(\mathbb{R})$. Since $s > 5/2$ this implies the required inequality with $K = C\nu M$ with M being the bound on $\|u_t\|_{s-1}^2$. □

We know prove the main theorem of this work by showing that the local-in-time solutions are global under the assumption that the initial energy is finite.

Theorem 4 *Assume that $s > 5/2$ and $u_0, u_1 \in H^s(\mathbb{R})$. Also assume that the initial energy $E(0)$ is finite. Then the system (15) has a global solution $u \in C^1([0, \infty), H^s(\mathbb{R}))$.*

Proof

$$\begin{aligned}
\frac{d}{dt}\|u_x(t)\|_{s-2}^2 &= 2\|u_x(t)\|_{s-2}\frac{d}{dt}\|u_x(t)\|_{s-2} \\
&= 2\|u(t)\|_{s-1}\frac{d}{dt}\|u(t)\|_{s-1} \\
&\leq 2\|u_x(t)\|_{s-2}\|u_t(t)\|_{s-1} \\
&\leq \|u_x(t)\|_{s-2}^2 + \|u_t(t)\|_{s-1}^2 \\
&\leq \|u_x(t)\|_{s-2}^2 + \|u_t(t)\|_2^2.
\end{aligned}$$

Now, by Proposition 24 we can have the estimate

$$\|u_t(t)\|_2^2 \leq \left(E(0) - \int_{\mathbb{R}} \psi(S)dx\right).$$

By Lemma 4 we obtain

$$\begin{aligned}
\frac{d}{dt}\|u_x(t)\|_{s-2}^2 &\leq \|u_x(t)\|_{s-2}^2 + \left(E(0) - \int_{\mathbb{R}} \psi(S)dx\right) \\
&\leq \|u_x(t)\|_{s-2}^2 + \left(E(0) + C\|u_x(t)\|_{s-2}^2 + K\right) \\
&\leq (E(0) + K) + (C + 1)\|u_x(t)\|_{s-2}^2.
\end{aligned}$$

By Grönwall's lemma we obtain $\|u_x(t)\|_{s-2}$ stays bounded in $[0, T)$. Since, $s > 5/2$ this implies $\|u_x(t)\|_\infty$ is bounded. Moreover, by

$$\|u_{xt}(t)\|_\infty \leq C\|u_{xt}(t)\|_{s-2} = C\|u_t(t)\|_{s-1} \leq C\|u_t(t)\|_2,$$

and by Theorem 24 we obtain $\|u_{xt}(t)\|_\infty$ is bounded. By Theorem 3, this implies a global solution. □

5 Revisiting the Smallness Assumptions

In order for our results to hold within the context of strain-limiting theory, we should revisit our assumptions on the smallness of the displacement gradient and its time derivative. In other words, while condition (23) is enough to provide global existence of solutions, in order for the theory to hold we must have that they are, in fact, sufficiently small. For this, we revisit the energy equality

$$\frac{d}{dt}\int_{\mathbb{R}} \left(\frac{1}{2}|u_t|^2 + \psi(S)\right) dx = -\nu \int_{\mathbb{R}} S_x^2\, dx. \tag{24}$$

As discussed earlier, this relation tells us that the energy, which is the integral on the left-hand side, decreases over time. Having the extra assumption on the initial energy allows us to show that our theory is compatible with the strain-limiting assumptions.

Proposition 2 *Assume that the initial energy satisfies*

$$E(0) < \delta, \tag{25}$$

for $\delta > 0$ sufficiently small, as well as the same being true for $\|u_0\|_2$. Then, $\|u_x(t)\|_\infty$ and $\|u_{xt}(t)\|_\infty$ are also sufficiently small.

Proof From (24) we can conclude that the energy is decreasing over time. By the assumption (25), we can conclude that $\|u_t(t)\|_2$ stays small for all times. However, for $s > 5/2$ we know that

$$\|u_{xt}\|_\infty \leq C\|u_{xt}\|_{s-2} = C\|u_t\|_{s-1} \leq C\|u_t\|_2,$$

which we know is small. Also, by the smallness assumption on the initial data, (14) implies $\|u(t)\|_2$ is sufficiently small. By a similar argument as above, we obtain $\|u_x(t)\|_\infty$ is sufficiently small as required. □

Theorem 5 *Assume that $\psi(z) \geq Cz$, $g(z) \geq Cz^2$, and $E(0) < \delta$. Then,*

$$\limsup_{t\to T^-}\big(\|u_x(t)\|_\infty + \|u_{xt}(t)\|_\infty\big) < \delta$$

implies $u \in X^s([0,\infty))$.

Proof From the smallness assumption on the initial energy and the fact that total energy is decreasing over time, we obtain

$$\int_{\mathbb{R}} \psi(S)dx < \delta.$$

By the assumptions we made on ψ and g we obtain $\|u_x + \nu u_{xt}\|_2^2 < \delta$. This gives

$$\|u_x\|_2^2 + \nu^2\|u_{xt}\|_2^2 + 2\nu \int_{\mathbb{R}} \frac{d}{dt}|u_x|^2 dx < \delta.$$

By Theorem 3 we know that $\|u_t\|_s$ is bounded for all times, which is enough for u to belong to $X^s([0,\infty))$ as long as $\|u\|_s$ stays small. Therefore, without loss of generality we can assume that $\|u_t\|_s \geq \sqrt{\delta}/\nu$. Then, we obtain

$$\|u_x\|_2^2 + \nu\frac{d}{dt}\|u_x\|^2 dx < \delta - \nu^2\|u_{xt}\|_2^2,$$

so that

$$\frac{d}{dt}\|u_x\|^2 dx \leq -\frac{1}{\nu}\|u_x\|_2^2,$$

which, by Grönwall lemma, implies that

$$\|u_x\|_2^2 \leq \|u_x(0,x)\|_2^2 e^{-\frac{t}{\nu}}.$$

Since $s > 5/2$, $\|u_x\|_2^2 \geq \|u\|_s^2$, and hence we can conclude that as t increases, $\|u\|_s$ stays small as required. □

References

1. M. Bulíček, J. Málek, K.R. Rajagopal, E. Süli, On elastic solids with limiting small strain: modelling and analysis. EMS Surv. Math. Sci. **1**(2), 283–332 (2014)
2. M. Bulíček, J. Málek, E. Süli, Analysis and approximation of a strain-limiting nonlinear elastic model. Math. Mech. Solids **20**(1), 92–118 (2015)
3. M. Bulíček, V. Patel, Y. Şengül, E. Süli, Existence of large-data global weak solutions to a model of a strain-limiting viscoelastic body. Commun. Pure Appl. Anal. **20**(5), 1931–1960 (2021)
4. M. Bulíček, V. Patel, Y. Şengül, E. Süli, Existence and uniqueness of global weak solutions to strain-limiting viscoelasticity with Dirichlet boundary data (submitted)
5. R. Bustamante, Some topics on a new class of elastic bodies. Proc. R. Soc. A **465**, 1377–1392 (2009)

6. R. Bustamante, K.R. Rajagopal, Solutions of some simple boundary value problems within the context of a new class of elastic materials. Int. J. Nonlinear Mech. **46**(2), 376–386 (2011)
7. J.C. Criscione, K.R. Rajagopal, On the modeling of the non-linear response of soft elastic bodies. Int. J. Nonlinear Mech. **56**, 20–24 (2013)
8. H.A. Erbay, Y. Şengül, Traveling waves in one-dimensional non-linear models of strain-limiting viscoelasticity. Int. J. Nonlinear Mech. **77**, 61–68 (2015)
9. H.A. Erbay, Y. Şengül, A thermodynamically consistent nonlinear model of one-dimensional strain-limiting viscoelasticity. Z. Angew. Math. Phys. **71**, 94 (2020)
10. H.A. Erbay, A. Erkip, Y. Şengül, Local existence of solutions to the initial-value problem for one-dimensional strain-limiting viscoelasticity. J. Differ. Equ. **269**, 9720–9739 (2020)
11. H. Itou, V.A. Kovtunenko, K.R. Rajagopal, Contacting crack faces within the context of bodies exhibiting limiting strains. JSIAM Lett. **9**, 61–64 (2017)
12. H. Itou, V.A. Kovtunenko, K.R. Rajagopal, On the states of stress and strain adjacent to a crack in a strain-limiting viscoelastic body. Math. Mech. Solids **23**(3), 433–444 (2018)
13. H. Itou, V.A. Kovtunenko, K.R. Rajagopal, Crack problem within the context of implicitly constituted quasi-linear viscoelasticity. Math. Mod. Methods Appl. Sci. **29**(2), 355–372 (2019)
14. K. Kannan, K.R. Rajagopal, G. Saccomandi, Unsteady motions of a new class of elastic solids. Wave Motion **51**, 833–843 (2014)
15. V. Kulvait, J. Málek, K.R. Rajagopal, Anti-plane stress state of a plate with a V-notch for a new class of elastic solids. Int. J. Fract. **179**(1–2), 59–73 (2013)
16. V. Kulvait, J. Málek, K.R. Rajagopal, Modeling gum metal and other newly developed titanium alloys within a new class of constitutive relations for elastic bodies. Arch. Mech. **69**(1), 223–241 (2017)
17. T. Mai, J.R. Walton, On monotonicity for strain-limiting theories of elasticity. Math. Mech. Solids **20**(2), 121–139 (2014)
18. R. Meneses, O. Orellana, R. Bustamante, A note on the wave equation for a class of constitutive relations for nonlinear elastic bodies that are not Green elastic. Math. Mech. Solids **23**(2), 148–158 (2018)
19. J. Merodio, K.R. Rajagopal, On constitutive equations for anisotropic nonlinearly viscoelastic solids. Math. Mech. Solids **12**, 131–147 (2007)
20. S. Montero, R. Bustamante, A. Ortiz-Bernardin, A finite element analysis of some boundary value problems for a new type of constitutive relation for elastic bodies. Acta Mech. **227**(2), 601–615 (2016)
21. A. Muliana, K.R. Rajagopal, A.S. Wineman, A new class of quasi-linear models for describing the nonlinear viscoelastic response of materials. Acta Mech. **224**, 2169–2183 (2013)
22. A. Ortiz, R. Bustamante, K.R. Rajagopal, A numerical study of a plate with a hole for a new class of elastic bodies. Acta Mech. **223**, 1971–1981 (2012)
23. A. Ortiz-Bernardin, R. Bustamante, K.R. Rajagopal, A numerical study of elastic bodies that are described by constitutive equations that exhibit limited strains. Int. J. Solids Struct. **51**, 875–885 (2014)
24. K.R. Rajagopal, On implicit constitutive theories. Appl. Math. **48**, 279–319 (2003)
25. K.R. Rajagopal, The elasticity of elasticity. Z. Angew. Math. Phys. **58**, 309–317 (2007)
26. K.R. Rajagopal, A note on a reappraisal and generalization of the Kelvin-Voigt model. Mech. Res. Commun. **36**, 232–235 (2009)
27. K.R. Rajagopal, On a new class of models in elasticity. J. Math. Comput. Appl. **15**(4), 506–528 (2010)
28. K.R. Rajagopal, Non-linear elastic bodies exhibiting limiting small strain. Math. Mech. Solids **16**(1), 122–139 (2011)
29. K.R. Rajagopal, Conspectus of concepts of elasticity. Math. Mech. Solids **16**, 536–562 (2011)
30. K.R. Rajagopal, On the nonlinear elastic response of bodies in the small strain range. Acta Mech. **225**, 1545–1553 (2014)
31. K.R. Rajagopal, A note on the linearization of the constitutive relations of non-linear elastic bodies. Mech. Res. Commun. **93**, 132–137 (2018)

32. K.R. Rajagopal, An implicit constitutive relation for describing the small strain response of porous elastic solids whose material moduli are dependent on the density. Math. Mech. Solids **26**(8), 1138–1146 (2021)
33. K.R. Rajagopal, G. Saccomandi, Shear waves in a class of nonlinear viscoelastic solids. Q. JI Mech. Appl. Math. **56**(2), 311–326 (2003)
34. K.R. Rajagopal, A.R. Srinivasa, A thermodynamic frame work for rate type fluid models. J. Non-Newtonian Fluid Mech. **88**, 207–227 (2000)
35. K.R. Rajagopal, J.R. Walton, Modeling fracture in the context of a strain-limiting theory of elasticity: a single anti-plane shear crack. Int. J. Fract. **169**, 39–48 (2011)
36. K.R. Rajagopal, A.S. Wineman, *Mechanical Response of Polymers: An Introduction* (Cambridge University Press, Cambridge, 2000)
37. K.R. Rajagopal, A.S. Wineman, A quasi-correspondence principle for quasi-linear viscoelastic solids. Mech. Time-Depend. Mater. **12**, 1–14 (2008)
38. Y. Şengül, One-dimensional strain-limiting viscoelasticity with an arctangent type nonlinearity. Appl. Eng. Sci. **7**, 100058 (2021)
39. Y. Şengül, Viscoelasticity with limiting strain. Discrete Contin. Dyn. Syst. S **14**(1), 57–70 (2021)

GENERIC for Dissipative Solids with Bulk–Interface Interaction

Marita Thomas and Martin Heida

1 Introduction

GENERIC, the acronym for General Equation of Non-Equilibrium Reversible–Irreversible Coupling, is a thermodynamical modeling framework originally introduced by Grmela and Öttinger in [11, 23] for thermodynamically closed systems with applications in fluid dynamics. In recent years, its versatility has been proved also for many other applications such as dissipative solids [13, 15, 17], complex and reactive fluids [21, 25, 33, 34], semiconductors and electro-chemistry [9, 14, 18], quantum mechanics [19], and thermodynamical multiscale processes [24]. A GENERIC system is characterized by a quintuple $(\mathcal{Q}, \mathcal{E}, \mathcal{S}, \mathbb{J}, \mathbb{K})$ consisting of a state space $\mathcal{Q}$, the two driving potentials: $\mathcal{E}$ the total energy and $\mathcal{S}$ the entropy, and two geometric structures: $\mathbb{J}$ a Poisson operator and $\mathbb{K}$ an Onsager operator. Herein, the triple $(\mathcal{Q}, \mathcal{E}, \mathbb{J})$ forms a Hamiltonian system characterizing the reversible contributions to the dynamics and the triple $(\mathcal{Q}, \mathcal{S}, \mathbb{K})$ forms an Onsager system accounting for the irreversible, dissipative contributions. These two triples are coupled in a GENERIC system under an additional constraint, the so-called noninteraction condition NIC, stating that $\mathbb{K}\mathrm{D}\mathcal{E} \equiv 0 \equiv \mathbb{J}\mathrm{D}\mathcal{S}$. In thermodynamically closed systems, the NIC automatically ensures conservation of energy and entropy production. The dynamics of the GENERIC system is then described by the evolution equation

$$\dot{q} = \mathbb{J}\mathrm{D}\mathcal{E}(q) + \mathbb{K}\mathrm{D}\mathcal{S}(q)\,,$$

M. Thomas (✉) · M. Heida
Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
e-mail: marita.thomas@wias-berlin.de; martin.heida@wias-berlin.de

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_15

which clearly displays the coupled evolution with reversible and dissipative contributions. The thermodynamical driving forces are the functional derivatives $\mathrm{D}\mathcal{E}(q)$ for reversible dynamics and $\mathrm{D}\mathcal{S}(q)$ for dissipative dynamics. In Sect. 2, we review the GENERIC framework for thermodynamically closed systems. It is the central aim of this work to extend the GENERIC framework to systems with bulk–interface interaction. These are systems composed of two (or more) subsystems $\Omega_\pm \subset \mathbb{R}^d$ coupled with each other along a joint interface $\Gamma \subset \mathbb{R}^{d-1}$ through which they exchange quantities like heat, stresses, mass, etc. Along Γ also additional processes may take place that are not only modeled by additional state variables solely defined on Γ with individual evolution laws on Γ but also driven by the interaction with the quantities from the bulk subdomains $\Omega_\pm$. While the compound $\Omega = \mathrm{int}\big(\overline{\Omega_+} \cup \overline{\Omega_-}\big)$ can be assumed to form a thermodynamically closed system, none of the two individual subsystems $\Omega_\pm$ nor the interface Γ do so. Each of these components alone is an open system. A first approach to the GENERIC framework for thermodynamically open systems was proposed in [22] using driving functionals and geometric structures for the bulk and the boundary components. Here, we follow this idea and, based on the definition of functional derivatives for functionals with bulk and interfacial contributions given in Sect. 3.1, we propose in Sect. 3.3 to regard the GENERIC formulation for bulk–interface processes in terms of a weak formulation. We also study the properties of geometric structures for systems with bulk–interface interaction in Sect. 3.2. It was observed in [15] for closed systems in dissipative solids that the validity of the NIC can be achieved with the aid of certain thermodynamic transformation maps. In Sect. 3.4, we show that this approach can also be applied to systems with bulk–interface interaction, again by exploiting the definition and structure of the functional derivatives involved in this transformation. We subsequently demonstrate the versatility of the weak formulation of GENERIC in Sect. 4 for thermo-viscoelastic materials experiencing delamination processes along Γ. It is shown that the weak formulation of GENERIC leads to well-known bulk equations and naturally provides interfacial coupling conditions along Γ.

2 The GENERIC Formalism for Closed Systems

Let $\mathcal{Q}$ be a Banach space and $\mathcal{V}$ a Hilbert space such that $\mathcal{Q} \hookrightarrow \mathcal{V} = \mathcal{V}^* \hookrightarrow \mathcal{Q}^*$ are dense. We denote for $q \in \mathcal{Q}$ and $q^* \in \mathcal{Q}^*$ the duality pairing by $\langle q^*, q\rangle_{\mathcal{Q}}$ and say that a linear operator $\mathbb{A} : \mathcal{Q}^* \to \mathcal{Q}$ is *symmetric*, respectively, *antisymmetric* if for every $q_1^*, q_2^* \in \mathcal{Q}^*$ it holds

$$\langle q_1^*, \mathbb{A}q_2^*\rangle_{\mathcal{Q}} = \langle q_2^*, \mathbb{A}q_1^*\rangle_{\mathcal{Q}}\,, \qquad \text{respectively,} \quad \langle q_1^*, \mathbb{A}q_2^*\rangle_{\mathcal{Q}} = -\langle q_2^*, \mathbb{A}q_1^*\rangle_{\mathcal{Q}}\,.$$

In most parts of the computations below, the reader may think of the Hilbert case $\mathcal{Q} = \mathcal{V}$. In this case, the above definitions coincide with the classical definitions $\mathbb{A} = \mathbb{A}^*$, respectively, $\mathbb{A} = -\mathbb{A}^*$. For a functional $\Phi : \mathcal{Q} \to \mathbb{R}$, we denote the

$$\text{Gâteaux derivative: } \delta\Phi : \boldsymbol{Q} \to \boldsymbol{Q}^*, \text{ i.e., the first variation,} \tag{1a}$$

$$\text{Fréchet derivative: } \mathrm{D}\Phi : \boldsymbol{Q} \to \boldsymbol{Q}^*\,, \tag{1b}$$

if they exist and we recall that $\mathrm{D}\Phi(q) = \delta\Phi(q)$ if $\mathrm{D}\Phi(q)$ exists. We finally define

$$\delta\Phi(q)[\tilde{q}] := \langle \delta\Phi(q), \tilde{q}\rangle_{\boldsymbol{Q}}\,, \qquad \mathrm{D}\Phi(q)[\tilde{q}] := \langle \mathrm{D}\Phi(q), \tilde{q}\rangle_{\boldsymbol{Q}}\,.$$

2.1 Hamiltonian Systems $(\boldsymbol{Q}, \mathcal{E}, \mathbb{J})$

In the spirit of Hamiltonian mechanics, a general Hamiltonian system accounts for reversible dynamics, only. The equations of motion are given by

$$\dot{q} = \mathbb{J}\mathrm{D}\mathcal{E}(q) \in \boldsymbol{Q}\,. \tag{2}$$

The driving potential of reversible dynamics is the total energy functional of the system $\mathcal{E} : \boldsymbol{Q} \to \mathbb{R}$, which may comprise kinetic, mechanical, chemical, electric, and thermal energy. The defining property for a Hamiltonian system is that the associated geometric structure $\mathbb{J}$ is a *Poisson structure*, i.e.,

$$\mathbb{J} : \boldsymbol{Q}^* \to \boldsymbol{Q} \text{ is antisymmetric and satisfies Jacobi's identity.} \tag{3}$$

More precisely, condition (3) ensures that the *Poisson bracket* $\{\cdot,\cdot\}$ defined by $\{\Phi_1, \Phi_2\} := \langle \mathrm{D}\Phi_1, \mathbb{J}\mathrm{D}\Phi_2\rangle_{\boldsymbol{Q}}$ for all $\Phi_j : \boldsymbol{Q} \to \mathbb{R}$ is an

$$\begin{aligned}&\text{antisymmetric bilinear form and satisfies Jacobi's identity, i.e.,}\\ &\forall \Phi_1, \Phi_2, \Phi_3 : \boldsymbol{Q} \to \mathbb{R} : \quad \{\Phi_1, \{\Phi_2, \Phi_3\}\} + \{\Phi_3, \{\Phi_1, \Phi_2\}\} + \{\Phi_2, \{\Phi_3, \Phi_1\}\} = 0\,.\end{aligned} \tag{4}$$

Moreover, the Poisson bracket fulfills the Leibniz rule:

$$\{\Phi_1\Phi_2, \Phi_3\} = \Phi_1\{\Phi_2, \Phi_3\} + \{\Phi_1, \Phi_3\}\Phi_2 \text{ for all } \Phi_1, \Phi_2, \Phi_3 : \boldsymbol{Q} \to \mathbb{R}\,. \tag{5}$$

Conditions (4) and (5) are the defining properties of a *symplectic* structure, which is the geometric structure underlying Hamiltonian mechanics, see, e.g., [1, Sect. 1.3]. Let us also mention that the requirement of Jacobi's identity provides a generalization of the commutativity of derivatives. Indeed, for $\boldsymbol{Q} = \boldsymbol{Q}_1 \times \boldsymbol{Q}_2$ and $\mathbb{J}$ in canonical form, i.e.,

$$\mathbb{J} := \begin{pmatrix} 0 & I_2 \\ -I_1 & 0 \end{pmatrix} \tag{6}$$

with $I_j : Q_j^* \to Q_j^*$ the identity operator and $Q_j^* = Q_k$ for $j \neq k \in \{1, 2\}$, it can be checked that fulfilling Jacobi's identity amounts to the validity of $\mathrm{D}_{q_1}\mathrm{D}_{q_2}\Phi_i = \mathrm{D}_{q_2}\mathrm{D}_{q_1}\Phi_i$. See, e.g., [6, 20, 34] for further discussion of Jacobi's identity. The antisymmetry of $\mathbb{J}$ implies $\langle q', \mathbb{J}q'\rangle_Q = -\langle q', \mathbb{J}q'\rangle_Q = 0$ for any $q' \in Q^*$ and conservation of energy along solutions of (2) follows immediately:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q(t)) = \langle \mathrm{D}\mathcal{E}(q), \dot{q}\rangle_Q = \langle \mathrm{D}\mathcal{E}(q), \mathbb{J}\mathrm{D}\mathcal{E}(q)\rangle_Q = 0\,. \tag{7}$$

2.2 Onsager Systems $(Q, \mathcal{S}, \mathbb{K})$ (Gradient Systems)

An Onsager system is related to the dynamics of irreversible, dissipative effects. The evolution equations read

$$\dot{q} = \mathbb{K}(q)\mathrm{D}\mathcal{S}(q) \in Q\,. \tag{8}$$

The driving functional is the total entropy $\mathcal{S}$, and the associated geometric structure is imposed by the so-called *Onsager operator* $\mathbb{K}$ with the properties:

$$\mathbb{K} \text{ is symmetric and positive semidefinite, i.e., } \langle \xi, \mathbb{K}\xi\rangle_Q \geq 0\,. \tag{9}$$

The symmetry of $\mathbb{K}$ reflects the *Onsager principle*, which states that the *rate* equals the *symmetric, positively semidefinite operator* $\mathbb{K}$ applied to the *thermodynamically conjugate force*. The positive semidefiniteness is a manifestation of the second law of thermodynamics, i.e., we have an increase of entropy via

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}(q(t)) = \langle \mathrm{D}\mathcal{S}(q), \dot{q}\rangle_Q = \langle \mathrm{D}\mathcal{S}(q), \mathbb{K}\mathrm{D}\mathcal{S}\rangle_Q \geq 0. \tag{10}$$

The properties of $\mathbb{K}$ are equivalent to the existence of a nonnegative, quadratic dual entropy-production (or dissipation) potential $\Psi^* = \Psi^*(q; \xi) = \frac{1}{2}\langle \xi, \mathbb{K}(q)\xi\rangle_Q$, see [16]. The dissipative structure can be generalized to non-quadratic potentials as follows:

$$\text{For all } q \in Q,\ \ \Psi^*(q; \cdot) \text{ is nonnegative, convex, and } \Psi^*(q; 0) = 0. \tag{11}$$

In particular, for all $q \in Q$, the potential Ψ^* is the convex conjugate of a nonnegative, convex dissipation potential $\Psi(q; \cdot) : Q \to [0, \infty]$ with the property $\Psi^*(q; 0) = 0$. The convex conjugate is defined by

$$\Psi^*(q; \xi) := \sup_{\tilde{q} \in Q} \big(\langle \xi, \tilde{q}\rangle_Q - \Psi(q; \tilde{q})\big) \quad \text{for all } (q, \xi) \in Q \times Q^*\,. \tag{12}$$

In this generalized setting, the evolution reads

$$\dot{q} \in \partial_\xi \Psi^*(q; \mathrm{D}\mathcal{S}(q)) \quad \text{in } Q$$

with $\partial_\xi \Psi^*(q, \xi)$ the (multivalued) subdifferential of $\Psi^*(q, \cdot)$ in $\xi \in Q^*$, i.e.,

$$\partial_\xi \Psi^*(q; \xi) = \{\tilde{q} \in Q,\ \Psi^*(q; \tilde{\xi}) - \Psi^*(q; \xi) \geq \langle \tilde{\xi} - \xi, \tilde{q} \rangle_Q \text{ for all } \tilde{\xi} \in Q^*\}\,. \tag{13}$$

We point to Sect. 3.2, where we discuss further implications of (11). Non-quadratic dual dissipation potentials arise, e.g., for generalized standard materials with a rate-independent evolution of the internal variable: here, $\Psi^*(q, \cdot)$ is positively 1-homogeneous. For further details, we refer to Sect. 4 as well as to [15].

2.3 GENERIC Systems $(Q, \mathcal{E}, \mathcal{S}, \mathbb{J}, \mathbb{K})$

A GENERIC system is a quintuple $(Q, \mathcal{E}, \mathcal{S}, \mathbb{J}, \mathbb{K})$, which couples a Hamiltonian system $(Q, \mathcal{E}, \mathbb{J})$ with an Onsager system $(Q, \mathcal{S}, \mathbb{K})$. The combined evolution equations have the form

$$\dot{q} = \mathbb{J}(q)\mathrm{D}\mathcal{E}(q) + \mathbb{K}(q)\mathrm{D}\mathcal{S}(q), \tag{14}$$

displaying the reversible and the irreversible part of the dynamics. Apart from the structural relations (3) and (9) of Hamiltonian and Onsager systems, a GENERIC system additionally has to satisfy the following crucial and nontrivial

$$\textbf{noninteraction condition, NIC}: \qquad \mathbb{K}\mathrm{D}\mathcal{E} \equiv 0 \qquad \text{and} \qquad \mathbb{J}\mathrm{D}\mathcal{S} \equiv 0\,. \tag{15}$$

If $\mathbb{K}$ arises from a (subdifferential of a) non-quadratic dual dissipation potential $\Psi^*(q; \cdot)$ as introduced in (11), then the NIC $\mathbb{K}\mathrm{D}\mathcal{E} = 0$ needs to be replaced by

$$\Psi^*(q; \xi + \lambda \mathrm{D}\mathcal{E}(q)) = \Psi^*(q; \xi) \ \text{ for all } q \in Q,\ \xi \in Q^*,\ \text{ and } \lambda \in \mathbb{R}\,. \tag{16}$$

We refer to Sect. 3.2 and to [15, Sec. 2.5] for more details.

Remark 1 (Direct Consequences of NIC) The NIC (15) ensures that the energy functional does not contribute to dissipative mechanisms and that the entropy functional does not contribute to reversible dynamics, i.e., every solution q of (14) satisfies:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q(t)) = \langle \mathrm{D}\mathcal{E}(q), \dot{q} \rangle = \langle \mathrm{D}\mathcal{E}(q), \mathbb{J}\mathrm{D}\mathcal{E} + \mathbb{K}\mathrm{D}\mathcal{S} \rangle = 0 + 0 = 0\,, \tag{17}$$

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{S}(q(t)) = \langle \mathrm{D}\mathcal{S}(q), \dot{q}\rangle = \langle \mathrm{D}\mathcal{S}(q), \mathbb{J}\mathrm{D}\mathcal{E} + \mathbb{K}\mathrm{D}\mathcal{S}\rangle = 0 + \langle \mathrm{D}\mathcal{S}, \mathbb{K}\mathrm{D}\mathcal{S}\rangle \geq 0\,. \tag{18}$$

Moreover, the NIC (15) guarantees the validity of the principle of maximum-entropy production. See, e.g., [11, 15, 23, 34] for more details. ⋆

3 GENERIC Formalism for Bulk–Interface Systems

Our investigations will be based on the following specific geometric setting:

Definition 1 (Geometric Setup and Notation) Let $\Omega_+, \Omega_- \subset \mathbb{R}^d$ be the reference domains of two bodies with $\Omega_+ \cap \Omega_- = \emptyset$, let $\partial\Omega_\pm$ their boundaries with outer unit normal $\mathrm{n}_\pm$ and joint interface $\Gamma := \partial\Omega_+ \cap \partial\Omega_-$. For better clarification, we further introduce the notation $\Gamma_\pm := \Gamma \cap \partial\Omega_\pm$, $\mathrm{B} := \Omega_+ \cup \Omega_-$, and $\Omega := \mathrm{int}(\overline{\Omega_+} \cup \overline{\Omega_-})$.

The geometric setup is described by one of the following two scenarios:

1. The two subdomains Ω_+ and Ω_- are connected with each other along $\Gamma \neq \emptyset$. The domain $\Omega := \mathrm{int}(\overline{\Omega_+} \cup \overline{\Omega_-})$ is again thermodynamically closed. Hence, the only exchange that $\Omega_\pm$ has with its surroundings is that with $\Omega_\mp$ along Γ.
2. As a special case of 1., the subdomain Ω_+ is surrounded by Ω_- so that $\Gamma = \partial\Omega_+$ and $\partial\Omega = \partial\Omega_- \backslash \Gamma$.

For a function $\phi : \Omega\backslash\Gamma \to \mathbb{R}$, we denote by $\phi_\pm = \phi|_{\Omega_\pm}$ its restriction to $\Omega_\pm$, by $\gamma_\pm\phi_\pm$ its trace from $\Omega_\pm$ onto $\partial\Omega_\pm$, and by $[\![\phi]\!] := (\gamma_+\phi_+ - \gamma_-\phi_-)$ its jump across Γ. Moreover, we use the short-hand notation $\phi_\gamma := (\gamma_+\phi_+, \gamma_-\phi_-)$, and if no confusion is possible, we abbreviate

$$\gamma_\pm\phi_\pm = \gamma\phi_\pm = \phi_\pm \quad \text{and} \quad \phi_\gamma = (\phi_+, \phi_-) \quad \text{on } \Gamma\,. \tag{19}$$

A similar notation is adopted for vector- and tensor-valued functions and has to be understood componentwise.

3.1 *Functional Calculus for Bulk–Interface Systems: Notation, Differentials, and ∗-Multiplication in the Setup of Definition 1*

States and Spaces In the setup of Definition 1, consider the function spaces $Q_\mathrm{B} := Q_+ \times Q_-$ and $Q = Q_\mathrm{B} \times Q_\Gamma$ with $Q_\pm$ being a Banach space (e.g., a Sobolev space) defined on $\Omega_\pm$ and with Q_Γ being a Banach space defined on Γ with dual spaces $Q_\pm^*, Q_\mathrm{B}^*$, and Q_Γ^*, and the dual pairings by $\langle\cdot,\cdot\rangle_{Q_\pm}$, respectively, $\langle\cdot,\cdot\rangle_{Q_\mathrm{B}}$, and $\langle\cdot,\cdot\rangle_{Q_\Gamma}$. The states $q = (q_\mathrm{B}, q_\Gamma)$ are composed of bulk states $q_\mathrm{B} \in Q_\mathrm{B}$ with

$q_\pm := q_{\mathrm{B}}|_{\Omega_\pm}$ and surface states $q_\Gamma \in \boldsymbol{Q}_\Gamma$. Here, a surface state q_Γ is supposed to have an own evolution equation on Γ so that its evolution is not solely governed from the bulk. For shorter notation, we also introduce the vector

$$q_{\Gamma\gamma} := (\gamma_+ q_+, \gamma_- q_-, q_\Gamma) \in \gamma\boldsymbol{Q}_{\mathrm{B}} \times \boldsymbol{Q}_\Gamma \,, \tag{20}$$

where $\gamma\boldsymbol{Q}_{\mathrm{B}}$ denotes the trace space corresponding to $\boldsymbol{Q}_{\mathrm{B}}$. More precisely, let the state $q = (q_{\mathrm{B}}, q_\Gamma) = (q_{\mathrm{B}_1}, \ldots, q_{\mathrm{B}_l}, q_{\Gamma_1}, \ldots, q_{\Gamma_m}) \in \boldsymbol{Q} = \boldsymbol{Q}_{\mathrm{B}} \times \boldsymbol{Q}_\Gamma$. If the state variable $q_{\mathrm{B}k}|_{\Omega_\pm}$ has a well-defined trace on Γ, then the kth component of $\gamma_\pm q_\pm$ in (20) is given by the trace $\gamma_\pm q_{k\pm}$. Instead, if the trace of this state variable does not exist, then we set the kth component of $\gamma_\pm q_\pm$ equal to zero in $q_{\Gamma\gamma}$. In this latter case, $\gamma_\pm \boldsymbol{Q}_{k\pm} = \{0\}$.

Example 1 (Notation (20)*)* Consider a system with bulk states $q_{\mathrm{B}} = (q_{\mathrm{B}1}, q_{\mathrm{B}2}) = (q_{1+}, q_{2+}, q_{1-}, q_{2-})$ and with one single interface state $q_{\Gamma 1}$. Assume $q_{1\pm}$ have well-defined traces on Γ, whereas $q_{2\pm}$ do not. Accordingly, $q_{\Gamma\gamma} = (\gamma_+ q_{1+}, 0, \gamma_- q_{1-}, 0, q_{\Gamma 1})$, where the entries 0 substitute the (non-existing) traces of $q_{2\pm}$. ⋆

Functionals and Their Derivatives Let $\Phi = \Phi_{\mathrm{B}} + \Phi_\Gamma : \boldsymbol{Q} \to \mathbb{R}$ denote an integral functional with density $\phi = (\phi_{\mathrm{B}}, \phi_\Gamma)$, which contains a bulk contribution ϕ_{B} and a surface contribution ϕ_Γ on Γ, i.e., for all states $q \in \boldsymbol{Q}$, it is

$$\Phi(q) = \int_{\Omega\setminus\Gamma} \phi_{\mathrm{B}}(q_{\mathrm{B}}, \nabla q_{\mathrm{B}})\,\mathrm{d}x + \int_\Gamma \phi_\Gamma(\gamma_+ q_+, \gamma_- q_-, q_\Gamma)\,\mathrm{d}\mathcal{H}^{d-1}\,. \tag{21}$$

Let again $q = (q_{\mathrm{B}}, q_\Gamma) = (q_{\mathrm{B}_1}, \ldots, q_{\mathrm{B}_l}, q_{\Gamma_1}, \ldots, q_{\Gamma_m}) = (q_j)_{j=1}^{l+m} \in \boldsymbol{Q} = \Pi_{j=1}^{l+m} \boldsymbol{Q}_j$ and q_k the kth component in this vector.

Restricting Φ to the affine space $\boldsymbol{Q}_k$, with $q_k \in \boldsymbol{Q}_k$, we use the following notation for the

$$\text{first variation wrt. } q_k : \quad \delta_{q_k}\Phi : \boldsymbol{Q} \to \boldsymbol{Q}_k^*, \tag{22a}$$

$$\text{functional derivative wrt. } q_k : \quad \mathrm{D}_{q_k}\Phi : \boldsymbol{Q} \to \boldsymbol{Q}_k^*, \tag{22b}$$

$$\text{partial derivative of a density } \phi = \phi(q) \text{ wrt. } q_k : \quad \partial_{q_k}\phi(q)\,. \tag{22c}$$

For Φ from (21) and a bulk state q_k, it is

$$\begin{aligned}\delta_{q_k}\Phi(q)[\tilde{q}_k] =& \delta_{q_k}\Phi_{\mathrm{B}}(q_{\mathrm{B}})[\tilde{q}_k] + \delta_{q_k}\Phi_\Gamma(q)[\tilde{q}_k] \\ =& \int_{\Omega\setminus\Gamma} \big(\partial_{q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}}, \nabla q_{\mathrm{B}})\tilde{q}_k + \partial_{\nabla q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}}, \nabla q_{\mathrm{B}}) \cdot \nabla\tilde{q}_k\big)\,\mathrm{d}x \\ &+ \sum_{i\in\{+,-\}} \int_\Gamma \partial_{q_{k_i}}\phi_\Gamma(\gamma_+ q_+, \gamma_- q_-, q_\Gamma)\gamma_{k_i}\tilde{q}_{k_i}\,\mathrm{d}\mathcal{H}^{d-1}\,,\end{aligned} \tag{22d}$$

and $\mathrm{D}_{q_k}\Phi(q)[\tilde{q}_k] = \mathrm{D}_{q_k}\Phi_{\mathrm{B}}(q_{\mathrm{B}})[\tilde{q}_k] + \mathrm{D}_{q_k}\Phi_\Gamma(q)[\tilde{q}_k]$

$$= \int_{\Omega\backslash\Gamma}\big(\partial_{q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}}) - \mathrm{div}\,\partial_{\nabla q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}})\big)\tilde{q}_k\,\mathrm{d}x \tag{22e}$$

$$+ \sum_{i\in\{+,-\}}\int_\Gamma\big(\partial_{\nabla q_k}\phi_{\mathrm{B}}(q_i,\nabla q_i)\cdot\mathrm{n}_i + 0\partial_{q_{k_i}}\phi_\Gamma(q_+,q_-,q_\Gamma)\big)\tilde{q}_{k_i}\,\mathrm{d}\mathcal{H}^{d-1},$$

i.e., by integration by parts, we have the equivalencec

$$\delta_{q_k}\Phi(q)[\tilde{q}_k] = \mathrm{D}_{q_k}\Phi(q)[\tilde{q}_k]\,. \tag{22f}$$

Similarly, for Φ from (21) and a surface state q_k, it is

$$\begin{aligned}\delta_{q_k}\Phi(q)[\tilde{q}_k] &= \int_\Gamma\big(\partial_{q_k}\phi_\Gamma(\gamma_+q_+,\gamma_-q_-,q_\Gamma)\tilde{q}_k + \partial_{q_{k_-}}\phi_\Gamma(q_+,q_-,q_\Gamma)\tilde{q}_{k_-}\big)\,\mathrm{d}\mathcal{H}^{d-1}\\ &= \mathrm{D}_{q_k}\Phi(q)[\tilde{q}_k]\,.\end{aligned} \tag{22g}$$

Given a sufficiently smooth function $\alpha:\Omega\backslash\Gamma\to\mathbb{R}$, we introduce the multiplication operation $*$ as follows:

$$\begin{aligned}\alpha{*}\mathrm{D}_{q_k}\Phi(q)[\tilde{q}_k] &= \int_{\Omega\backslash\Gamma}\big(\alpha\partial_{q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}}) - \mathrm{div}\,\big(\alpha\partial_{\nabla q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}})\big)\big)\tilde{q}_k\,\mathrm{d}x\\ &+ \sum_{i\in\{+,-\}}\int_\Gamma\big(\alpha\partial_{q_{k_i}}\phi_\Gamma(q_+,q_-,q_\Gamma) + \alpha\partial_{\nabla q_k}\phi_{\mathrm{B}}(q_i,\nabla q_i)\cdot\mathrm{n}_i\big)\tilde{q}_{k_+}\,\mathrm{d}\mathcal{H}^{d-1}\,.\end{aligned} \tag{23}$$

We will use the above notation for differentials and $*$ multiplication also for densities ϕ themselves. Exemplarily, we indicate this here for a bulk state q_k:

$$\begin{aligned}\delta_{q_k}\phi(q)[\Box] &= \big[\partial_{q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}})\Box + \partial_{\nabla q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}})\cdot\nabla\Box\big]_{\Omega\backslash\Gamma}\\ &+ \sum_{i\in\{+,-\}}\Big[\partial_{q_{k_i}}\phi_\Gamma(q_+,q_-,q_\Gamma)\Box\Big]_\Gamma\,,\end{aligned} \tag{24a}$$

$$\begin{aligned}\mathrm{D}_{q_k}\phi(q)[\Box] &= \big[\big(\partial_{q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}}) - \mathrm{div}\,\partial_{\nabla q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}})\big)\Box\big]_{\Omega\backslash\Gamma}\\ &+ \sum_{i\in\{+,-\}}\Big[\big(\partial_{q_{k_i}}\phi_\Gamma(q_+,q_-,q_\Gamma) + \partial_{\nabla q_k}\phi_{\mathrm{B}}(q_i,\nabla q_i)\cdot\mathrm{n}_i\big)\Box\Big]_\Gamma\,,\end{aligned} \tag{24b}$$

$$\begin{aligned}\alpha*\mathrm{D}_{q_k}\phi(q)[\Box] &= \big[\big(\alpha\partial_{q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}}) - \mathrm{div}\,\big(\alpha\partial_{\nabla q_k}\phi_{\mathrm{B}}(q_{\mathrm{B}},\nabla q_{\mathrm{B}})\big)\big)\Box\big]_{\Omega\backslash\Gamma}\\ &+ \sum_{i\in\{+,-\}}\Big[\big(\alpha\partial_{q_{k_i}}\phi_\Gamma(q_+,q_-,q_\Gamma) + \alpha\partial_{\nabla q_k}\phi_{\mathrm{B}}(q_i,\nabla q_i)\cdot\mathrm{n}_i\big)\Box\Big]_\Gamma\,.\end{aligned} \tag{24c}$$

In general, there is the following equivalence in a weak sense

$$\alpha * \mathrm{D}_{q_k}\phi(q)[\tilde{q}_k] = \alpha\delta_{q_k}\phi(q)[\tilde{q}_k]\,. \tag{25}$$

Dual Dissipation Potentials In the same manner, the notation introduced in (21)–(24) is also applied for dual dissipation potentials $\Psi^* = \Psi^*_{\mathrm{B}} + \Psi^*_{\Gamma} : \boldsymbol{Q} \times \boldsymbol{Q}^* \to [0, \infty]$ and the dual states $\xi := (\xi_{\mathrm{B}}, \xi_{\Gamma}) \in \boldsymbol{Q}^*$ with $\xi_{\mathrm{B}} \in \boldsymbol{Q}^*_{\mathrm{B}}$ and $\xi_{\Gamma} \in \boldsymbol{Q}^*_{\Gamma}$. In particular, we will see that the corresponding Onsager operator is thus given by

$$\mathbb{K}(q;\xi) := \mathrm{D}_{\xi_{\mathrm{B}}}\Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\xi_{\mathrm{B}}) + \mathrm{D}_{\xi}\Psi^*_{\Gamma}(q;\xi)\,. \tag{26}$$

Geometric Structures Also state-dependent geometric structures $\mathbb{J} = \mathbb{J}_{\mathrm{B}} + \mathbb{J}_{\Gamma} : \boldsymbol{Q} \times \boldsymbol{Q}^* \to \boldsymbol{Q}$, and $\mathbb{K} = \mathbb{K}_{\mathrm{B}} + \mathbb{K}_{\Gamma} : \boldsymbol{Q} \times \boldsymbol{Q}^* \to \boldsymbol{Q}$ are composed of bulk and interfacial contributions with $\boldsymbol{Q} = \boldsymbol{Q}_{\mathrm{B}} \times \boldsymbol{Q}_{\Gamma}$.

3.2 *Direct Implications for Geometric Structures*

Next, we discuss the defining properties of the geometric structures with bulk–interface interaction in more detail. In particular, we have the following:

Lemma 1 (Properties of Dual Dissipation Potentials) *Let the setup of Definition 1 and Sect. 3.1 be satisfied. Consider a dual dissipation potential* $\Psi^* = \Psi^*_{\mathrm{B}} + \Psi^*_{\Gamma} : \boldsymbol{Q} \times \boldsymbol{Q}^* \to [0, \infty]$ *of the form*

$$\Psi^*(q;\xi) = \int_{\Omega\backslash\Gamma} \psi^*_{\mathrm{B}}(q_{\mathrm{B}};\xi_{\mathrm{B}},\nabla\xi_{\mathrm{B}})\,\mathrm{d}x + \int_{\Gamma} \psi^*_{\Gamma}(q_{\Gamma\gamma};\xi_{\Gamma\gamma})\mathrm{d}\mathcal{H}^{d-1}\,. \tag{27}$$

1. *Assume that* $\Psi^*(q;\cdot) : \boldsymbol{Q}^* \to [0,\infty]$ *is convex for all* $q \in \boldsymbol{Q}$. *Hence, both* $\Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\cdot) : \boldsymbol{Q}^*_{\mathrm{B}} \to [0,\infty]$ *and* $\Psi^*_{\Gamma}(q;\cdot) : \boldsymbol{Q}^* \to [0,\infty]$ *are convex for all* $q = (q_{\mathrm{B}}, q_{\Gamma}) \in \boldsymbol{Q}$.
2. *Assume that* $\Psi^*(q;\cdot) : \boldsymbol{Q}^* \to [0,\infty]$ *is convex for all* $q \in \boldsymbol{Q}$ *and in addition also that* $\Psi^*(q;0) = 0$ *for all* $q \in \boldsymbol{Q}$. *Then, there holds*

$$\langle \xi, \tilde{q}\rangle_{\boldsymbol{Q}} \geq 0 \textit{ for all } (q,\xi) \in \boldsymbol{Q} \times \boldsymbol{Q}^* \textit{ and for all } \tilde{q} \in \partial_\xi \Psi^*(q;\xi)\,. \tag{28a}$$

Moreover, also $\Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\cdot)$ *and* $\Psi^*_{\Gamma}(q;\cdot)$ *satisfy*

$$\Psi^*_{\mathrm{B}}(q_{\mathrm{B}};0) = 0 \textit{ for all } q_{\mathrm{B}} \in \boldsymbol{Q}_{\mathrm{B}}\,, \qquad \Psi^*_{\Gamma}(q;0) = 0 \textit{ for all } q \in \boldsymbol{Q} \tag{28b}$$

$$\langle \xi_{\mathrm{B}}, \tilde{q}_{\mathrm{B}}\rangle_{\boldsymbol{Q}_{\mathrm{B}}} \geq 0 \textit{ for all } \xi_{\mathrm{B}} \in \boldsymbol{Q}^*_{\mathrm{B}},\ \tilde{q}_{\mathrm{B}} \in \partial_{\xi_{\mathrm{B}}} \Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\xi_{\mathrm{B}})\,, \tag{28c}$$

$$\langle \xi, \tilde{q}\rangle_{\boldsymbol{Q}} \geq 0 \textit{ for all } \xi \in \boldsymbol{Q}^*,\ \tilde{q} \in \partial_\xi \Psi^*_{\Gamma}(q;\xi)\,. \tag{28d}$$

Furthermore, if $\Psi^*(q;\cdot)$ is Gâteaux-differentiable for all $q \in \boldsymbol{Q}$, then also $\Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\cdot)$ and $\Psi^*_\Gamma(q;\cdot)$ are so, and vice versa.

3. *In addition to the prerequisites of 2., assume that for all $q \in \boldsymbol{Q}$, the potential $\Psi^*(q;\cdot) : \boldsymbol{Q}^* \to [0,\infty]$ is quadratic and Gâteaux-differentiable. Then $\mathbb{K}(q) = \mathrm{D}_\xi \Psi^*(q;\cdot) : \boldsymbol{Q}^* \to \boldsymbol{Q}$ is a linear, symmetric, and positively semidefinite operator and so are its bulk part $\mathbb{K}_{\mathrm{B}}(q_{\mathrm{B}}) = \mathrm{D}_{\xi_{\mathrm{B}}} \Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\cdot)$ and its boundary part $\mathbb{K}_\Gamma(q) = \mathrm{D}_\xi \Psi^*_\Gamma(q;\cdot)$.*

Proof

To 1. Convexity of $\Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\cdot)$ and $\Psi^*_\Gamma(q;\cdot)$ is equivalent to the convexity of the densities $\psi^*_{\mathrm{B}}(q_{\mathrm{B}};\cdot,\cdot)$, $\psi^*_\Gamma(q_{\Gamma\gamma};\cdot)$. Since these two densities have different supports, the assertion follows.

To 2. By the definition of the subdifferential for convex potentials (13), we deduce that $\Psi^*(q;0) - \Psi^*(q;\xi) \geq \langle \tilde{q}, -\xi\rangle_{\boldsymbol{Q}}$ for all $\xi \in \boldsymbol{Q}^*$, $\tilde{q} \in \partial_\xi \Psi^*(q;\xi)$. Using that $\Psi^*(q;0) = 0$ and rearranging terms result in (28a). Moreover, since the densities ψ^*_{B} and ψ^*_Γ have different supports, the first statement of (28c) and (28d) follows. Furthermore, by 1., both potentials $\Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\cdot)$ and $\Psi^*_\Gamma(q;\cdot)$ are convex. Thus, the second statement of (28c) and (28d) is obtained by repeating the argument for (28a). Again, since the densities ψ^*_{B} and ψ^*_Γ have different supports, the Gâteaux-differentiability of $\Psi^*(q;\cdot)$ is equivalent to the Gâteaux-differentiability of $\Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\cdot)$ and $\Psi^*_\Gamma(q;\cdot)$.

To 3. The potential $\Psi^*(q;\cdot)$ is quadratic and Gâteaux-differentiable if and only if $\Psi^*_{\mathrm{B}}(q_{\mathrm{B}};\cdot)$ and $\Psi^*_\Gamma(q;\cdot)$ are so. Hence, their derivatives are linear, symmetric operators. Positive semidefiniteness follows from (28c), respectively, (28d). □

At this point, we also address canonical Poisson structures for bulk–interface interaction with an immediate statement, cf. (6). With the aid of transformation maps, this finding will be transferred to the non-canonical case in Sect. 3.4, see Lemma 4.

Lemma 2 (Properties of Canonical Poisson Structures) *Let the setup of Definition 1 and Sect. 2 be satisfied. Furthermore, let $\mathbb{J}_{\mathrm{B}} : \boldsymbol{Q}^*_{\mathrm{B}} \to \boldsymbol{Q}_{\mathrm{B}}$ and $\mathbb{J}_\Gamma : \boldsymbol{Q}^*_{\mathrm{B}} \times \boldsymbol{Q}^*_\Gamma \to \boldsymbol{Q}_{\mathrm{B}} \times \boldsymbol{Q}_\Gamma$ be both in canonical form. Hence, $\mathbb{J}_{\mathrm{B}}$ and $\mathbb{J}_\Gamma$ are antisymmetric and satisfy the Leibniz rule as well as Jacobi's identity. Moreover, also $\mathbb{J} = \mathbb{J}_{\mathrm{B}} + \mathbb{J}_\Gamma$ is in canonical form, thus antisymmetric, and satisfies the Leibniz rule as well as Jacobi's identity.*

In addition, also the NIC 15 can be shown to hold true for geometric structures and functionals with bulk–interface interaction of the type introduced in Sect. 3.1. This also results in the validity of energy conservation and entropy production and is consistent with the fact that the coupled bulk–interface system is assumed to be thermodynamically closed, cf. Definition 1.

Lemma 3 (NIC for GENERIC Bulk–Interface Systems) *Under the prerequisites of Lemma 1, consider energy and entropy functionals of the form (21). Furthermore, assume that Ψ^* and $\mathcal{E}$ satisfy the generalized NIC (16). Then,*

$$\langle \mathrm{D}\mathcal{E}(q), \tilde{q}\rangle_{\boldsymbol{Q}} = 0 \textit{ for all } q \in \boldsymbol{Q}, \xi \in \boldsymbol{Q}^{*,} \textit{ and all } \tilde{q} \in \partial_\xi \Psi^*(q;\xi), \tag{29}$$

and thus, for solutions $q \in L^2(0,T;\boldsymbol{Q}) \cap H^1(0,T;\boldsymbol{Q}^*)$ *of* (14)*, the energy conservation and entropy production hold true, i.e.,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q(t)) = 0 \quad \textit{and} \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}(q(t)) \geq 0 \quad \textit{for all } t \in [0, T]. \tag{30}$$

Moreover, the generalized NIC (16) *as well as properties* (30) *hold true even separately for the bulk and boundary contributions of* $\mathcal{E} = \mathcal{E}_{\mathrm{B}} + \mathcal{E}_\Gamma$, $\mathcal{S} = \mathcal{S}_{\mathrm{B}} + \mathcal{S}_\Gamma$, *and* $\Psi^* = \Psi^*_{\mathrm{B}} + \Psi^*_\Gamma$.

Proof By the generalized NIC (16), we have for all $q \in \boldsymbol{Q}$, for all $\xi \in \boldsymbol{Q}^*$, for all $\tilde{q} \in \partial_\xi \Psi^*(q;\xi)$, and for all $\lambda \in \mathbb{R}$,

$$0 = \Psi^*(q;\xi + \lambda \mathrm{D}\mathcal{E}(q)) - \Psi^*(q;\xi) \geq \langle \lambda \mathrm{D}\mathcal{E}(q), \tilde{q}\rangle_{\boldsymbol{Q}}.$$

Choosing $\lambda = 1$ and $\lambda = -1$ gives (29). Now, energy conservation follows by direct calculation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q(t)) = \langle \mathrm{D}\mathcal{E}(q), \dot{q}\rangle_{\boldsymbol{Q}} = \langle \mathrm{D}\mathcal{E}(q), \mathbb{J}\mathrm{D}\mathcal{E}(q) + \tilde{q}\rangle_{\boldsymbol{Q}} = 0 + 0 = 0$$

using the chain rule, the antisymmetry of $\mathbb{J}$, cf. (3), and (29). With similar arguments, also the entropy production of the system is verified:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}(q(t)) = \langle \mathrm{D}\mathcal{S}(q), \dot{q}\rangle_{\boldsymbol{Q}} = \langle \mathrm{D}\mathcal{S}(q), \mathbb{J}\mathrm{D}\mathcal{E}(q) + \tilde{q}\rangle_{\boldsymbol{Q}} \geq 0 + 0,$$

where the first 0 arises by the antisymmetry of $\mathbb{J}$ together with the NIC $\mathbb{J}\mathrm{D}\mathcal{S}(q) = 0$ and the inequality is due to (28a). This finishes the proof of (30).

Since the densities ψ^*_{B} and ψ^*_Γ as well as E_{B} and E_Γ have different supports, the generalized NIC (16) has to be satisfied separately for $\Psi^*(q_{\mathrm{B}}; \mathrm{D}_{\mathrm{B}}\mathcal{E}_{\mathrm{B}}(q_{\mathrm{B}})$ and $\Psi^*_\Gamma(q; \mathrm{D}_q\mathcal{E}_\Gamma(q))$ in order to hold true for Ψ^* and $\mathcal{E}$. Accordingly, also the relations $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_{\mathrm{B}}(q_{\mathrm{B}}(t)) = 0$ and $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_\Gamma(q(t)) = 0$ as well as $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}_{\mathrm{B}}(q_{\mathrm{B}}(t)) \geq 0$ and $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}_\Gamma(q(t)) \geq 0$ are obtained separately for the bulk and boundary contributions. □

Remark 2 (Comparison with [22]) Lemmata 1 and 2 show that with the setup of Definition 1 and Sect. 3.1, the characteristic properties of GENERIC systems $(\boldsymbol{Q}, \mathcal{E}, \mathcal{S}, \mathbb{J}, \mathbb{K})$ are satisfied separately by the bulk system $(\boldsymbol{Q}_{\mathrm{B}}, \mathcal{E}_{\mathrm{B}}, \mathcal{S}_{\mathrm{B}}, \mathbb{J}_{\mathrm{B}}, \mathbb{K}_{\mathrm{B}})$ and by the surface system $(\boldsymbol{Q}, \mathcal{E}_\Gamma, \mathcal{S}_\Gamma, \mathbb{J}_\Gamma, \mathbb{K}_\Gamma)$. This finding essentially rests on our definition of the derivatives and operations (22)–(25) for functionals with bulk and boundary contributions, i.e., considering the derivatives as distributions rather than classical functions. When starting from the abstract definition of variations and functional derivatives (22a)– (22c) of functionals defined on Banach spaces, relations (22d)– (25) arise as a natural consequence. We refer to [22], where the boundary terms arising from the bulk contributions by integration by parts are attributed to the boundary part of the system. Consequently, neither the bulk nor the boundary system satisfies the characteristic properties of a GENERIC system, but the sum of the two does. We stress that our approach and [22] lead to the same bulk–interface systems. ⋆

3.3 Weak Form of GENERIC as a Formalism for Bulk–Interface Systems

Based on the definitions given in Definition 1 and Sect. 3.1, we now introduce a GENERIC formalism for systems with bulk–interface interaction and open systems in terms of a weak formulation. In this way, the bulk equations have to hold in a weak sense and the coupling conditions along Γ naturally appear also in a weak sense.

Definition 2 (Weak form of GENERIC for Bulk–Interface Systems) Let $(\mathcal{Q}, \mathcal{E}, \mathcal{S}, \mathbb{J}, \mathbb{K})$ be a system with bulk and surface contributions as described in Definition 1 and Sect. 3.1, with the mapping properties $\mathcal{E} = \mathcal{E}_{\mathrm{B}} + \mathcal{E}_\Gamma : \mathcal{Q} \to \mathbb{R}$, $\mathcal{S} = \mathcal{S}_{\mathrm{B}} + \mathcal{S}_\Gamma : \mathcal{Q} \to \mathbb{R}$, $\mathbb{J} = \mathbb{J}_{\mathrm{B}} + \mathbb{J}_\Gamma : \mathcal{Q}^* \to \mathcal{Q}$, and $\mathbb{K} = \mathbb{K}_{\mathrm{B}} + \mathbb{K}_\Gamma : \mathcal{Q}^* \to \mathcal{Q}$ with $\mathcal{Q} = \mathcal{Q}_{\mathrm{B}} \times \mathcal{Q}_\Gamma$. A weak formulation for $(\mathcal{Q}, \mathcal{E}, \mathcal{S}, \mathbb{J}, \mathbb{K})$ is given by

$$\begin{aligned}\langle \tilde{\xi}, \dot{q} \rangle_{\mathcal{Q}} = \langle \tilde{\xi}_{\mathrm{B}}, \dot{q}_{\mathrm{B}} \rangle_{\mathcal{Q}_{\mathrm{B}}} + \langle \tilde{\xi}_\Gamma, \dot{q}_\Gamma \rangle_{\mathcal{Q}_\Gamma} &= \langle \tilde{\xi}, \mathbb{J}\mathrm{D}\mathcal{E}(q) + \mathbb{K}\mathrm{D}\mathcal{S}(q) \rangle_{\mathcal{Q}} \\ &= \langle \tilde{\xi}_{\mathrm{B}}, \mathbb{J}_{\mathrm{B}}(q_{\mathrm{B}})\mathrm{D}_{q_{\mathrm{B}}}\mathcal{E}_{\mathrm{B}}(q_{\mathrm{B}}) + \mathbb{K}_{\mathrm{B}}(q_{\mathrm{B}})\mathrm{D}_{q_{\mathrm{B}}}\mathcal{S}_{\mathrm{B}}(q_{\mathrm{B}}) \rangle_{\mathcal{Q}_{\mathrm{B}}} \\ &\quad + \langle \tilde{\xi}_{\Gamma\gamma}, \mathbb{J}_\Gamma(q_{\Gamma\gamma})\mathrm{D}_{q_{\Gamma\gamma}}\mathcal{E}_\Gamma(q_{\Gamma\gamma}) + \mathbb{K}_\Gamma(q_{\Gamma\gamma})\mathrm{D}_{q_{\Gamma\gamma}}\mathcal{S}_\Gamma(q_{\Gamma\gamma}) \rangle_{\mathcal{Q}_{\Gamma\gamma}}\end{aligned} \tag{31}$$

for all $\tilde{\xi} = (\tilde{\xi}_{\mathrm{B}}, \tilde{\xi}_\Gamma) \in \tilde{\mathcal{Q}} \subset \mathcal{Q}^*$ for with $\tilde{\mathcal{Q}} = \tilde{\mathcal{Q}}_{\mathrm{B}} \times \tilde{\mathcal{Q}}_\Gamma$ a suitable space of test functions. Here, $q_{\Gamma\gamma}$ is defined as in (20) and $\tilde{\xi}_{\Gamma\gamma} = (\tilde{\xi}_{+\gamma}, \tilde{\xi}_{-\gamma}, \tilde{\xi}_\Gamma) \in (\gamma\mathcal{Q})^* \times \mathcal{Q}_\Gamma^*$.

In the example of heat conduction, we now illustrate how the weak form of GENERIC arises from the definition of functional derivatives for functionals with bulk–interface interaction and how the interfacial coupling naturally emerges from this weak form.

Example 2 (Heat Transfer for Bulk–Interface and Open Systems) In the following, we discuss the Onsager structure for heat conduction taking into account different interfacial Osager operators along Γ, thus resulting in different coupling conditions.

Heat Conduction in the Bulk $\Omega\backslash\Gamma = \Omega_+ \cup \Omega_-$ The dual dissipation potential in the bulk is defined as

$$\Psi_{\mathrm{B}}^*(\theta;\cdot) : \mathcal{Q}^* \to \mathbb{R}, \ \ \Psi_{\mathrm{B}}^*(\theta;\xi) := \sum_{i\in\{+,-\}} \int_{\Omega_i} \tfrac{\theta^2\kappa(\theta)}{2} \left|\nabla\left(\tfrac{\xi}{\mathrm{D}_\theta E}\right)\right|^2 \mathrm{d}x\,. \tag{32}$$

Hence, the first variation in ξ reads

$$\delta_\xi \Psi^*_{\mathrm{B}}(\theta;\xi)[\tilde{\xi}] := \sum_{i\in\{+,-\}} \int_{\Omega_i} \theta^2\kappa(\theta)\nabla\big(\tfrac{\xi}{\mathrm{D}_\theta E}\big)\cdot\nabla\big(\tfrac{\tilde{\xi}}{\mathrm{D}_\theta E}\big)\,\mathrm{d}x\,. \tag{33}$$

We formally distinguish it from the functional derivative $\mathrm{D}_\xi \Psi^*_{\mathrm{B}}(\theta;\xi)$, which is obtained from $\delta_\xi \Psi^*_{\mathrm{B}}$ by an integration by parts, i.e.,

$$\begin{aligned}\mathrm{D}_\xi \Psi^*_{\mathrm{B}}(\theta;\xi)[\tilde{\xi}] = &\sum_{i\in\{+,-\}} \int_{\Omega_i} -\operatorname{div}\theta^2\kappa(\theta)\nabla\big(\tfrac{\xi}{\mathrm{D}_\theta E}\big)\tfrac{\tilde{\xi}}{\mathrm{D}_\theta E}\,\mathrm{d}x\\ &+\sum_{i\in\{+,-\}} \int_{\Gamma_i} \gamma_i\Big(\theta_i^2\kappa(\theta_i)\nabla\big(\tfrac{\xi_i}{\mathrm{D}_\theta E_i}\big)\Big)\cdot \mathrm{n}_i\,\gamma_i\big(\tfrac{\tilde{\xi}_i}{\mathrm{D}_\theta E_i}\big)\,\mathrm{d}\mathcal{H}^{d-1}\\ &=:\langle \mathbb{K}_{\mathrm{B}}(\theta)\,\xi,\tilde{\xi}\rangle_{\boldsymbol{Q}}\,,\end{aligned}$$

and we introduce the Onsager operator

$$\mathbb{K}_{\mathrm{B}}(\theta) = \sum_{i\in\{+,-\}} \Big[\tfrac{-1}{\mathrm{D}_\theta E}\operatorname{div}\Big(\theta^2\kappa(\theta)\nabla\big(\tfrac{\square}{\mathrm{D}_\theta E}\big)\Big)\Big]_{\Omega_i} + \Big[\gamma_i\Big(\tfrac{\theta_i^2\kappa_i(\theta_i)}{\mathrm{D}_\theta E_i}\nabla\big(\tfrac{\square}{\mathrm{D}_\theta E}\big)\Big)\cdot \mathrm{n}_i\Big]_{\Gamma_i}\,. \tag{34}$$

Ideal Heat Transfer Across the Perfectly Conducting Interface Γ At the perfectly conducting interface Γ, all quantities are continuous, which implies

$$\gamma_+\tilde{\xi}_+ = \gamma_-\tilde{\xi}_- \;\text{ for all } \tilde{\xi}\in \boldsymbol{Q}^*, \tag{35a}$$

$$\gamma_+\Big(\tfrac{\theta_+^2\kappa_+(\theta_+)}{\mathrm{D}_\theta E_+}\nabla\big(\tfrac{\xi_+}{\mathrm{D}_\theta E_+}\big)\Big)\cdot \mathrm{n}_+ = -\gamma_-\Big(\tfrac{\theta_-^2\kappa_-(\theta_-)}{\mathrm{D}_\theta E_-}\nabla\big(\tfrac{\xi_-}{\mathrm{D}_\theta E_-}\big)\Big)\cdot \mathrm{n}_-\,. \tag{35b}$$

Furthermore, $\mathbb{K}_{\mathrm{B}}$ satisfies properties (9) as well as NIC (15).

Heat Transfer Across the Imperfect Interface Γ We assume that the heat transfer through Γ is regulated by the heat transfer coefficient $\hat{\kappa}_\Gamma(\gamma_+\theta_+,\gamma_-\theta_-)$. In this spirit, we introduce the quadratic dual dissipation potential along Γ, for every $\xi_\gamma\in\operatorname{dom}(\Psi_\Gamma(\theta_\gamma;\cdot))$,

$$\Psi^*_\Gamma(\theta_\gamma;\xi_\gamma) := \int_\Gamma \tfrac{\hat{\kappa}_\Gamma(\gamma_+\theta_+,\gamma_-\theta_-)}{2}\Big|\gamma_+\big(\tfrac{\xi_+}{\mathrm{D}_\theta E_+}\big)-\gamma_-\big(\tfrac{\xi_-}{\mathrm{D}_\theta E_-}\big)\Big|^2\,\mathrm{d}\mathcal{H}^{d-1}, \tag{36}$$

and we find for all $\xi_\gamma,\tilde{\xi}_\gamma\in\operatorname{dom}(\Psi_\Gamma(\theta_\gamma;\cdot))$ that

$$\begin{aligned}\mathrm{D}_{\xi_\gamma}\Psi^*_\Gamma(\theta_\gamma;\xi_\gamma)[\tilde{\xi}_\gamma] &= \int_\Gamma \hat{\kappa}_\Gamma(\theta_+,\theta_-)\Big(\big(\tfrac{\xi_+}{\mathrm{D}_\theta E_+}\big)-\big(\tfrac{\xi_-}{\mathrm{D}_\theta E_-}\big)\Big)\Big(\big(\tfrac{\tilde{\xi}_+}{\mathrm{D}_\theta E_+}\big)-\big(\tfrac{\tilde{\xi}_-}{\mathrm{D}_\theta E_-}\big)\Big)\,\mathrm{d}\mathcal{H}^{d-1}\\ &= \langle \mathbb{K}_\Gamma(\theta_\gamma)\xi_\gamma,\tilde{\xi}_\gamma\rangle_{\operatorname{dom}(\Psi_\Gamma(\theta_\gamma;\cdot))}\,.\end{aligned} \tag{37}$$

Clearly, $\mathbb{K}_\Gamma(\theta_\gamma)$ is symmetric and positively semidefinite provided that $\hat{\kappa}_\Gamma(\theta_\Gamma) \geq 0$. Also, NIC (15) holds true since for all $\tilde{\xi}_\gamma = (\tilde{\xi}_+, \tilde{\xi}_-)^\top$, we have

$$\begin{aligned} &\langle \mathbb{K}_\Gamma(\theta_\gamma)(\mathrm{D}_\theta E)_\Gamma, \tilde{\xi}_\gamma \rangle_{\mathrm{dom}(\Psi_\Gamma(\theta_\gamma;\cdot))} \\ &\quad = \int_\Gamma \hat{\kappa}_\Gamma(\theta_+, \theta_-)\Big(\tfrac{\mathrm{D}_\theta E_+}{\mathrm{D}_\theta E_+} - \tfrac{\mathrm{D}_\theta E_-}{\mathrm{D}_\theta E_-}\Big)\Big(\tfrac{\tilde{\xi}_+}{\mathrm{D}_\theta E_+} - \tfrac{\tilde{\xi}_-}{\mathrm{D}_\theta E_-}\Big)\,\mathrm{d}\mathcal{H}^{d-1} = 0\,. \end{aligned} \tag{38}$$

Thus, in view of (34) and (38), the Onsager operator of the full coupled system is

$$\mathbb{K}(\theta) = \mathbb{K}_\mathrm{B}(\theta) + \mathbb{K}_\Gamma(\theta_\gamma), \tag{39}$$

and $\mathbb{K}(\theta)$ is symmetric, positively semidefinite and satisfies the NIC (15).

Now, the evolution equation (8) can be understood in a weak form such that for a.a. $t \in (0, \mathsf{T})$ and for all $\tilde{\xi} \in \tilde{Q} = H^1(\Omega\backslash\Gamma)$, there holds

$$\begin{aligned} \langle \dot{\theta}, \tilde{\xi} \rangle_{\tilde{Q}} &= \langle \mathbb{K}(\theta)\mathrm{D}_\theta \mathcal{S}(\theta), \tilde{\xi} \rangle_{\tilde{Q}} \\ &= \langle \mathbb{K}_\mathrm{B}(\theta)\mathrm{D}_\theta \mathcal{S}(\theta), \tilde{\xi} \rangle_{H^1(\Omega\backslash\Gamma)} + \langle \mathbb{K}_\Gamma(\theta_\Gamma)\mathrm{D}_\theta \mathcal{S}_\Gamma(\theta), \tilde{\xi}_\Gamma \rangle_{H^{1/2}(\Gamma)}\,. \end{aligned} \tag{40}$$

For a closed system, the heat flux through the boundary is 0 pointwise, i.e.,

$$\theta^2\kappa(\theta)\nabla\big(\tfrac{\xi}{\mathrm{D}_\theta E}\big)\tfrac{1}{\mathrm{D}_\theta E} \cdot \nu_{\partial\Omega} = 0 \qquad \text{on} \quad \partial\Omega\,.$$

Hence, choosing test functions $\tilde{\xi} = \mathrm{D}_\theta E\,\hat{\xi}$ with $\hat{\xi} \in \tilde{Q}$ and using the Gibbs relation, there holds in a weak sense in $Q = \tilde{Q}^*$

$$\mathrm{D}_\theta E\,\dot{\theta} = -\operatorname{div}\big(\theta^2\kappa(\theta)\nabla\tfrac{1}{\theta}\big) \;\text{ in } \Omega\backslash\Gamma \tag{41a}$$

for a.a. $t \in (0, \mathsf{T})$, together with the following transmission conditions along Γ

$$\gamma_+\Big(\tfrac{\theta_+^2\kappa_+(\theta_+)}{\mathrm{D}_\theta E_+}\nabla\big(\tfrac{1}{\theta_+}\big)\Big) \cdot \mathrm{n}_+ = -\gamma_-\Big(\tfrac{\theta_-^2\kappa_-(\theta_-)}{\mathrm{D}_\theta E_-}\nabla\big(\tfrac{1}{\theta_-}\big)\Big) \cdot \mathrm{n}_-\,, \tag{41b}$$

$$\gamma_+\Big(\tfrac{\theta_+^2\kappa_+(\theta_+)}{\mathrm{D}_\theta E_+}\nabla\big(\tfrac{1}{\theta_+}\big)\Big) \cdot \mathrm{n}_+ = -\hat{\kappa}_\Gamma(\theta_\gamma)\big(\tfrac{1}{\theta_+} - \tfrac{1}{\theta_-}\big) = \tfrac{\hat{\kappa}_\Gamma(\theta_\gamma)}{\theta_+\theta_-}[\![\theta]\!]\,, \tag{41c}$$

complemented by homogeneous boundary conditions along $\partial\Omega$ and by an initial condition. We point out that the transmission conditions (41b) and (41c) are also obtained, e.g., in [8, 30] for interfaces in local equilibrium.

Ideal Heat Transfer Across the External Boundary $\partial\Omega_+ = \Gamma$ In the setting of scenario 2 from Definition 1, above considerations help to formulate proper boundary conditions for non-closed systems. In this case, Ω_+ is a bounded domain which is connected to a reservoir Ω_-. Evolution equation (41) then has to be satisfied only in Ω_+, whereas the part of the system on Ω_- is not of interest. For a perfectly conducting boundary Γ and for a given function h, we thus set

$$-\gamma_-\Big(\tfrac{\theta_-^2\kappa_-(\theta_-)}{\mathrm{D}_\theta E_-}\nabla\big(\tfrac{\xi_-}{\mathrm{D}_\theta E_-}\big)\Big)\cdot \mathrm{n}_- := h. \tag{42}$$

In other words, the inhomogeneous Neumann boundary condition h is implemented in the above deduced GENERIC system by appropriately adjusting the functions κ_- on Ω_-, $E_-(\theta_-)$ and by making an appropriate choice for ξ_-.

In case of an imperfectly conducting boundary, we are free to choose θ_- and the coefficient functions $\kappa_-(\theta_-)$, $E_-(\theta_-)$, and $\hat{\kappa}_\Gamma(\theta_\Gamma)$ for a given function h such that

$$-\tfrac{\theta_-^2\kappa_-(\theta_-)}{\mathrm{D}_\theta E_-}\nabla\big(\tfrac{1}{\theta_-}\big)\cdot \mathrm{n}_- = h \quad \text{and} \quad \tfrac{\hat{\kappa}_\Gamma(\theta_\Gamma)}{\theta_+\theta_-}[\![\theta]\!] = h\,. \tag{43}$$

Again, the inhomogeneous Neumann boundary condition is implemented in the above deduced GENERIC system by appropriately adjusting the coefficient functions on Ω_- and Γ and by making an appropriate choice for ξ_-.

In both cases (perfect or imperfect), this neither interferes with the symmetry of $\mathbb{K}$ nor with the validity of NIC (15), but to ensure positive semidefiniteness of $\mathbb{K}$, it may be necessary to restrict the choices of h. ⋆

We refer to Sect. 4 to see (31) in application for specific examples of bulk–interface systems related to delamination processes.

3.4 Tools for Dissipative Solids with Bulk–Interface Interaction

In [15, Sec. 2.4] and [34, Sec. 3.4], it was established that the GENERIC structure of thermodynamically closed systems is preserved under similarity transformations. In particular, this approach can be used to facilitate the verification of the structural properties of the system, such as the NIC (15). For this, first consider a thermodynamically closed system described by the states $q_\tau = (w, \tau) \in Q_\tau$, where τ represents the thermodynamic variable and $w \in \mathbb{R}^N$ collects the remaining state variables. Suitable choices for the thermodynamic variable $\tau \in \{\theta, U, S, E\}$ are the temperature θ, the internal energy density U, the entropy density S, or the total energy density E, which is given by the sum of kinetic and internal energy density. For an integral functional $\mathcal{H} : Q_\tau \to \mathbb{R}$ with density H (and H as a placeholder for E, U, S), we introduce the map

$$\mathrm{T}_{\tau\to H} : Q_\tau \to Q_H,\ q_\tau := (w, \tau) \mapsto q_H := (w, H) \tag{44}$$

and its inverse $\mathrm{T}_{H\to\tau} = \mathrm{T}_{\tau\to H}^{-1}$. The calculation of the Fréchet derivative of $\mathrm{T}_{H\to\tau}$ thus gives

$$\mathbb{L}_H := \mathrm{DT}_{H\to\tau}(q_H) = \mathrm{DT}_{\tau\to H}(q_\tau)^{-1} = \begin{pmatrix} I & 0 \\ \delta_w H(q_\tau) & \partial_\tau H(q_\tau) \end{pmatrix}^{-1} \tag{45}$$

and leads to the relations

$$\mathbb{L}_H = \begin{pmatrix} I & 0 \\ -\frac{1}{\partial_\tau H}\delta_w H & \frac{1}{\partial_\tau H} \end{pmatrix} \quad \text{and} \quad \mathbb{L}_H^* = \begin{pmatrix} I & -\frac{\square}{\partial_\tau H} * \mathrm{D}_w H \\ 0 & \frac{1}{\partial_\tau H} \end{pmatrix}.$$

More generally, for some linear operator $\mathbb{A}_H : \boldsymbol{Q}_w^* \to \boldsymbol{Q}_w^*$ with adjoint $\mathbb{A}_H^*$, we set

$$\mathbb{L}_H = \begin{pmatrix} \mathbb{A}_H & 0 \\ -\frac{\delta_w H \circ \mathbb{A}_H}{\partial_\tau H} & \frac{1}{\partial_\tau H} \end{pmatrix} \quad \text{and} \quad \mathbb{L}_H^* = \begin{pmatrix} \mathbb{A}_H^* & \mathbb{A}_H^* \circ (-\frac{\square}{\partial_\tau H} * \mathrm{D}_w H) \\ 0 & \frac{1}{\partial_\tau H} \end{pmatrix}. \tag{46}$$

In this way, there clearly holds

$$\mathbb{L}_H^* \mathrm{D}\mathcal{H} \equiv \mathbb{L}_H^* \mathrm{D}H = (0, 1)^\top, \tag{47}$$

and the NIC (15) is ensured by assuming that the Poisson and the Onsager operator of a GENERIC system in the variables q_τ can be composed as

$$\mathbb{J}(q_\tau) := \mathbb{L}_S \mathbb{J}^0 \mathbb{L}_S^* \quad \text{and} \quad \mathbb{K}(q_\tau) := \mathbb{L}_E \mathbb{K}^0 \mathbb{L}_E^* \quad \text{with} \quad \mathbb{J}^0 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0 = \mathbb{K}^0 \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{48}$$

Here, $\mathbb{J}^0 : \boldsymbol{Q}_S^* \to \boldsymbol{Q}_S$ is a Poisson structure and $\mathbb{K}^0 : \boldsymbol{Q}_E^* \to \boldsymbol{Q}_E$ is an Onsager operator on the state spaces $\boldsymbol{Q}_S$ and $\boldsymbol{Q}_E$ with homogeneous boundary conditions.

We observe now that similar relations can also be established for systems with bulk–interface interaction: following the notation of Sect. 3.1, we assume that the bulk energy and entropy densities are given through $E_{\mathrm{B}}(w_{\mathrm{B}}, \nabla w_{\mathrm{B}}, \tau_{\mathrm{B}}) = E_+(w_+, \nabla w_+, \tau_+) + E_-(w_-, \nabla w_-, \tau_-)$ and $S_{\mathrm{B}}(w_{\mathrm{B}}, \nabla w_{\mathrm{B}}, \tau_{\mathrm{B}}) = S_+(w_+, \nabla w_+, \tau_+) + S_-(w_-, \nabla w_-, \tau_-)$. For simplicity, we assume that all variables, particularly $\tau_\pm$, have well-defined traces on Γ. Since Ω_+ and Ω_- are disjoint, we can follow (48) and find the bulk operators

$$\mathbb{J}_{\mathrm{B}}(q_{\tau_{\mathrm{B}}}) := \mathbb{L}_{S_+} \mathbb{J}_+^0 \mathbb{L}_{S_+}^* + \mathbb{L}_{S_-} \mathbb{J}_-^0 \mathbb{L}_{S_-}^* \quad \text{with} \quad \mathbb{J}_\pm^0 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0 \quad \text{and} \tag{49a}$$

$$\mathbb{K}_{\mathrm{B}}(q_{\tau_{\mathrm{B}}}) := \mathbb{L}_{E_+} \mathbb{K}_+^0 \mathbb{L}_{E_+}^* + \mathbb{L}_{E_-} \mathbb{K}_-^0 \mathbb{L}_{E_-}^* \quad \text{with} \quad \mathbb{K}_\pm^0 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0\,. \tag{49b}$$

Now, $\mathbb{J}_\pm^0 : \boldsymbol{Q}_{S_\pm}^* \to \boldsymbol{Q}_{S_\pm}$ is a Poisson structure and $\mathbb{K}_\pm^0 : \boldsymbol{Q}_{E_\pm}^* \to \boldsymbol{Q}_{E_\pm}$ is an Onsager operator on the state spaces $\boldsymbol{Q}_{S_\pm}^* \to \boldsymbol{Q}_{S_\pm}$ and $\boldsymbol{Q}_{E_\pm}^* \to \boldsymbol{Q}_{E_\pm}$ that allows for inhomogeneous boundary conditions along Γ for the state variables $w_\pm$. In a similar spirit, also, attention has to be paid for the entries of $\mathbb{L}_{H_\pm}$ and $\mathbb{L}^*_{H_\pm}$: the operator $\mathbb{A}^*_H : \mathbf{Q}^*_w \to \mathbf{Q}^*_{H,w}$ acts on the dual of the state space $\mathbf{Q}_w$ which can accommodate inhomogeneous boundary conditions. Elements $\delta_{w_\pm} H_\pm, \mathrm{D}_{w_\pm} H_\pm$, or $\frac{1}{\partial_{\tau_\pm} H_\pm} * \mathrm{D}_{w_\pm} H_\pm$ represent functionals from the dual space $\mathbf{Q}^*_w$ and are thus characterized by bulk and trace terms by explicitly making use of relations (22e) and (23). In this way, the bulk operators also generate a trace contribution on Γ

$$\mathbb{J}_\pm(q_{\tau_\pm}) := \left[\mathbb{L}_{S_\pm}\mathbb{J}^0_\pm\mathbb{L}^*_{S_\pm}\right]_{\Omega_\pm} + \left[\mathbb{L}_{S_\pm}\mathbb{J}^0_\pm\mathbb{L}^*_{S_\pm}\right]_{\Gamma_\pm}, \tag{50a}$$

$$\mathbb{K}_\pm(q_{\tau_\pm}) := \left[\mathbb{L}_{E_\pm}\mathbb{K}^0_\pm\mathbb{L}^*_{E_\pm}\right]_{\Omega_\pm} + \left[\mathbb{L}_{E_\pm}\mathbb{K}^0_\pm\mathbb{L}^*_{E_\pm}\right]_{\Gamma_\pm}. \tag{50b}$$

On Γ, the interfacial energy and entropy densities E_Γ and S_Γ depend on the traces of the bulk states $\gamma_\pm q_\pm := (\gamma_\pm w_\pm, \gamma_\pm \tau_\pm)$ as well as on additional surface states $q_\Gamma := (w_\Gamma, \tau_\Gamma)$, collected in $q_{\Gamma\gamma} := (q_+, q_-, q_\Gamma)^\top$, cf. (20). Here, we wrote $q_\pm = \gamma_\pm q_\pm$ according to (19). In analogy to (44)– (48), using the surface density $H_\Gamma \in \{E_\Gamma, S_\Gamma\}$, we introduce for $i \in \{+, -, \Gamma\}$ with operators $\mathbb{A}_{w_i}$

$$\mathbb{L}_{H_\Gamma,i} = \begin{pmatrix} \mathbb{A}_{w_i} & 0 \\ -\frac{\delta_{w_i} H_\Gamma \circ \mathbb{A}_{w_i}}{\partial_{\tau_i} H_\Gamma} & \frac{1}{\partial_{\tau_i} H_\Gamma} \end{pmatrix}, \quad \mathbb{L}^*_{H_\Gamma,i} = \begin{pmatrix} \mathbb{A}^*_{w_i} & \mathbb{A}^*_{w_i}(-\frac{\Box}{\partial_{\tau_i} H_\Gamma} * \mathrm{D}_{w_i} H_\Gamma) \\ 0 & \frac{1}{\partial_{\tau_i} H_\Gamma} \end{pmatrix}, \tag{51}$$

where we typically choose $\mathbb{A}_{w_i} = I$, with I the identity, and define the following matrices as a cartesian product:

$$\mathbb{L}_{H_\Gamma} = \mathrm{diag}\left(\mathbb{L}_{H_\Gamma,+}, \mathbb{L}_{H_\Gamma,-}, \mathbb{L}_{H_\Gamma,\Gamma}\right), \tag{52a}$$

$$\mathbb{L}^*_{H_\Gamma} = \mathrm{diag}\left(\mathbb{L}^*_{H_\Gamma,+}, \mathbb{L}^*_{H_\Gamma,-}, \mathbb{L}^*_{H_\Gamma,\Gamma}\right). \tag{52b}$$

Similar to (47), this construction provides

$$\mathbb{L}^*_{H_\Gamma}\mathrm{D}\mathcal{H}_\Gamma \equiv \mathbb{L}^*_{H_\Gamma}\mathrm{D}H_\Gamma = (0, 1, 0, 1, 0, 1)^\top, \tag{53}$$

so that the NIC (15) for the interfacial geometric operators

$$\mathbb{J}(q_{\Gamma\gamma}) := \mathbb{L}_{S_\Gamma}\mathbb{J}^0_\Gamma\mathbb{L}^*_{S_\Gamma} \quad \text{and} \quad \mathbb{K}(q_{\Gamma\gamma}) := \mathbb{L}_{E_\Gamma}\mathbb{K}^0_\Gamma\mathbb{L}^*_{E_\Gamma} \tag{54}$$

can be ensured for interfacial Poisson and Onsager operators $\mathbb{J}^0_\Gamma, \mathbb{K}^0_\Gamma : (\gamma\mathbf{Q})^* \times \mathbf{Q}^*_\Gamma \to (\gamma\mathbf{Q}) \times \mathbf{Q}_\Gamma$ with the property

$$\mathbb{J}^0_\Gamma(0, 1, 0, 1, 0, 1)^\top = 0 = \mathbb{K}^0_\Gamma(0, 1, 0, 1, 0, 1)^\top. \tag{55}$$

We further observe that the above strategy can be extended to more general dissipative mechanisms modelled by (non-quadratic) convex dual dissipation potentials as discussed in Lemma 1. In this setting, NIC (15) is ensured if

$$\partial_{q_{\mathrm{B}}}\Psi^{0*}_{\mathrm{B}}(q_{\mathrm{B}};(0,1)^\top)\ni 0 \quad\text{and}\quad \partial_{q_{\Gamma\gamma}}\Psi^{0*}_{\Gamma}(q_{\Gamma\gamma};(0,1,0,1,0,1)^\top)\ni 0 \tag{56}$$

for some dual dissipation potentials Ψ^{0*}_{B} and Ψ^{0*}_{Γ} with the properties established in Lemma 1. Moreover, it should be mentioned that also gradients of surface states can be taken into account in interfacial densities E_Γ and S_Γ. This is addressed in Sect. 4.

In view of [34, Sec. 3.4] and [15] and based on the afore discussion as well as on the results obtained in Sect. 3.2, the following statements can be concluded.

Lemma 4 (Properties of Geometric Structures Under Transformations) *Let the prerequisites of Lemmata 1 and 2 be satisfied, and consider transformation operators $\mathbb{L}_{H_\pm}$ and $\mathbb{L}_{H_\Gamma}$ of the type (46)– (49) and (52). Then, the following statements hold true:*

1. *Let $\mathbb{J}^0=\mathbb{J}^0_{\mathrm{B}}+\mathbb{J}^0_{\Gamma}$ be a Poisson operator in canonical form. Then, the transformed operators $\mathbb{J}_{\mathrm{B}}(q_{\tau_{\mathrm{B}}})=\mathbb{L}_{S_{\mathrm{B}}}\mathbb{J}^0_{\mathrm{B}}\mathbb{L}^*_{S_{\mathrm{B}}}$ and $\mathbb{J}_\Gamma(q_{\tau_{\Gamma\gamma}})=\mathbb{L}_{S_\Gamma}\mathbb{J}^0_\Gamma\mathbb{L}^*_{S_\Gamma}$ are also Poisson operators. If $\mathbb{J}^0_{\mathrm{B}}$ and $\mathbb{J}^0_\Gamma$ have properties (49a) and (55), then $\mathbb{J}_{\mathrm{B}}(q_{\tau_{\mathrm{B}}})$ and $\mathbb{J}_\Gamma(q_{\tau_{\Gamma\gamma}})$ satisfy the NIC (15) for $S=S_{\mathrm{B}}+S_\Gamma$.*
2. *Let $\Psi^{0*}=\Psi^{0*}_{\mathrm{B}}+\Psi^{0*}_{\Gamma}$ be a dual dissipation potential with the properties of L. 1, Items 1 and 2. Then, the transformed potentials $\Psi^*_{\mathrm{B}}(q_{\tau_{\mathrm{B}}};\xi_{\mathrm{B}})=\Psi^{0*}_{\mathrm{B}}(\mathbb{L}^*_{E_{\mathrm{B}}}q_{\tau_{\mathrm{B}}};\mathbb{L}^*_{E_{\mathrm{B}}}\xi_{\mathrm{B}})$ and $\Psi^*_\Gamma(\mathbb{L}_{E_\Gamma}q_{\tau_{\Gamma\gamma}};\mathbb{L}_{E_\Gamma}\xi_{\Gamma\gamma})=\Psi^{0*}_\Gamma(\mathbb{L}^*_{E_\Gamma}q_{\tau_{\Gamma\gamma}};\mathbb{L}^*_{E_\Gamma}\xi_{\Gamma\gamma})$ also have properties of L. 1, Items 1 & 2. If Ψ^{0*}_{B} and Ψ^{0*}_Γ have properties (56), then the transformed $\Psi^*_{\mathrm{B}}(q_{\tau_{\mathrm{B}}})$ and $\Psi^*_\Gamma(q_{\tau_{\Gamma\gamma}})$ satisfy the NIC (15) for $E=E_{\mathrm{B}}+E_\Gamma$.*

Remark 3 (Gibbs' Relation) Calculating the products $\mathbb{L}^*_S\mathrm{D}\mathcal{E}$ and $\mathbb{L}^*_E\mathrm{D}\mathcal{S}$, which appear in the above GENERIC formalism, leads to expressions of the form $\frac{\partial_\tau E}{\partial_\tau S}$. For $\tau=\theta$, this is the so-called

$$\text{Gibbs' relation: } \frac{\partial_\theta\mathcal{S}}{\partial_\theta\mathcal{E}}=\frac{1}{\theta}\,. \tag{57}$$

Mielke [15] demonstrates the generalization $\frac{\partial_\tau\mathcal{S}}{\partial_\tau\mathcal{E}}=\frac{1}{\theta}$, which we will frequently use in Sect. 4 and see Example 3 below. ⋆

Example 3 (Specific Choice of Driving Functionals in Thermoelasticity) In case of $\tau=\theta$, a specific choice for the bulk and surface functionals matching with Gibbs' relation (57) is given by

$$U_{\mathrm{B}}(e,\theta)=W(e)-\phi_0(\theta)+\theta\phi_0'(\theta),\qquad S_{\mathrm{B}}(e,\theta)=\phi_0'(\theta)-\mathbb{B}:e, \tag{58a}$$

$$U_\Gamma(\gamma_+\theta_+,\gamma_-\theta_-)=\tfrac{1}{4}(\gamma_+\theta_+^2+\gamma_-\theta_-^2),\qquad S_\Gamma(\gamma_+\theta_+,\gamma_-\theta_-)=\tfrac{1}{2}(\gamma_+\theta_++\gamma_-\theta_-), \tag{58b}$$

e.g., with $\phi_0'(\theta) = c_\mathrm{V}\ln\theta$ and a matrix $\mathbb{B} \in \mathbb{R}^{d\times d}$. Clearly, both the bulk functionals in (58a) and the surface functionals (58b) satisfy Gibbs' relation (57).

For the transformation $\mathrm{T}_{\theta\to S}$, we have for the bulk terms $\tilde\theta_\mathrm{B} = \mathrm{T}^{-1}_{\theta\to S}(S_\mathrm{B}) = (\phi_0')^{-1}(S_\mathrm{B} + \mathbb{B} : e)$ and

$$\tilde U_\mathrm{B}(e, S_\mathrm{B}) = W(e) - \phi_0\big((\phi_0')^{-1}(S_\mathrm{B} + \mathbb{B} : e)\big) + \big((\phi_0')^{-1}(S_\mathrm{B} + \mathbb{B} : e)\big)(S_\mathrm{B} + \mathbb{B} : e),$$

which again results in Gibbs' relation $\partial_{S_\mathrm{B}}\tilde U_\mathrm{B}(e, S_\mathrm{B}) = (\phi_0')^{-1}(S_\mathrm{B} + \mathbb{B} : e) = \tilde\theta_\mathrm{B}$. We further deduce that $\gamma_\pm\tilde\theta_\pm = 2\gamma_\pm S_\pm$ and $\tilde U_\Gamma(\gamma_\pm S_+, \gamma_\pm S_-) = S_+^2 + S_-^2$. This also provides the interfacial Gibbs' relation $\partial_{\gamma_\pm S_\pm} U_\Gamma = 2\gamma_\pm S_\pm = \gamma_\pm\tilde\theta_\pm$.

Similarly, for the transformation $\mathrm{T}_{\theta\to U}$ it is $\hat\theta_\mathrm{B} = \mathrm{T}^{-1}_{\theta\to U}(U_\mathrm{B}) = h^{-1}(U_\mathrm{B} - W(e))$, where we have set $h(\theta_\mathrm{B}) := -\phi_0(\theta) + \theta\phi_0'(\theta)$, and $\hat S_\mathrm{B}(e, U_\mathrm{B}) = \phi_0'(h^{-1}(\tilde U_\mathrm{B} - W(e)))$. Direct calculation again gives the Gibbs' relation for the bulk, since $\partial_{U_\mathrm{B}}\hat S_\mathrm{B}(e, U_\mathrm{B}) = \phi_0''(\hat\theta_\mathrm{B})\partial_{U_\mathrm{B}}\hat\theta_\mathrm{B}$ and $\partial_{U_\mathrm{B}}\hat\theta_\mathrm{B} = 1/h' = 1/(\hat\theta_\mathrm{B}\phi_0''(\hat\theta_\mathrm{B}))$. Along Γ, we now have $\gamma_\pm\hat\theta_\pm = 2\sqrt{\gamma_\pm U_\pm}$ and $S_\Gamma(\gamma_+ U_+, \gamma_- U_-) = \sqrt{\gamma_+ U_+} + \sqrt{\gamma_- U_-}$, so that indeed $\partial_{\gamma_\pm U_\pm} S_\Gamma + 1/(2\sqrt{\gamma_\pm U_\pm}) = 1/(\gamma_\pm\hat\theta_\pm)$, which confirms the interfacial Gibbs' relation also for the choice $\tau = U$. Since $E = E_\mathrm{kin} + U$, the same calculations also verify Gibbs' relation for $\tau = E$. ⋆

4 Delamination Processes in Thermo-viscoelastic Materials

We now consider a composite consisting of two thermo-viscoelastic bodies glued together along the interface Γ by an elastic adhesive. This adhesive can experience damage; in other words, delamination may evolve along Γ. In the spirit of generalized standard materials [12], this process is modeled with the aid of an internal variable, the delamination variable $z : [0, \mathsf{T}] \times \Gamma \to [0, 1]$ solely defined on Γ to account for the degradation state of the glue. In particular, $z(t, x) = 1$ means that the glue is fully intact in the interfacial point $x \in \Gamma$ at time $t \in [0, \mathsf{T}]$, whereas $z(t, x) = 0$ means that the glue is completely ineffective in $(t, x) \in [0, \mathsf{T}] \times \Gamma$. In this way, the set $\mathrm{C}(t) := \{x \in \Gamma,\ z(t, x) = 0\}$ describes the crack set. This meaning of z is connected to adhesive delamination and brittle fracture processes in the sense of Griffith [10], see, e.g., also [7], rather than to so-called cohesive zone models in the sense of Barenblatt [2]. For the latter type of models the internal variable has a different meaning as it is introduced to keep track of the history of displacement jumps across Γ, for example, in the form $z(t, x) := \max_{s\in[0,t]}[\![u(s, x)]\!] \cdot \mathrm{n}$. We refer to [31, 32] for a comparison of these different types of modeling approaches and to the references therein for an analytical treatment of cohesive zone models in absence of thermal effects.

The State Vector The vector of bulk state variables is given by $q_\mathrm{B} := (u, p, \tau)$. Here, $u : \Omega\backslash\Gamma \to \mathbb{R}^d$ are the displacements, $p = \varrho\dot u : \Omega\backslash\Gamma \to \mathbb{R}^d$ denotes the momentum with a given mass density $\varrho > 0$, and $\tau : \Omega\backslash\Gamma \to \mathbb{R}$, $\tau \in \{E, S, U, \theta\}$, is a thermodynamic state variable, where we omit the index $_\mathrm{B}$ on the individual variables. Along Γ, we have the traces $u_\pm$, $p_\pm$, and $\tau_\pm$ and the surface state variables, the delamination variable z, and the interfacial thermodynamic variable τ_Γ. Hence, writing $w := (u, p)$, the vector of interfacial variables is composed by

$$q_{\Gamma\gamma} = \gamma q_\mathrm{B} = (w_+, \tau_+, w_-, \tau_-, z, \tau_\Gamma)^\top = (u_+, p_+, \tau_+, u_-, p_-, \tau_-, z, \tau_\Gamma)^\top \,. \tag{59}$$

Prototypical Driving Functionals This choice of variables is complemented by an ansatz for the potentials of the form

$$E_\mathrm{B}(q_\mathrm{B}) := \tfrac{1}{2\varrho}|p|^2 + U_\mathrm{B}(e(u), \tau) - f \cdot u \qquad \text{in } \Omega\backslash\Gamma, \tag{60a}$$

$$U_\mathrm{B}(e(u), \tau) := U_\mathrm{B}^\mathrm{el}(e(u)) + U_\mathrm{B}^\mathrm{th}(\tau) \qquad \text{in } \Omega\backslash\Gamma, \tag{60b}$$

$$S_\mathrm{B}(q_\mathrm{B}) := S_\mathrm{B}^\mathrm{th}(e(u), \tau) \qquad \text{in } \Omega\backslash\Gamma, \tag{60c}$$

$$E_\Gamma(\gamma q_\mathrm{B}) := \tfrac{1}{2}\big(\gamma_+ U_\mathrm{B}^\mathrm{th}(\tau) + \gamma_- U_\mathrm{B}^\mathrm{th}(\tau)\big) + U_\Gamma^\mathrm{el}(z, [\![u]\!]) + U_\Gamma^\mathrm{th}(\tau_\Gamma) \qquad \text{on } \Gamma\,, \tag{60d}$$

$$S_\Gamma(\gamma q_\mathrm{B}) := \tfrac{1}{2}\big(\gamma_+ S_\mathrm{B}^\mathrm{th}(\tau) + \gamma_- S_\mathrm{B}^\mathrm{th}(\tau)) + S_\Gamma^\mathrm{th}(\tau_\Gamma\big) \qquad \text{on } \Gamma\,, \tag{60e}$$

where $e(u) := \frac{1}{2}(\nabla u + \nabla u^\top)$ is the linearized strain tensor and $[\![u]\!]$ is the displacement jump as introduced in Definition 1. We assume that Gibbs' relation is satisfied both in the bulk and on the interface, i.e.,

$$\frac{\partial_\tau E_\mathrm{B}}{\partial_\tau S_\mathrm{B}} = \theta_\mathrm{B}\,, \quad \frac{\partial_{\tau_\Gamma} E_\Gamma}{\partial_{\tau_\Gamma} S_\Gamma} = \theta_\Gamma \quad \text{and} \quad \frac{\partial_{\tau_\pm} E_\Gamma}{\partial_{\tau_\pm} S_\Gamma} = \gamma_\pm\theta_\pm\,, \tag{61}$$

see also Example 3 below for a specific choice matching with (61).

Underlying Poisson and Onsager Structures We introduce the bulk Poisson operator and bulk dual dissipation potential

$$\mathbb{J}_\mathrm{B}^0 := \begin{pmatrix} 0 & I & 0 \\ -I & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \Psi_\mathrm{B}^{0*}(q_\mathrm{B}; \xi_\mathrm{B}) := \Psi_\mathrm{V}^{0*}(q_\mathrm{B}; \xi_p) + \Psi_\mathrm{H}^{0*}(q_\mathrm{B}; \xi_\tau)\,. \tag{62a}$$

Here, $\mathbb{J}_\mathrm{B}^0$ features the reversible contribution to the evolution of the pair (u, p), whereas the bulk thermodynamic variable τ evolves solely dissipative and thus has no nonzero entries in $\mathbb{J}_\mathrm{B}^0$. In this way, there clearly holds $\mathbb{J}_\mathrm{B}^0(0, 0, 1)^\top = 0$, as a prerequisite to satisfy NIC according to (49). In turn, the dissipative evolution of τ is ruled by the dual dissipation potential $\Psi^*_{\mathrm{H}}(q_{\mathrm{B}};\xi_\tau)$ with density ψ^*_{H} of the form

$$\Psi^{0*}_{\mathrm{H}}(\theta;\xi_\tau) := \int_{\Omega\setminus\Gamma} \frac{\theta^2\kappa(\theta)}{2}\big|\nabla\xi_\tau\big|^2\,\mathrm{d}x\,, \tag{62b}$$

resulting in the bulk Onsager operator for heat transport

$$\mathbb{K}^0_{\mathrm{H_B}}(\theta)\xi_{\tau_{\mathrm{B}}} = \sum_{i\in\{+,-\}} \Big[-\operatorname{div}\big(\theta^2\kappa(\theta)\nabla\xi_\tau\big)\Big]_{\Omega_i} + \Big[\gamma_i\big(\theta^2\kappa(\theta)\nabla\xi_\tau\big)\mathrm{n}_i\Big]_{\Gamma_i}. \tag{62c}$$

This operator satisfies $\mathbb{K}^0_{\mathrm{H_B}}(\theta)1 = 0$, a prerequisite to satisfy the NIC (49). In addition, the bulk dual dissipation potential features Kelvin–Voigt viscosity of the form

$$\Psi^{0*}_{\mathrm{V}}(q_{\mathrm{B}};\xi_p) := \int_{\Omega\setminus\Gamma} \frac{1}{2}\mathbb{D}(q_{\mathrm{B}})\xi_p : \xi_p\,\mathrm{d}x \tag{62d}$$

$$\text{with}\quad \mathbb{K}^0_{\mathrm{V}}(q_{\mathrm{B}})\xi_p := \sum_{i\in\{+,-\}} \Big[\mathbb{D}(q_{\mathrm{B}})\xi_p\Big]_{\Omega_i}\,, \tag{62e}$$

the viscous Onsager operator. The full bulk Onsager operator is thus given as

$$\mathbb{K}^0_{\mathrm{B}}(q_{\mathrm{B}}) = \operatorname{diag}\Big(0,\ \mathbb{K}^0_{\mathrm{V}}(q_{\mathrm{B}}),\ \mathbb{K}^0_{\mathrm{H_B}}(\theta)\Big)\,. \tag{62f}$$

On the interface Γ, dissipation occurs due to heat exchange between $\Omega_\pm$ and Γ. Writing $\xi_{\gamma\tau} = (\xi_{\tau+},\xi_{\tau-},\xi_{\tau\Gamma})$, the dissipation for imperfect heat transfer has the general form

$$\begin{aligned}\Psi^{0*}_{\mathrm{H}\Gamma}(\theta_\gamma;\xi_{\gamma\tau}) &:= \int_\Gamma \tfrac{1}{2}\,\xi_{\gamma\tau}\cdot\hat{\kappa}_\Gamma(\theta_+,\theta_-,\theta_\Gamma)\,\xi_{\gamma\tau}\,\mathrm{d}\mathcal{H}^{d-1}\,,\\ \text{where } \hat{\kappa}_\Gamma &:= \begin{pmatrix}\kappa_{\Gamma,++} & \kappa_{\Gamma,+-} & \kappa_{\Gamma,+\Gamma}\\ \kappa_{\Gamma,-+} & \kappa_{\Gamma,--} & \kappa_{\Gamma,-\Gamma}\\ \kappa_{\Gamma,\Gamma+} & \kappa_{\Gamma,\Gamma-} & \kappa_{\Gamma,\Gamma\Gamma}\end{pmatrix}\ \text{fulfills}\ \hat{\kappa}_\Gamma\begin{pmatrix}1\\1\\1\end{pmatrix} = 0\,,\end{aligned} \tag{63}$$

and $\hat{\kappa}_\Gamma$ is positive semidefinite. We additionally account for dissipation due to processes involving z on the interface by the dual dissipation potential

$$\mathcal{R}^*_z(q_\Gamma,\xi_z) := \int_\Gamma R^*_{\mathrm{D}}(q_\Gamma,\xi_z)\,\mathrm{d}\mathcal{H}^{d-1}\,, \tag{64}$$

where for fixed q_Γ the function $\xi_z \mapsto R^*_{\mathrm{D}}(q_\Gamma,\xi_z)$ is convex and lower semicontinuous, see Sect. 4.2 for more details. Writing $\eta(\cdot) := \eta(q_\Gamma,\cdot) = \partial_{\xi_z}R^*_{\mathrm{D}}(q_\Gamma,\cdot)$, we find the following form of $\mathbb{K}^0_\Gamma$ with entries $\kappa_{\Gamma,ij}$ from (63) for $i, j \in \{+,-,\Gamma\}$

$$\mathbb{K}^0_\Gamma := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \kappa_{\Gamma,++} & 0 & \kappa_{\Gamma,+-} & 0 & \kappa_{\Gamma,+\Gamma} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \kappa_{\Gamma,-+} & 0 & \kappa_{\Gamma,--} & 0 & \kappa_{\Gamma,-\Gamma} \\ 0 & 0 & 0 & 0 & \eta(\square) & 0 \\ 0 & \kappa_{\Gamma,\Gamma+} & 0 & \kappa_{\Gamma,\Gamma-} & 0 & \kappa_{\Gamma,\Gamma\Gamma} \end{pmatrix}, \tag{65}$$

where the first and the third lines collect the entries for $\xi_{w_\pm}$ with $w = (u, p)$. Again, we confirm that $\mathbb{K}^0_\Gamma(0, 1, 0, 1, 0, 1)^\top = 0$ as a prerequisite for NIC by (55).

Since the interface functionals in (60) account for a mutual interaction of the traces of w (i.e., of u), there is a conservative contribution to the evolution along Γ, and hence we set

$$\mathbb{J}^0_{\mathrm{HD}\Gamma} := \begin{pmatrix} 0 & 0\,1\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0 \\ -1 & 0\,0\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0 \\ 0 & 0\,0\,0\,0\,0 \end{pmatrix}. \tag{66}$$

Noninteraction Conditions (15) In order to ensure the NIC (15), we follow the approach of Sect. 3.4 to find the Poisson structures $\mathbb{J}_{\mathrm{B}}$ and $\mathbb{J}_\Gamma$ and the Onsager operators $\mathbb{K}_{\mathrm{B}}$ and $\mathbb{K}_\Gamma$. In particular, we introduce the bulk Poisson structure according to (50) and (46) in the specific form

$$\mathbb{J}_{\mathrm{B}}(q_\tau) = \mathbb{L}_{S_{\mathrm{B}}}\mathbb{J}^0_{\mathrm{B}}\mathbb{L}^*_{S_{\mathrm{B}}}, \quad \text{where} \quad \mathbb{L}^*_{S_{\mathrm{B}}} = \begin{pmatrix} I & 0 & -\frac{\square}{\partial_\tau S_{\mathrm{B}}} * \mathrm{D}_u S_{\mathrm{B}} \\ 0 & I & -\frac{\square}{\partial_\tau S_{\mathrm{B}}} * \mathrm{D}_p S_{\mathrm{B}} \\ 0 & 0 & \frac{1}{\partial_\tau S_{\mathrm{B}}} \end{pmatrix} \tag{67a}$$

$$\text{with } \delta_p S_{\mathrm{B}} = (0, \dots, 0), \ \text{and} \mathrm{D}_p S_{\mathrm{B}} = (0, \dots, 0)^\top \in \mathbb{R}^d$$
$$\text{and } -\frac{\square}{\partial_\tau S_{\mathrm{B}}} * \mathrm{D}_u S_{\mathrm{B}} = \sum_{i\in\{+,-\}} \big[\mathrm{div}\big(\tfrac{\square}{\partial_\tau S_{\mathrm{B}}}\partial_e S_{\mathrm{B}}\big)\big]_{\Omega_i} - \big[\gamma_i\big(\tfrac{\square}{\partial_\tau S_{\mathrm{B}}}\partial_e S_{\mathrm{B}}\big)\mathrm{n}_i\big]_{\Gamma_i}, \tag{67b}$$

so that this entry in $\mathbb{L}^*_{S_{\mathrm{B}}}$ also generates a nonzero trace contribution on $\Gamma_\pm$. It can be readily checked that $\mathbb{L}^*_{S_{\mathrm{B}}}\mathrm{D}S_{\mathrm{B}} = (0, 0, 1)^\top$, which ensures the NIC (15) $\mathbb{J}_{\mathrm{B}}(q_{\tau_{\mathrm{B}}})\mathrm{D}S_{\mathrm{B}} = \mathbb{L}_{S_{\mathrm{B}}}\mathbb{J}^0_{\mathrm{B}}\mathbb{L}^*_{S_{\mathrm{B}}}\mathrm{D}S_{\mathrm{B}} = 0$ by the form of $\mathbb{J}^0_{\mathrm{B}}$ from (62a).

For the bulk Onsager contribution, we also follow (48) and (46), i.e.,

$$\mathbb{K}_{\mathrm{B}}(q_\tau) = \mathbb{L}_{E_{\mathrm{B}}}\mathbb{K}^0_{\mathrm{B}}\mathbb{L}^*_{E_{\mathrm{B}}}, \text{ where } \quad \mathbb{L}^*_{E_{\mathrm{B}}} = \begin{pmatrix} I & 0 & -\frac{\square}{\partial_\tau E_{\mathrm{B}}} * \mathrm{D}_u E_{\mathrm{B}} \\ 0 & e(\square) & e \circ \Big(-\frac{\square}{\partial_\tau E_{\mathrm{B}}} * \mathrm{D}_p E_{\mathrm{B}}\Big) \\ 0 & 0 & \frac{1}{\partial_\tau E_{\mathrm{B}}} \end{pmatrix} \tag{67c}$$

$$\begin{aligned} &\text{with } \delta_p E_{\mathrm{B}} = \partial_p E_{\mathrm{B}}^\top = p/\varrho, \\ &\quad -\tfrac{\square}{\partial_\tau E_{\mathrm{B}}} * \mathrm{D}_u E_{\mathrm{B}} = \sum_{i\in\{+,-\}} \big[\mathrm{div}\big(\tfrac{\square}{\partial_\tau E_{\mathrm{B}}}\partial_e E_{\mathrm{B}}\big)\big]_{\Omega_i} - \big[\big(\tfrac{\square}{\partial_\tau E_{\mathrm{B}}}\partial_e E_{\mathrm{B}}\big)\mathrm{n}_i\big]_{\Gamma_i}, \\ &\text{and } e^*(\xi_p) = \sum_{i\in\{+,-\}} \big[-\mathrm{div}\,\xi_p\big]_{\Omega_i} + \big[\gamma_i\xi_p\cdot\mathrm{n}_i\big]_{\Gamma_i}. \end{aligned} \tag{67d}$$

Comparing with (46), this means that here $\mathbb{A}_H = e^*$. The operators e and e^* together with the viscous Onsager operator $\mathbb{K}^0_{\mathrm{V}}$ generate the dissipative contributions for Kelvin–Voigt viscoelasticity.

Similarly, the interfacial geometric structures $\mathbb{J}_\Gamma$ and $\mathbb{K}_\Gamma$ are deduced according to (49) by transforming $\mathbb{K}^0_\Gamma$ and $\mathbb{J}^0_\Gamma$ using the transformation maps $\mathbb{L}_{E_\Gamma}$ and $\mathbb{L}^*_{E_\Gamma}$ as well as $\mathbb{L}_{S_\Gamma}$ and $\mathbb{L}^*_{S_\Gamma}$ obtained by (52).

Elimination of Surface Thermodynamic Variable τ_Γ by Local Equilibrium We specify $\hat\kappa_\Gamma$ from (63) as

$$\hat\kappa_\Gamma(\theta_\gamma) = \kappa_\Gamma(\theta_\gamma)\begin{pmatrix} 1+\frac{1}{4\epsilon} & -1+\frac{1}{4\epsilon} & -\frac{1}{2\epsilon} \\ -1+\frac{1}{4\epsilon} & 1+\frac{1}{4\epsilon} & -\frac{1}{2\epsilon} \\ -\frac{1}{2\epsilon} & -\frac{1}{2\epsilon} & \frac{1}{\epsilon} \end{pmatrix},$$

$$\text{i.e.,} \quad \xi\cdot\hat\kappa_\Gamma\xi = \kappa_\Gamma(\xi_{\tau+} - \xi_{\tau\_})^2 + \epsilon^{-1}(\xi_{\tau\Gamma} - \tfrac12(\xi_{\tau+} + \xi_{\tau-}))^2$$

and consider $\mathbb{K}_\Gamma = \mathbb{L}_{S_\Gamma}\mathbb{K}^0_\Gamma\mathbb{L}^*_{S_\Gamma}$ with $\mathbb{K}^0_\Gamma$ from (65). Here, $\epsilon > 0$ plays the role of a relaxation parameter. Using Gibbs' relation $\frac{\partial_{\tau i} S_\Gamma}{\partial_{\tau i} E_\Gamma} = \theta_i^{-1}$ for $i \in \{+,-,\Gamma\}$, we calculate the entries of $\mathbb{K}_\Gamma \mathrm{D}\mathcal{S}_\Gamma$ as follows:

$$\begin{aligned} &(\mathbb{K}_\Gamma\mathrm{D}\mathcal{S}_\Gamma)_{w+} = (\mathbb{K}_\Gamma\mathrm{D}\mathcal{S}_\Gamma)_{w-} = 0, \\ &(\mathbb{K}_\Gamma\mathrm{D}\mathcal{S}_\Gamma)_{\tau\pm} = \kappa_\Gamma\tfrac{1}{\partial_{\tau_\pm}E_\Gamma}\Big((\theta_\pm^{-1} - \theta_\mp^{-1}) - \frac{1}{2\epsilon}\Big(\theta_\Gamma^{-1} - \tfrac12(\theta_+^{-1} + \theta_-^{-1})\Big)\Big) \\ &(\mathbb{K}_\Gamma\mathrm{D}\mathcal{S}_\Gamma)_z = \eta\Big(\partial_z S_\Gamma - \theta_\Gamma^{-1} * \partial_z E_\Gamma\Big) \\ &(\mathbb{K}_\Gamma\mathrm{D}\mathcal{S}_\Gamma)_{\tau_\Gamma} = -\tfrac{1}{\partial_{\tau_\Gamma}E_\Gamma}\eta\Big(\partial_z S_\Gamma - \theta_\Gamma^{-1} * \partial_z E_\Gamma\Big) + \tfrac{1}{\partial_{\tau_\Gamma}E_\Gamma}\frac{\kappa_\Gamma}{\epsilon}\Big(\theta_\Gamma^{-1} - \tfrac12(\theta_+^{-1} + \theta_-^{-1})\Big). \end{aligned}$$

Under the assumption that the system is quasi-stationary with respect to τ_Γ, we have $(\mathbb{K}_\Gamma \mathrm{D}\mathcal{S}_\Gamma)_{\tau_\Gamma} = 0$ and hence in the second row

$$(\mathbb{K}_\Gamma \mathrm{D}\mathcal{S}_\Gamma)_{\tau_\pm} = \kappa_\Gamma \tfrac{1}{\partial_{\tau_\pm} E_\Gamma} \left((\theta_\pm^{-1} - \theta_\mp^{-1}) \right) - \tfrac{1}{2\,\partial_{\tau_\pm} E_\Gamma} \eta \left(\partial_z S_\Gamma - \theta_\Gamma^{-1} * \partial_z E_\Gamma \right) .$$

The last expression cannot yet be rephrased in terms of a gradient flow. However, the limit $\epsilon \to 0$ formally enforces that $\frac{1}{\theta_\Gamma} \approx \frac{1}{2}(\frac{1}{\theta_+} + \frac{1}{\theta_-})$. Then, we find the new GENERIC system by eliminating the row for τ_Γ in $\mathbb{K}^0_\Gamma$ and $\mathbb{J}^0_\Gamma$ in (65) and (66) and modifying $\mathbb{L}_{E_\Gamma}$ and $\mathbb{L}^*_{E_\Gamma}$ in the following way:

$$\mathbb{L}_{E_\Gamma} = \begin{pmatrix} I_{w_+} & 0 & 0 & 0 & 0 \\ -\frac{1}{\partial_{\tau_+} E_\Gamma} \delta_{w_+} E_\Gamma & \frac{1}{\partial_{\tau_+} E_\Gamma} & 0 & 0 & -\frac{1}{2\,\partial_{\tau_+} E_\Gamma} \delta_z E_\Gamma \\ 0 & 0 & I_{w_-} & 0 & 0 \\ 0 & 0 & -\frac{1}{\partial_{\tau_-} E_\Gamma} \delta_{w_-} E_\Gamma & \frac{1}{\partial_{\tau_-} E_\Gamma} & -\frac{1}{2\,\partial_{\tau_-} E_\Gamma} \delta_z E_\Gamma \\ 0 & 0 & 0 & 0 & I_z \end{pmatrix} , \tag{68a}$$

$$\mathbb{L}^*_{E_\Gamma} = \begin{pmatrix} I_{w_+} & -\frac{\square}{\partial_{\tau_+} E_\Gamma} * \mathrm{D}_{w_+} E_\Gamma & 0 & 0 & 0 \\ 0 & \frac{1}{\partial_{\tau_+} E_\Gamma} & 0 & 0 & 0 \\ 0 & 0 & I_{w_-} & -\frac{\square}{\partial_{\tau_-} E_\Gamma} * \mathrm{D}_{w_-} E_\Gamma & 0 \\ 0 & 0 & 0 & \frac{1}{\partial_{\tau_-} E_\Gamma} & 0 \\ 0 & -\frac{\square}{2\,\partial_{\tau_+} E_\Gamma} * \mathrm{D}_z E_\Gamma & 0 & -\frac{\square}{2\,\partial_{\tau_-} E_\Gamma} * \mathrm{D}_z E_\Gamma & I_z \end{pmatrix} , \tag{68b}$$

and similarly for $\mathbb{L}_{S_\Gamma}$ and $\mathbb{L}^*_{S_\Gamma}$.

Weak Formulation of the GENERIC System With Bulk–Interface Coupling Altogether, the weak formulation of the GENERIC evolution system

$$\langle \tilde\xi, \dot q \rangle_{\boldsymbol{Q}} = \langle \tilde\xi, \mathbb{J}_{\mathrm{B}} \mathrm{D}\mathcal{E}_{\mathrm{B}} + \mathbb{K}_{\mathrm{B}} \mathrm{D}\mathcal{S}_{\mathrm{B}} \rangle_{\boldsymbol{Q}_{\mathrm{B}}} + \langle \tilde\xi_{\Gamma\gamma}, \mathbb{J}_\Gamma \mathrm{D}\mathcal{E}_\Gamma + \mathbb{K}_\Gamma \mathrm{D}\mathcal{S}_\Gamma \rangle_{\boldsymbol{Q}_\Gamma} \tag{69}$$

for all admissible test functions $\tilde\xi = (\tilde\xi_{\mathrm{B}}, \tilde\xi_z) = (\tilde\xi_u, \tilde\xi_p, \tilde\xi_\tau, \tilde\xi_z) \in \boldsymbol{Q}^*$ and the interfacial test functions $\tilde\xi_{\Gamma\gamma} = (\gamma_+ \tilde\xi_{\mathrm{B}}, \gamma_- \tilde\xi_{\mathrm{B}}, \tilde\xi_z) \in \boldsymbol{Q}^*_{\Gamma\gamma}$ can be written as follows when collecting terms that use the same test function:

$$\langle \tilde\xi_u, \dot u \rangle_{\boldsymbol{Q}_u} = \langle \tilde\xi_u, p/\varrho \rangle_{\boldsymbol{Q}_u} , \tag{70a}$$

$$\begin{aligned} \langle \tilde\xi_p, \dot p \rangle_{\boldsymbol{Q}_p} = {} & \langle \tilde\xi_p, \operatorname{div}(\partial_e W(e(u)) + \mathbb{D}e(\dot u) - \tfrac{\partial_\tau E_{\mathrm{B}}}{\partial_\tau S_{\mathrm{B}}} \partial_e S_{\mathrm{B}}) + f \rangle_{\Omega \setminus \Gamma} \\ & + \sum_{i \in \{+,-\}} - \langle \gamma_i \tilde\xi_p, \gamma_i \big(\partial_e W_{\mathrm{B}}(e(u)) + \mathbb{D}e(\dot u) - \tfrac{\partial_\tau E_{\mathrm{B}}}{\partial_\tau S_{\mathrm{B}}} \partial_e S_{\mathrm{B}} \big) \mathrm{n}_i \rangle_{\Gamma_i} \\ & + \langle \gamma_i \tilde\xi_p, \partial_{\gamma_i u} U^{\mathrm{el}}_\Gamma(z, [\![u]\!]) \rangle_\Gamma , \end{aligned} \tag{70b}$$

$$\langle\tilde{\xi}_z,\dot{z}\rangle_{Q_z}=\langle\tilde{\xi}_z,\eta_z\rangle_{Q_z}\text{ and }\eta_z\in\partial_{\xi_z}R^*_{\mathrm{D}}\big(q_{\Gamma\gamma};-\tfrac{1}{2}\big(\tfrac{1}{\theta_+}+\tfrac{1}{\theta_-}\big)*\mathrm{D}_zU^{\mathrm{el}}_\Gamma(z,[\![u]\!])\big)\,,\tag{70c}$$

$$\begin{aligned}\langle\tilde{\xi}_\tau,\dot{\tau}\rangle_{Q_\tau}=&\langle\tfrac{\tilde{\xi}_\tau}{\partial_\tau E_{\mathrm{B}}},-\operatorname{div}\Big(\kappa(\theta)\nabla\tfrac{\partial_\tau S_{\mathrm{B}}}{\partial_\tau E_{\mathrm{B}}}\Big)-\tfrac{\partial_\tau E_{\mathrm{B}}}{\partial_\tau S_{\mathrm{B}}}\partial_e S_{\mathrm{B}}:e(\dot{u})+e(\dot{u}):\mathbb{D}e(\dot{u})\rangle_{\Omega\backslash\Gamma}\\&+\sum_{i\in\{+,-\}}\langle\gamma_i\big(\tfrac{\tilde{\xi}_\tau}{\partial_\tau E_{\mathrm{B}}}\big),\gamma_i\big(\kappa(\theta)\nabla\tfrac{\partial_\tau S_{\mathrm{B}}}{\partial_\tau E_{\mathrm{B}}}\big)\mathrm{n}_i\rangle_{\Gamma_i}+\langle\gamma_i\tilde{\xi}_\tau,-\tfrac{1}{2\partial_{\gamma_i\tau}E_\Gamma}\delta_zU^{\mathrm{el}}_\Gamma\cdot\eta_z\rangle_\Gamma\\&+\langle\big(\tfrac{\gamma_+\tilde{\xi}_\tau}{\partial_{\gamma_+\tau}E_\Gamma}-\tfrac{\gamma_-\tilde{\xi}_\tau}{\partial_{\gamma_-\tau}E_\Gamma}\big),\kappa_\Gamma([\![u]\!],z)[\![\theta]\!]\rangle_\Gamma\,,\end{aligned}\tag{70d}$$

where $\eta_z=\dot{z}\in\partial_{\xi_z}R^*_{\mathrm{D}}\big(q_{\Gamma\gamma};-\frac{1}{2}\big(\frac{1}{\theta_+}+\frac{1}{\theta_-}\big)*\mathrm{D}_zU^{\mathrm{el}}_\Gamma\big)$ by (70c). Using for the test functions in (70d) the ansatz $\tilde{\xi}_\tau=\partial_\tau E_{\mathrm{B}}\hat{\xi}_\tau$ for any suitable $\hat{\xi}_\tau$ and exploiting the Gibbs' relation, Eq. (70d) can be rewritten as

$$\begin{aligned}\langle\partial_\tau E_{\mathrm{B}}\hat{\xi}_\tau,\dot{\tau}\rangle_{Q_\tau}=&\langle\hat{\xi}_\tau,-\operatorname{div}\big(\kappa(\theta)\nabla\tfrac{1}{\theta}\big)-\theta\partial_e S_{\mathrm{B}}:e(\dot{u})+e(\dot{u}):\mathbb{D}e(\dot{u})\rangle_{\Omega\backslash\Gamma}\\&+\sum_{i\in\{+,-\}}\langle\gamma_i\hat{\xi}_\tau,\gamma_i\big(\kappa(\theta)\nabla\tfrac{1}{\theta}\big)\mathrm{n}_i\rangle_{\Gamma_i}+\langle\gamma_i\hat{\xi}_\tau,-\tfrac{1}{2}\delta_zU^{\mathrm{el}}_\Gamma\cdot\dot{z}\rangle_\Gamma+\langle[\![\hat{\xi}_\tau]\!],\kappa_\Gamma[\![\theta]\!]\rangle_\Gamma\end{aligned}$$

for all suitable test functions $\hat{\xi}_\tau$. This entails the condition

$$\sum_{i\in\{+,-\}}\langle\gamma_i\hat{\xi}_\tau,\gamma_i\big(\kappa(\theta)\nabla\tfrac{1}{\theta}\big)\mathrm{n}_i\rangle_{\Gamma_i}+\langle\gamma_i\hat{\xi}_\tau,-\tfrac{1}{2}\delta_zU^{\mathrm{el}}_\Gamma\cdot\dot{z}\rangle_\Gamma+\langle[\![\hat{\xi}_\tau]\!],\kappa_\Gamma[\![\theta]\!]\rangle_\Gamma=0\,.\tag{71}$$

Derivation of Interfacial Coupling Conditions From this, we are now going to derive interfacial coupling conditions in strong form. For shorter notation, we here set $J:=\kappa(\theta)\nabla\frac{1}{\theta}$, and with the relation $\mathrm{n}_-=-\mathrm{n}_+$, we calculate

$$\begin{aligned}\sum_{i\in\{+,-\}}\langle\gamma_i\hat{\xi}_\tau,\gamma_i\big(\kappa(\theta)\nabla\tfrac{1}{\theta}\big)\mathrm{n}_i\rangle_{\Gamma_i}&=\langle\gamma_+\hat{\xi}_\tau,\gamma_+J\mathrm{n}_+\rangle_\Gamma-\langle\gamma_-\hat{\xi}_\tau,\gamma_-J\mathrm{n}_+\rangle_\Gamma\\&=\langle[\![\hat{\xi}_\tau]\!],\tfrac{1}{2}(\gamma_+J+\gamma_-J)\cdot\mathrm{n}_+\rangle_\Gamma+\langle\tfrac{1}{2}(\gamma_+\hat{\xi}_\tau+\gamma_-\hat{\xi}_\tau),[\![J]\!]\cdot\mathrm{n}_+\rangle_\Gamma\,.\end{aligned}$$

Comparison with the remaining terms in (71) results in the interfacial coupling conditions (73g) and (73h) below.

Similar arguments also allow us to deduce interfacial coupling conditions from the weak momentum balance (70b): noting that $\langle\gamma_\pm\tilde{\xi}_p,\partial_{\gamma_+u_+}U^{\mathrm{el}}_\Gamma\rangle_\Gamma=\langle-\gamma_\pm\tilde{\xi}_p,\partial_{\gamma_-u_-}U^{\mathrm{el}}_\Gamma\rangle_\Gamma$, we may rewrite the terms stemming from the interfacial energy in (70b) as follows:

$$\langle\gamma_+\tilde{\xi}_p,\partial_{\gamma_+u}U^{\mathrm{el}}_\Gamma\rangle_\Gamma+\langle\gamma_-\tilde{\xi}_p,\partial_{\gamma_-u}U^{\mathrm{el}}_\Gamma\rangle_\Gamma=\langle[\![\tilde{\xi}_p]\!],\partial_{[\![u]\!]}U^{\mathrm{el}}_\Gamma\rangle_\Gamma\,.$$

Furthermore, using the abbreviation $\sigma_\pm := \big[\partial_e W_{\mathrm{B}}(e(u)) + \mathbb{D}e(\dot u) - \frac{\partial_\tau E_{\mathrm{B}}}{\partial_\tau S_{\mathrm{B}}} \partial_e S_{\mathrm{B}}\big]_{\Omega_\pm}$, we now read from (70b) that the terms on Γ have to satisfy the condition

$$0 = \langle [\![\tilde\xi_p]\!], \gamma_+\sigma_+\mathrm{n}_+ + \partial_{[\![u]\!]} U_\Gamma^{\mathrm{el}} \rangle_\Gamma + \langle \gamma_-\tilde\xi_p, \gamma_+\sigma_+\mathrm{n}_+ + \gamma_-\sigma_-\mathrm{n}_- \rangle_\Gamma \tag{72}$$

for all test functions $\tilde\xi_p \in Q_p^*$ with traces $\gamma_i\tilde\xi_p,\ i \in \{+,-\}$ and jump $[\![\tilde\xi_p]\!]$ across Γ. This provides the interfacial coupling conditions (73e) and (73f) below.

Strong form of the GENERIC System with Bulk–Interface Coupling From above considerations, we conclude that (70) corresponds to the following strong formulation:

$$\dot u = p/\varrho \quad \text{in } \Omega_\pm\,, \tag{73a}$$

$$\dot p = \mathrm{div}(\partial_e U_{\mathrm{B}} + \mathbb{D}e(\dot u) - \theta\partial_e S_{\mathrm{B}}) + f \quad \text{in } \Omega_\pm\,, \tag{73b}$$

$$\dot\tau = \tfrac{1}{\partial_\tau E_{\mathrm{B}}}\Big(\mathrm{div}(\kappa(\theta)\nabla\theta) - \theta\partial_e S_{\mathrm{B}} : e(\dot u) + e(\dot u) : \mathbb{D}e(\dot u)\Big) \quad \text{in } \Omega_\pm\,, \tag{73c}$$

$$\dot z \in \partial_{\xi_z} R_{\mathrm{D}}^*\big(q_{\Gamma\gamma}; -\tfrac12(\tfrac{1}{\theta_+} + \tfrac{1}{\theta_-}) * \mathrm{D}_z U_\Gamma^{\mathrm{el}}(z, [\![u]\!])\big) \quad \text{on } \Gamma\,, \tag{73d}$$

complemented by the following interfacial coupling conditions along Γ:

$$\gamma_+\sigma_+\mathrm{n}_+ + \gamma_-\sigma_-\mathrm{n}_- = 0\,, \tag{73e}$$

$$\gamma_+\sigma_+\mathrm{n}_+ + \partial_{[\![u]\!]} U_\Gamma^{\mathrm{el}}(z, [\![u]\!]) = 0\,, \tag{73f}$$

$$[\![\kappa(\theta)\nabla\tfrac{1}{\theta}]\!]\cdot\mathrm{n}_+ - \delta_z U_\Gamma^{\mathrm{el}}\cdot\dot z = 0\,, \tag{73g}$$

$$\tfrac12\big(\gamma_+(\kappa(\theta)\nabla\tfrac{1}{\theta}) + \gamma_-(\kappa(\theta)\nabla\tfrac{1}{\theta})\big)\cdot\mathrm{n}_+ + \kappa_\Gamma([\![u]\!], z)[\![\theta]\!] = 0\,, \tag{73h}$$

and by homogeneous boundary conditions on $\partial\Omega$ and suitable initial conditions.

In the next sections, we introduce typical choices used in mathematical literature for the dissipation potential for delamination R_{D} and its conjugate R_{D}^* in (64) and for the interfacial mechanical energy U_Γ^{el} from (60d). For these choices, we discuss the resulting form of the interfacial coupling conditions (73g) and (73h) and thus reveal the GENERIC structure of the models previously studied in the literature with analytical methods.

4.1 Typical Choices for Interfacial Mechanical Energies for Delamination

Interfacial mechanical energies for delamination are typically of the type

$$U_\Gamma^{\mathrm{el}}(z, [\![u]\!]) = W_\Gamma(z, [\![u]\!]) + I_{[0,1]}(z) + I_K([\![u]\!])\,. \tag{74}$$

Here, $I_{[0,1]}$ is the indicator function of the set $[0,1]$ to feature the constraint $z \in [0,1]$, i.e., $I_{[0,1]}(z) = 0$ if $z \in [0,1]$ and $I_{[0,1]}(z) = \infty$ otherwise. Moreover, the indicator function I_K of the convex cone $K := \{v,\ [\![v]\!] \cdot \mathrm{n}_+ \geq 0\}$ ensures non-penetration of the material along the interface. Most importantly, the term $W_\Gamma(z, [\![u]\!])$ takes into account that displacement discontinuities along Γ are energetically more costly as long as the glue is effective, i.e., for $z(t,x) > 0$. If displacement jumps are only penalized but not excluded, one speaks of *adhesive contact* and a typical energy density takes the form

$$W_\Gamma^k(z, [\![u]\!]) := \frac{k}{2} z |[\![u]\!]|^2 \tag{75a}$$

with a constant $k > 0$. In the case of *brittle delamination*, displacement jumps are excluded in points where the glue active with the energy density

$$W_\Gamma^\infty(z, [\![u]\!]) := I_{\mathrm{C_b}}(z, [\![u]\!]) \quad \text{with } \mathrm{C_b} := \{(\tilde{z}, [\![\tilde{u}]\!]),\ \tilde{z}|[\![\tilde{u}]\!]| = 0 \text{ a.e. on } \Gamma\} \tag{75b}$$

the set accounting for the non-smooth, *brittle constraint*. In combination with a unidirectional, rate-independent dissipation potential, cf. (77) below with $r = 1$, this provides a model for fracture in the spirit of Griffith [10].

For analytical reasons, some works additionally consider in $U_\Gamma^{\mathrm{el}}(z, [\![u]\!])$ a gradient term for the delamination variable. For example, [3–5] consider the gradient term

$$G(\nabla z) := \tfrac{1}{2}|\nabla z|^2, \tag{76}$$

and [28] uses a Modica–Mortola type gradient term $G_{\mathrm{M}}(\nabla z) := \frac{1}{2\mathrm{M}}|\nabla z|^2 + \frac{\mathrm{M}}{2} z^2 (1-z)^2$ to approximate as $\mathrm{M} \to \infty$ a model which only accounts for the fully intact $z(t,x) = 1$ and the fully broken state $z(t,x) = 0$. In this limit, the interfacial gradient term is given by the relative perimeter of the set $Z(t) := \{x \in \Gamma,\ z(t,x) = 1\}$ in Γ.

4.2 Typical Choices of Dissipation Potentials for Delamination

Delamination in non-living materials is a unidirectional process, i.e., once the glue has weakened in an interfacial point, it cannot heal and will ultimately break. This property can be modeled by a dissipation potential of the form

$$\begin{aligned} &\mathcal{R}_{\mathrm{D}}(q; v) := \int_\Gamma R_{\mathrm{D}}(q; v)\, \mathrm{d}\mathcal{H}^{d-1} \\ &\text{with } R_{\mathrm{D}}(q; v) := a(q) R_r(v) + I_{(-\infty,0]}(v) \quad \text{and } R_r(v) := \frac{1}{r}|v|^r, \end{aligned} \tag{77}$$

for some strictly positive, state-dependent function $a(q) > 0$ for all q, the integrability exponent $r \in [1, \infty)$, and $I_{(-\infty,0]}$ the indicator function of the set $(-\infty, 0]$, i.e., $I_{(-\infty,0]}(v) = 0$ if $v \in (-\infty, 0]$ and $I_{(-\infty,0]}(v) = \infty$ otherwise. Rate-independent delamination corresponds to the case $r = 1$, see, e.g., [26–29], while $r > 1$ describes rate-dependent delamination and is most commonly treated in the literature with the exponent $r = 2$, cf., e.g., [4]. When choosing $v = \dot z$, we note that the indicator function $I_{(-\infty,0]}$ in (77) entails the constraint $\dot z \le 0$, which ensures that delamination cannot heal, as $z = 1$ is the intact state and $z = 0$ denotes the broken state. The dual dissipation potential $\mathcal{R}^*_{\mathrm{D}}$, respectively, its density R^*_{D}, is the convex conjugate of $\mathcal{R}_{\mathrm{D}}$ obtained by (12). In the case $r = 1$, this is

$$R^*_{\mathrm{D}}(q; \xi_z) = \begin{cases} 0 & \text{if } \xi_z \in [-a(q), \infty), \\ \infty & \text{otherwise}, \end{cases} \tag{78a}$$

i.e., $R^*_{\mathrm{D}}(q; \cdot) = I_{\partial_v R_{\mathrm{D}}(q;0)}$ is given by the indicator function of the convex set $\partial_v R_{\mathrm{D}}(q; 0)$. For $r \in (1, \infty)$, the convex conjugate is given by

$$R^*_{\mathrm{D}}(q; \xi_z) = \begin{cases} 0 & \text{if } \xi_z > 0, \\ a(q) R_{r'}\big(\frac{\xi}{a(q)}\big) & \text{otherwise}, \end{cases} \quad \text{where } \tfrac{1}{r} + \tfrac{1}{r'} = 1\,. \tag{78b}$$

We now discuss (formally) equivalent formulations for the flow rule (73d) and their implication on the coupling conditions (73g)– (73h). For this, let us first assume that $U^{\mathrm{el}}_\Gamma(z, [\![u]\!])$ is smooth and does not feature the gradient ∇z. Then, $\frac12\big(\frac{1}{\theta_+} + \frac{1}{\theta_-}\big) *$ $\mathrm{D}_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!]) = \frac12\big(\frac{1}{\theta_+} + \frac{1}{\theta_-}\big)\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!])$ in (73d). By convex duality, the flow rule (73d) is equivalent to the force balance

$$-\tfrac12\big(\tfrac{1}{\theta_+} + \tfrac{1}{\theta_-}\big)\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!]) \in \partial_{\dot z} R_{\mathrm{D}}(q_{\Gamma\gamma}; \dot z) \quad \text{on } \Gamma\,, \tag{79}$$

and to the Fenchel equality

$$R_{\mathrm{D}}(q_{\Gamma\gamma}; \dot z) + R^*_{\mathrm{D}}\big(q_{\Gamma\gamma}; -\tfrac12\big(\tfrac{1}{\theta_+}+\tfrac{1}{\theta_-}\big)\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!])\big) = \langle -\tfrac12\big(\tfrac{1}{\theta_+}+\tfrac{1}{\theta_-}\big)\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!]), \dot z\rangle_\Gamma\,. \tag{80}$$

Comparing (79) with (77), we make the choice

$$a(q) := \tfrac12\big(\tfrac{1}{\theta_+} + \tfrac{1}{\theta_-}\big), \tag{81}$$

so that (79) results in the state-independent force balance

$$-\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!]) \in \partial_{\dot z}\big(R_r(\dot z) + I_{(-\infty,0]}(\dot z)\big)\,. \tag{82}$$

This type of force balance featuring a dissipation potential independent of θ is, e.g., considered in [26, 28] for $r = 1$ and in [3–5] for $r = 2$. Of course, the choice (81) also specifies $a(q) = \frac12\big(\frac{1}{\theta_+} + \frac{1}{\theta_-}\big)$ in (78b). With this ansatz, by formally dividing (80) by $a(q)$, coupling condition (73g) can be further rewritten as

$$[\![\kappa(\theta)\nabla\tfrac{1}{\theta}]\!] + R_r(\dot z) + \tfrac{1}{a(q)}R^*_{\mathrm{D}}\big(q; -\tfrac{1}{2}(\tfrac{1}{\theta_+} + \tfrac{1}{\theta_-})\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!])\big) = 0\,. \tag{83}$$

In the rate-independent case $r = 1$, when taking into account that $R^*_{\mathrm{D}}(q; -\frac{1}{2}(\frac{1}{\theta_+} + \frac{1}{\theta_-})\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!])) = 0$ by (78a) and (73d), we find for (83) in particular

$$[\![\kappa(\theta)\nabla\tfrac{1}{\theta}]\!] + R_1(\dot z) + I_{(-\infty,0]}(\dot z) = 0\,. \tag{84a}$$

Moreover, in the rate-dependent case $r > 1$, it is $R^*_{\mathrm{D}}\big(q; -\frac{1}{2}(\frac{1}{\theta_+} + \frac{1}{\theta_-})\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!])\big)$ $= a(q)R_{r'}\big(-\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!])\big)$ by (78b) given that $\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!]) \geq 0$, and hence (83) provides

$$[\![\kappa(\theta)\nabla\tfrac{1}{\theta}]\!] + R_r(\dot z) + I_{(-\infty,0]}(\dot z) + R_{r'}\big(-\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!])\big) = 0\,. \tag{84b}$$

In (74)–(75), it was discussed that mechanical energies for adhesive contact and brittle delamination for modeling reasons in general feature non-smooth but convex terms. Thus, $\partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!])$ in (79)–(84) indeed is the subdifferential of a convex function.

Now, we turn to the case that U^{el}_Γ from (74) additionally also contains a quadratic gradient term G as in (76). Then,

$$a(q) * \mathrm{D}_z\Big(U^{\mathrm{el}}_\Gamma(z, [\![u]\!]) + G(\nabla z)\Big) = a(q)\zeta_z + \operatorname{div} a(q)\nabla z \quad \text{with } \zeta_z \in \partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!]) \tag{85}$$

in (73d). Thus, the ansatz (77) and repeating above calculations do not help to remove $a(q)$ from the divergence term. In order to find for the delamination variable a force balance that is independent of $a(q)$, one rather has to modify the ansatz used to ensure the NIC for z in $\mathbb{L}_{E_\Gamma}$ and $\mathbb{L}^*_{E_\Gamma}$, see (68). More precisely, in the fifth line of $\mathbb{L}^*_{E_\Gamma}$, we replace

$$-\tfrac{\square}{2\partial_{\tau_\pm}E_\Gamma} * \mathrm{D}_z E_\Gamma \quad \text{by} \quad -\tfrac{\square}{2\partial_{\tau_\pm}E_\Gamma}\mathrm{D}_z E_\Gamma\,, \tag{86}$$

which is

$$-\tfrac{\square}{2\partial_{\tau_\pm}E_\Gamma}\mathrm{D}_z E_\Gamma = -\tfrac{\square}{2\partial_{\tau_\pm}E_\Gamma}\big(\zeta_z - \Delta_\Gamma z\big) \quad \text{with } \zeta_z \in \partial_z U^{\mathrm{el}}_\Gamma(z, [\![u]\!])\big)$$

and where Δ_Γ denotes the Laplace–Beltrami operator on Γ. This choice gives (79) with $\partial_z U^{\mathrm{el}}_\Gamma$ replaced by $\mathrm{D}_z\big(U^{\mathrm{el}}_\Gamma + G\big)$ and thus also results in a force balance alike (82)

$$0 \in \mathrm{D}_z\big(U^{\mathrm{el}}_\Gamma(z, [\![u]\!]) + G(\nabla z)\big) + \partial_{\dot z}\big(R_r(\dot z) + I_{(-\infty,0]}(\dot z)\big). \tag{87}$$

Moreover, we find for the fifth column in $\mathbb{L}_{E_\Gamma}$ that

$$-\tfrac{1}{2\partial_{\tau_\pm}E_\Gamma}\delta_z E_\Gamma \quad \text{is replaced by} \quad -\delta_z E_\Gamma\big[\tfrac{\square}{2\partial_{\tau_\pm}E_\Gamma}\big]\,, \tag{88}$$

where we have assumed homogeneous Neumann boundary conditions to hold along $\partial\Gamma$, here $\nabla z \cdot \mathrm{n}_\Gamma = 0$ on $\partial\Gamma$. This gives

$$-\,\delta_z E_\Gamma\big[\tfrac{\dot z}{2\partial_{\tau_\pm} E_\Gamma}\big] = -\Big(\zeta_z \cdot \tfrac{\dot z}{2\partial_{\tau_\pm} E_\Gamma} + \nabla_\Gamma z \cdot \nabla_\Gamma\big(\tfrac{\dot z}{2\partial_{\tau_\pm} E_\Gamma}\big)\Big) \tag{89}$$

to appear in (70d). To further process this expression, we now assume that $\nabla z \cdot \nabla\big(\frac{\dot z}{2\partial_{\tau_\pm} E_\Gamma}\big)$ is formally equivalent to $-\Delta_\Gamma z\big(\frac{\dot z}{2\partial_{\tau_\pm} E_\Gamma}\big)$ thanks to the homogeneous Neumann boundary conditions. Hence, $-\delta_z E_\Gamma\big[\frac{\dot z}{2\partial_{\tau_\pm} E_\Gamma}\big]$ is formally replaced by $-\frac{1}{2\partial_{\tau_\pm} E_\Gamma}\mathrm{D}_z E_\Gamma \cdot \dot z$ in (70d). By repeating the arguments subsequent to (70d), we arrive at (73g) with $\delta_z E_\Gamma \cdot \dot z$ replaced by $\mathrm{D}_z E_\Gamma \cdot \dot z$. In this way, we can again arrive at the interfacial coupling conditions (84) by exploiting the Fenchel equality, which now reads

$$\begin{aligned} &R_{\mathrm{D}}(q_{\Gamma\gamma};\dot z) + R^*_{\mathrm{D}}\Big(q_{\Gamma\gamma}; -\tfrac12\big(\tfrac{1}{\theta_+}+\tfrac{1}{\theta_-}\big)\mathrm{D}_z\big(U^{\mathrm{el}}_\Gamma(z,[\![u]\!]) + G(\nabla z)\big)\Big)\\ &= \langle -\tfrac12\big(\tfrac{1}{\theta_+}+\tfrac{1}{\theta_-}\big)\mathrm{D}_z\big(U^{\mathrm{el}}_\Gamma(z,[\![u]\!]) + G(\nabla z)\big), \dot z\rangle_\Gamma\,. \end{aligned}$$

Remark 4 We have exemplarily shown for delamination processes that the GENERIC structure of bulk–interface systems can be given in a weak sense based on thermodynamic functionals and geometric operators with bulk and interfacial contributions. The interfacial coupling conditions arise naturally from this weak form of GENERIC. The delamination models studied for their well-posedness, e.g., in the works [26, 28] with $r = 1$ and in [4] with $r = 2$, are obtained from thermodynamic functionals as discussed in Sects. 4.1 and 4.2. Hence, the above derivation confirms the GENERIC structure of these models. Yet, it has to be stressed that coupling conditions (73e)–(73h) and the reformulations made in Sect. 4.2 to arrive at (84) hold true on a formal level, only, since they require additional regularity of the terms involved, for example, for the term $\delta_z U^{\mathrm{el}}_\Gamma, \dot z$ appearing, e.g., in (70d). However, for $r = 1$, this cannot be guaranteed, since $\dot z$ is a Radon measure, only, and to have good duality would thus require $\delta_z U^{\mathrm{el}}_\Gamma$ to be continuous, which clearly is not to be expected. To circumvent this problem, [26, 28] derive a weak formulation directly based on (84).

Finally, we remark that also sensitivity with respect to the fracture mode can be added to the model by decomposing the displacement jump $[\![u]\!] = [\![u]\!]\cdot\mathrm{n}_+ + [\![u]\!]\cdot\mathrm{t}$ into its normal and tangential components and by considering the $a(q)$ in (77) not only to depend on $\theta_\pm$ but also on a function $\alpha([\![u]\!]\cdot\mathrm{n}_+, [\![u]\!]\cdot\mathrm{t})$. In [27], the analysis of such a mode-sensitive adhesive contact model with thermal effects requires the use of higher order gradients of $\dot u$. ⋆

Acknowledgments M.T. acknowledges the partial funding by the DFG through project C09 *Dynamics of rock dehydration on multiple scales* (project number 235221301) within CRC 1114 *Scaling cascades in complex systems* and project *Nonlinear fracture dynamics: Modeling, Analysis, Approximation, and Applications* (project number 441212523) within SPP 2256. M.H. acknowledges the funding by the DFG through SPP 2256 project HE 8716/1-1 *Fractal and stochastic homogenization using variational techniques*.

References

1. V. Arnold, V. Kozlov, A. Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics* (Springer, Berlin, 2006)
2. G. Barenblatt, The mathematical theory of equilibrium of cracks in brittle fracture. Adv. Appl. Mech. **7**, 55–129 (1962)
3. E. Bonetti, G. Bonfanti, R. Rossi, Thermal effects in adhesive contact: modelling and analysis. Nonlinearity **22**(11), 2697–2731 (2009)
4. E. Bonetti, G. Bonfanti, R. Rossi, Analysis of a model coupling volume and surface processes in thermoviscoelasticity. Nonlinear Anal. Real World Appl. **35**(6), 2349–2403 (2015)
5. E. Bonetti, G. Bonfanti, R. Rossi, Modeling via the internal energy balance and analysis of adhesive contact with friction in thermoviscoeleasticity. Nonlinear Anal. Real World Appl. **22**, 473–507 (2015)
6. E.C. D'Avignon, Physical consequences of the Jacobi identity (2015). https://arxiv.org/abs/1510.06455
7. M. Frémond, *Non-Smooth Thermomechanics* (Springer, Berlin, 2002)
8. K. Glavatskiy, D. Bedeaux, Non-equilibrium thermodynamics for surfaces; square gradient theory. Eur. Phys. J. Special Top. **222**, 161–175 (2013)
9. A. Glitzky, A. Mielke, A gradient structure for systems coupling reaction-diffusion effects in bulk and interfaces. ZAMP Z. Angew. Math. Phys. **64**, 29–52 (2013)
10. A. Griffith, The phenomena of rupture and flow in solids. Philos. Trans. R. Soc. Lond. A **221**, 163–198 (1921)
11. M. Grmela, H. Öttinger, Dynamics and thermodynamics of complex fluids. I. Development of a general formalism. Phys. Rev. E **56**(6), 6620–6632 (1997)
12. B. Halphen, Q. Nguyen, Sur les matériaux standards généralisés. J. Mécanique **14**, 39–63 (1975)
13. M. Hütter, B. Svendsen, Thermodynamic model formulation for viscoplastic solids as general equations for non-equilibrium reversible-irreversible coupling. Continuum Mech. Thermodyn. **24**, 211–227 (2012)
14. J. Maas, A. Mielke, Modeling of chemical reaction systems with detailed balance using gradient structures. J. Stat. Phys. **181**, 2257–2303 (2020)
15. A. Mielke, Formulation of thermoelastic dissipative material behavior using GENERIC. Continuum Mech. Thermodyn. **23**(3), 233–256 (2011)
16. A. Mielke, A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems. Nonlinearity **24**(4), 1329–1346 (2011). https://doi.org/10.1088/0951-7715/24/4/016
17. A. Mielke, Free energy, free entropy, and a gradient structure for thermoplasticity, in *Innovative Numerical Approaches for Multi-Field and Multi-Scale Problems*, ed. by K. Weinberg, A. Pandolfi. In Honor of Michael Ortiz's 60th Birthday. Lecture Notes in Applied and Computational Mechanics, vol. 81 (Springer, Cham, 2016), pp. 135–160
18. A. Mielke, D. Peschka, N. Rotundo, M. Thomas, On some extension of energy-drift-diffusion models: gradient structure for optoelectronic models of semiconductors, in *Progress in Industrial Mathematics at ECMI 2016*, ed. by P. Quintela et al. (ed.) Mathematics in Industry, vol. 26 (Springer, Berlin, 2017), pp. 291–298
19. M. Mittnenzweig, A. Mielke, An entropic gradient structure for Lindblad equations and GENERIC for quantum systems coupled to macroscopic models. J. Stat. Phys. **167**, 205–233 (2017)
20. P.J. Morrison, Hamiltonian description of the ideal fluid. Rev. Mod. Phys. **70**(2), 467–521 (1998). https://link.aps.org/doi/10.1103/RevModPhys.70.467
21. A. Moses Badlyan, C. Zimmer, Operator-GENERIC formulation of thermodynamics of irreversible processes (2018). https://arxiv.org/abs/1807.09822
22. H. Öttinger, Nonequilibrium thermodynamics for open systems. Phys. Rev. E **73**, 036126 (2006)

23. H. Öttinger, M. Grmela, Dynamics and thermodynamics of complex fluids. II. Illustrations of a general formalism. Phys. Rev. E **56**(6), 6633–6655 (1997)
24. M. Pavelka, V. Klika, M. Grmela, *Multiscale Thermo-Dynamics* (De Gruyter, Berlin, 2018). https://doi.org/10.1515/9783110350951
25. D. Peschka, M. Thomas, T. Ahnert, A. Münch, B. Wagner, *Gradient Structures for Flows of Concentrated Suspensions*. CIM Series in Mathematical Sciences (Springer, Berlin, 2019), pp. 295–318
26. R. Rossi, T. Roubíček, Thermodynamics and analysis of rate-independent adhesive contact at small strains. Nonlinear Anal. **74**(10), 3159–3190 (2011)
27. R. Rossi, T. Roubíček, Adhesive contact delaminating at mixed mode, its thermodynamics and analysis. Interfaces Free Bound. **15**(1), 1–37 (2013)
28. R. Rossi, M. Thomas, From an adhesive to a brittle delamination model in thermo-visco-elasticity. ESAIM Control Optim. Calc. Var. **21**, 1–59 (2015)
29. R. Rossi, M. Thomas, From adhesive to brittle delamination in visco-elastodynamics. Math. Models Methods Appl. Sci. **27**, 1489–1546 (2017)
30. R. Rurali, L. Colombo, X. Cartoixà, Ø. Wilhelmsen, T. Trinh, D. Bedeaux, S. Kjelstrup, Heat transport through a solid–solid junction: the interface as an autonomous thermodynamic system. Phys. Chem. Chem. Phys. **18**, 13741 (2016)
31. M. Thomas, A comparison of delamination models: modeling, properties, and applications, in *Mathematical Analysis of Continuum Mechanics and Industrial Applications II, Proceedings of the International Conference CoMFoS16*, ed. by P. van Meurs, M. Kimura, H. Notsu. Mathematics for Industry, vol. 30 (Springer, Singapore, 2017), pp. 27–38
32. M. Thomas, C. Zanini, Cohesive zone-type delamination in visco-elasticity. Discrete Contin. Dyn. Syst. Ser. S **10**, 1487–1517 (2017)
33. P. Vágner, M. Pavelka, O. Esen, Multiscale thermodynamics of charged mixtures. Contin. Mech. Thermodyn. **33**, 237–268 (2021)
34. A. Zafferi, D. Peschka, M. Thomas, GENERIC framework for reactive fluid flows. ZAMM Z. Angew. Math. Mech., published online 9.5.2022. https://doi.org/10.1002/zamm.202100254

Part II
Review Papers

Phase Separation in Heterogeneous Media

Riccardo Cristoferi, Irene Fonseca, and Raghavendra Venkatraman

1 Introduction

The study of pattern formation in equilibrium configurations phase separation is an extremely complex phenomenon that has attracted the interest of many mathematicians. In the case of homogeneous substances, variational models such as the Modica–Mortola functional (see [28, 31]) and its vectorial (see [4, 24]), anisotropic (see [5, 23]), and non-isothermal variants (see [11]) have been proven capable of describing the stable configurations observed in experiments. For composite materials, it has been realized experimentally (see [6]) that the microscopic scale heterogeneities can affect the macroscopic equilibrium configurations as well as the dynamics of interfaces. Therefore, physics requires the mathematical models to include these microscopic effects.

In this chapter, we consider a variational approach to the study of phase transitions in heterogeneous media in the case where the scale of the heterogeneities is the same as those at which the phase transitions phenomenon takes place. In particular, we study a Modica–Mortola-like phase field model where the heterogeneities are modeled by oscillations in the potential. To be precise, let $d, N \geq 1$, fix an open bounded set $\Omega \subset \mathbb{R}^N$ with Lipschitz boundary, and, for $\varepsilon > 0$, define the energy

R. Cristoferi
Radboud University, IMAPP - Mathematics, Nijmegen, The Netherlands
e-mail: riccardo.cristoferi@ru.nl

I. Fonseca
Carnegie Mellon University, Department of Mathematical Sciences, Pittsburgh, PA, USA
e-mail: fonseca@andrew.cmu.edu

R. Venkatraman (✉)
Courant Institute, New York University, New York, NY, USA
e-mail: raghav@cims.nyu.edu

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_16

$\mathcal{F}_\varepsilon : H^1(\Omega;\mathbb{R}^d) \to [0,\infty]$ as

$$\mathcal{F}_\varepsilon(u) := \int_\Omega \left[\frac{1}{\varepsilon} W\left(\frac{x}{\varepsilon}, u(x)\right) + \varepsilon |\nabla u(x)|^2 \right] \mathrm{d}x \,. \tag{1}$$

Here $u \in H^1(\Omega;\mathbb{R}^d)$ represents the phase field variable. The assumptions that the double well potential $W : \mathbb{R}^N \times \mathbb{R}^d \to [0,\infty)$ has to satisfy differ according to the questions addressed, and therefore, we will present them in each section.

We are interested in understanding what is the sharp interface limit as the parameter $\varepsilon \to 0$. Local minimizers of this limit under a mass constraint will describe equilibrium configurations.

Previous investigations on models related to the one considered in this chapter have been undertaken by several authors. In particular, in [2] (see also [1]), Ansini, Braides and Chiadò Piat considered the case where oscillations are in the forcing term $f(\nabla u)$ (which generalizes $|\nabla u|^2$), while in [17] and [18] by Dirr, Lucia and Novaga investigated the interaction of the fluid with a periodic mean zero external field. Moreover, in [7], Braides and Zeppieri studied the Γ expansion of the scalar one-dimensional case, allowing the zeros of the potential to jump in a specific way. Finally, the case of higher-order derivatives is examined in [25] by Francfort and Müller.

2 Phase Field Model

In this section, we present the results obtained in [9, 10, 12, 13].

2.1 Sharp Interface Limit

In order to study the sharp interface limit of the energy (1), we assume that the double well potential $W : \mathbb{R}^N \times \mathbb{R}^d \to [0,\infty)$ satisfies the following properties:

(A1) For all $p \in \mathbb{R}^d$, $x \mapsto W(x,p)$ is Q-periodic, where $Q := (-1/2, 1/2)^N$.

(A2) W is a Carathéodory function, i.e.:

  (i) For all $p \in \mathbb{R}^d$, the function $x \mapsto W(x,p)$ is measurable.
  (ii) For a.e. $x \in Q$, the function $p \mapsto W(x,p)$ is continuous.

(A3) There exist $z_1, z_2 \in \mathbb{R}^d$ such that, for a.e. $x \in Q$, $W(x,p) = 0$ if and only if $p \in \{z_1, z_2\}$.

(A4) There exists a continuous function $\widetilde{W} : \mathbb{R}^d \to [0,\infty)$, vanishing only at $p = z_1$ and at $p = z_2$, such that $\widetilde{W}(p) \leq W(x,p)$ for a.e. $x \in Q$.

(A5) There exist $C > 0$ and $q \geq 2$ such that

$$\frac{1}{C}|p|^q - C \leq W(x,p) \leq C(1+|p|^q)$$

for a.e. $x \in Q$ and all $p \in \mathbb{R}^d$.

Remark 1 The assumption (A2)(i) above is the strongest we can ask when modeling periodic inclusions of different materials. Indeed, when each cell Q is composed of k different inclusions of materials each in a region $E_1, \dots, E_k \subset Q$, the potential W takes the form

$$W(x,p) := \sum_{i=1}^{k} W_i(p)\chi_{E_i}(x)\,,$$

where $W_i : \mathbb{R}^d \to [0,\infty)$ are continuous functions with quadratic growth at infinity and such that $W_i(p) = 0$ if and only if $p \in \{z_1, z_2\}$. Therefore, the function W in the first variable is, in general, only measurable. Moreover, the continuity of W in the second variable, as well as the non-degeneracy of the potential (A4) and the growth at infinity in the second variable (A5), is compatible with what is usually assumed in the physical literature.

The limiting functional will be an interfacial energy whose energy density is defined via a cell formula as follows.

Definition 1 For $\nu \in \mathbb{S}^{N-1}$, let $u_{0,\nu} : \mathbb{R}^N \to \mathbb{R}^d$ be the function

$$u_{0,\nu}(x) := \begin{cases} z_1 & \text{if } x\cdot\nu \leq 0\,, \\ z_2 & \text{if } x\cdot\nu > 0\,, \end{cases}$$

and denote by $\mathcal{Q}_\nu$ the family of cubes centered at the origin with unit length sides and having two faces orthogonal to ν. For $T > 0$, $Q_\nu \in \mathcal{Q}_\nu$, and $\rho \in C_c^\infty(B(0,1))$ with $\int_{\mathbb{R}^N} \rho(x)\mathrm{d}x = 1$, where $B(0,1)$ is the unit ball in $\mathbb{R}^N$, consider the class of functions

$$\mathcal{C}(\rho, Q_\nu, T) := \Big\{ u \in H^1(TQ_\nu;\mathbb{R}^d) \;:\; u = u_{0,\nu} * \rho \text{ on } \partial(TQ_\nu) \Big\}\,.$$

We define the function $\sigma : \mathbb{S}^{N-1} \to [0,\infty)$ as

$$\sigma(\nu) := \lim_{T\to\infty} g(\nu, T)\,,$$

where, for each $\nu \in \mathbb{S}^{N-1}$ and $T > 0$,

$$g(\nu,T) := \frac{1}{T^{N-1}} \inf\left\{ \int_{TQ_\nu} \Big[W(y,u(y)) + |\nabla u|^2 \Big] \mathrm{d}y \;:\; Q_\nu \in \mathcal{Q}_\nu,\; u \in \mathcal{C}(\rho, Q_\nu, T) \right\}.$$

Remark 2 It was observed by Müller in [29] that, in the case the potential W is vectorial, in the definition of the cell formula it is not enough to take the minimum only on a single cell, but to consider the sequence of minima taken on larger and larger cells TQ_ν. In case the potential W is scalar, it is possible to reduce to a single-cell problem with W replaced by W^{**} (see Lemma 4.1 and the remark after that, in [29]).

The main properties of the function $\sigma : \mathbb{S}^{N-1} \to [0,\infty)$ that are relevant for our study are collected in the following result. For the proof, see [12, Lemma 4.1, Remark 4.2, Lemma 4.3, Proposition 4.4].

Lemma 1 *The following hold:*

(i) For every $\nu \in \mathbb{S}^{N-1}$, the quantity $\sigma(\nu)$ is well-defined and finite.
(ii) The value of $\sigma(\nu)$ does not depend on the choice of the mollifier ρ.
(iii) The map $\nu \mapsto \sigma(\nu)$ is upper semi-continuous on $\mathbb{S}^{N-1}$.
(iv) The infimum in the definition of $g(\nu,T)$ may be taken with respect to one fixed cube $Q_\nu \in \mathcal{Q}_\nu$. Namely, given $\nu \in \mathbb{S}^{N-1}$, for any $Q_\nu \in \mathcal{Q}_\nu$, it holds

$$\sigma(\nu) = \lim_{T\to\infty} \frac{1}{T^{N-1}} \inf\left\{ \int_{TQ_\nu} \left[W(y,u(y)) + |\nabla u|^2 \right] \mathrm{d}y \, : \, u \in C(\rho, Q_\nu, T) \right\}.$$

We are now in position to introduce the limiting functional.

Definition 2 Define the functional $\mathcal{F}_0 : L^1(\Omega;\mathbb{R}^d) \to [0,\infty]$ as

$$\mathcal{F}_0(u) := \begin{cases} \displaystyle\int_{\partial^* A} \sigma(\nu_A(x))\, \mathrm{d}\mathcal{H}^{N-1}(x) & \text{if } u \in BV(\Omega;\{z_1,z_2\}), \\ +\infty & \text{else,} \end{cases} \tag{2}$$

where $A := \{u = z_1\}$ and $\nu_A(x)$ denotes the measure theoretic external unit normal to the reduced boundary $\partial^* A$ of A at the point x.

Remark 3 Note that by Lemma 1(i), it holds $\mathcal{F}_0(u) < \infty$ for all $u \in BV(\Omega;\{z_1,z_2\})$, and by Lemma 1(ii), the definition does not depend on the choice of the mollifier ρ.

Theorem 1 *Let $\{\varepsilon_n\}_{n\in\mathbb{N}} \subset (0,1)$ be a sequence such that $\varepsilon_n \to 0^+$ as $n \to \infty$. Assume that (A1), (A2), (A3), (A4), and (A5) hold:*

(i) If $\{u_n\}_{n\in\mathbb{N}} \subset H^1(\Omega;\mathbb{R}^d)$ is such that

$$\sup_{n\in\mathbb{N}} \mathcal{F}_{\varepsilon_n}(u_n) < +\infty,$$

then up to a subsequence (not relabeled), $u_n \to u$ in $L^1(\Omega;\mathbb{R}^d)$, for some function $u \in BV(\Omega;\{z_1,z_2\})$.

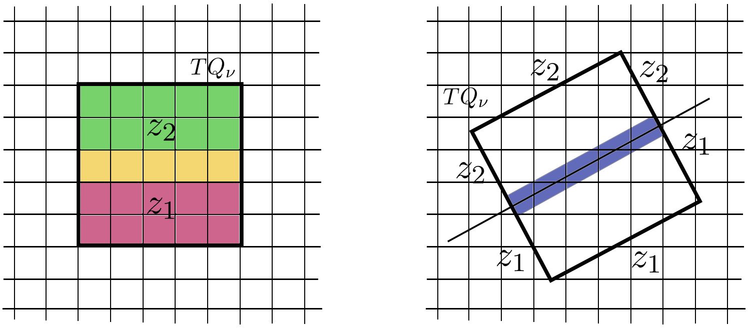


Fig. 1 The source of anisotropy for the limiting functional. If $\nu_A(x)$ is oriented with a direction of periodicity of W, the (local) recovery sequence would simply be obtained by using a rescaled version of the recovery sequence for $\sigma(\nu_A(x))$ in each yellow cube and by setting z_1 in the green region, and z_2 in the pink one. If, instead, $\nu_A(x)$ is not oriented with a direction of periodicity of W, the above procedure does not guarantee that we recover the desired energy since the energy of such functions is *not* the sum of the energy of each cube

(ii) The functional $\mathcal{F}_0$ is the Γ-limit in the L^1 topology of the family of functionals $\{\mathcal{F}_{\varepsilon_n}\}_{n\in\mathbb{N}}$.

Remark 4 The most interesting aspect of the above result is the anisotropic character of the limiting functional. This might come as a surprise since the initial functional $\mathcal{F}_\varepsilon$ is isotropic, but there is a hidden anisotropy: the possible mismatch between the directions of periodicity of W and the local orientation of the limiting interface $\partial^* A$ (see Fig. 1).

We would like to comment on the main ideas behind the proof of Theorem 1. Compactness follows by using classical arguments (see [24]) since the non-degeneracy assumption (A4) allows to reduce to the case of a non-oscillating potential

$$\mathcal{F}_{\varepsilon_n}(u_n) \geq \int_\Omega \left[\frac{1}{\varepsilon_n} \widetilde{W}(u_n(x)) + \varepsilon_n |\nabla_n u(x)|^2 \right] \mathrm{d}x.$$

The liminf inequality (see [12, Proposition 6.1]) is based on a standard blow-up argument (see [22]) at a point $x_0 \in \partial^* A$ to reduce to the case where the limiting function is $u_{0,\nu}$ and the domain is $Q_\nu \in \mathcal{Q}_\nu$, where $\nu = \nu_A(x_0)$. Then, a technical lemma (see [12, Lemma 3.1]) in the spirit of De Giorgi's slicing method (see [15]) allows to modify the given sequence $\{u_n\}_{n\in\mathbb{N}} \subset H^1(Q_\nu;\mathbb{R}^d)$ into a new sequence $\{v_n\}_{n\in\mathbb{N}} \subset H^1(Q_\nu;\mathbb{R}^d)$ with $v_n \to u_{0,\nu}$ in L^1, such that

$$\liminf_{n\to\infty} \mathcal{F}_{\varepsilon_n}(u_n) \geq \limsup_{n\to\infty} \mathcal{F}_{\varepsilon_n}(v_n),$$

and $v_n = \rho_n * u_{0,\nu}$ on ∂Q_ν, where $\rho_n(x) := \varepsilon_n^{-N}\rho(x/\varepsilon_n)$. The required inequality then follows by using a change of variable, and the definition of $\sigma(\nu)$ together with Lemma 1(iv).

The main challenges are related to the proof of the limsup inequality (see [12, Proposition 7.1]) for a function $u \in BV(\Omega, \{a, b\})$, which requires new geometric arguments. The idea is first to prove the result for functions $u \in BV(\Omega; \{a, b\})$ whose outer normals to the reduce boundary have rational coordinates, and then use the density of this class of functions in $BV(\Omega; \{a, b\})$ together with Reshetnyak's upper semi-continuity theorem (by Lemma 1(iii) the function $\nu \mapsto \sigma(\nu)$ is upper semi-continuous on $\mathbb{S}^{N-1}$) to conclude in the general case. In order to tackle the first step, we use a general strategy developed by De Giorgi, which can be seen as a sort of *reverse* blow-up argument: we consider the localized Γ-limsup as a map on Borel sets, and we prove that it is indeed a Radon measure λ. This is done by using a simplification of the De Giorgi–Letta coincidence criterion for Borel measures (see [16]) by Dal Maso, Fonseca, and Leoni (see [14, Corollary 5.2]). Next, we show that λ is absolutely continuous with respect to the measure $\mu := \mathcal{H}^{N-1} \llcorner \partial^* A$. The result follows by proving that the density of λ with respect to μ at a point $x_0 \in \partial^* A$ is bounded above by $\sigma(\nu_A(x_0))$. It is in this step that we exploit the fact that $\nu_A(x_0) \in \mathbb{S}^{N-1} \cap \mathbb{Q}^{N-1}$: indeed, by using the fact that W is periodic (with a different period) also as a function on any cube Q whose faces are normal to directions in $\mathbb{S}^{N-1} \cap \mathbb{Q}^{N-1}$, we can estimate the energy of a configuration similar to that in Fig. 1 on the left.

Remark 5 The strategy used to prove the above result is robust enough to be easily adapted to prove the analogous result when a mass constraint is enforced. Moreover, as a consequence of the Γ-limit result, we get that the function $\sigma : \mathbb{S}^{N-1} \to [0, \infty)$ is continuous, and its 1-homogeneous extension is convex.

The upshot of the foregoing result is that microscopic heterogeneities during phase transitions result in anisotropic surface tensions at the macroscopic level. Natural follow-up questions are:

1. Beyond convexity, what can one say about the effective surface tension σ? What functions σ are *attainable* as effective surface tensions of phase transitions in periodic media?
2. Considering the *gradient flow dynamics* of an energy as in (1), what are the $\varepsilon \to 0^+$ asymptotics? Does one indeed obtain a suitable weak formulation of anisotropic mean curvature flow, by analogy with the isotropic setting?

In [9], we provide partial answers to the first question above, by relating it to a geometry problem. In [10], we address dynamics. In the rest of this survey, we will summarize the results of [9], and a similar review of the results on dynamics will appear elsewhere [21].

In the sequel, we assume the product form of the potential W:

$$W(y, \xi) := a(y)(1 - u^2)^2, \quad y \in \mathbb{R}^N, u \in \mathbb{R}. \tag{3}$$

Here $a : \mathbb{R}^N \to \mathbb{R}$ is Q-periodic and non-degenerate in the sense that

$$\theta \leqslant a(y) \leqslant \Theta, \qquad y \in \mathbb{R}^N, \tag{4}$$

for some $0 < \theta < \Theta < \infty$. Note that assumptions (A1)–(A5) of Sect. 2.1 are satisfied with $z_1 = -1$, $z_2 = 1$, and $\widetilde{W} = W$. The fact that u is scalar-valued is crucial for a number of the results proven in [9, 10] since we use arguments based on the maximum principle. However, this is not true of all the results, and we will indicate this as appropriate.

2.2 Bounds on the Anisotropic Surface Tension σ

2.2.1 A Geometric Framework

Consider the periodic Riemannian metric on $\mathbb{R}^N$ that is conformal to the Euclidean one, defined as follows: given points $x, y \in \mathbb{R}^N$, we set

$$d_{\sqrt{a}}(x, y) := \inf_{\gamma} \int_0^1 \sqrt{a(\gamma(t))}|\dot{\gamma}(t)|\, dt,$$

where the infimum is taken over Lipschitz continuous curves $\gamma : [0, 1] \to \mathbb{R}^N$ such that $\gamma(0) = x, \gamma(1) = y$. It is easily seen that the formula defining $d_{\sqrt{a}}$ is independent of the parameterization of the competitor curves γ. Furthermore, standard arguments via the Hopf–Rinow theorem imply that $\mathbb{R}^N$ with the metric $d_{\sqrt{a}}$ is a complete metric space. Equivalently, geodesically complete: given any pair of points $x, y \in \mathbb{R}^N$, there exists a distance-minimizing geodesic joining them, whose length is equal to $d_{\sqrt{a}}(x, y)$ (see [31] for details). Now fix a direction $\nu \in \mathbb{S}^{N-1}$, and consider the plane Σ_ν through the origin with normal ν,

$$\Sigma_\nu := \{y \in \mathbb{R}^N : y \cdot \nu = 0\}.$$

Next, define the signed distance function in the $d_{\sqrt{a}}$-metric to the plane Σ_ν, via

$$h_\nu(y) := \operatorname{sgn}(y \cdot \nu) \inf_{z \in \Sigma_\nu} d_{\sqrt{a}}(y, z),$$

where the signum function is defined as

$$\operatorname{sgn}(t) := \begin{cases} 1 & t \geqslant 0, \\ -1 & t < 0. \end{cases}$$

It is easily shown (see [9, Lemma 2.2]) that h_ν is Lipschitz continuous, with

$$|\nabla h_\nu(y)| = \sqrt{a(y)} \qquad \text{at a.e. } y \in \mathbb{R}^N. \tag{5}$$

These observations, together with (4), yield

$$\begin{aligned} \sqrt{\theta}(y\cdot\nu) \leqslant h_\nu(y) \leqslant \sqrt{\Theta}(y\cdot\nu), \quad & y\cdot\nu \geqslant 0, \\ \sqrt{\Theta}(y\cdot\nu) \leqslant h_\nu(y) \leqslant \sqrt{\theta}(y\cdot\nu), \quad & y\cdot\nu < 0. \end{aligned} \tag{6}$$

In order to explain the relationship that the $d_{\sqrt{a}}$-metric bears with the anisotropic surface tension σ, it is useful to revisit the case $a \equiv 1$, and the celebrated Modica–Mortola example. In this case,

$$\sigma(\nu) = \lim_{T\to\infty} \frac{1}{T^{N-1}} \inf\left\{ \int_{TQ_\nu} \left[W(u(y)) + |\nabla u|^2 \right] : u \in C(\rho, Q_\nu, T) \right\}.$$

Elementary algebraic manipulations that effectively boil down to completing the square yield that the infimum above is asymptotically reached by the one-dimensional profile satisfying *equipartition of energy*. This entails, in the model case of (3), that the optimal cost is achieved by the choice $u(y) = q \circ (y\cdot\nu)$, where $q := \tanh$. The associated cost is given by

$$\sigma(\nu) \equiv \sigma_0 := \int_{-\infty}^{\infty} \left[W(q\circ(y\cdot\nu)) + |\nabla(q\circ(y\cdot\nu))|^2 \right] d(y\cdot\nu) = 2\int_{-1}^{1} \sqrt{W(s)}\, ds.$$

To make the connection to the $\sqrt{a}$-metric, we begin by noting that when $a \equiv 1$, we have $h_\nu(y) \equiv y\cdot\nu$. Our main motivation, then, is to obtain a similar formula that is exact when a is non-constant, or at least supplies reasonable bounds for the non-constant $\nu \mapsto \sigma(\nu)$. We do so by encoding the heterogeneous effects of a into the geometry of the underlying space, i.e., by working in the $\sqrt{a}$-metric. We turn to making these comments precise.

Fix $\nu \in \mathbb{S}^{N-1}$. Then, the cell formula defining $\sigma(\nu)$, proven in [12, 13] and specialized to our setting, reads (see Lemma 1 (iv))

$$\begin{aligned} \sigma(\nu) = \lim_{T\to\infty} \frac{1}{T^{N-1}} \inf\bigg\{ \int_{TQ_\nu} \left[a(y)W(u) + |\nabla u|^2 \right] dy : u \in H^1(TQ_\nu), \\ u = \rho * u_{0,\nu} \text{ on } \partial(TQ_\nu) \bigg\}. \end{aligned}$$

Here, we recall that $u_{0,\nu}(y) := \operatorname{sgn}(y\cdot\nu)$ and ρ is any standard smooth normalized mollifier (it is shown in Lemma 1(ii) that $\sigma(\nu)$ is independent of this choice). A preliminary step is to observe, by De Giorgi's slicing method (see [9, Lemma A.1]), that, equivalently,

$$\sigma(\nu) = \lim_{T\to\infty} \frac{1}{T^{N-1}} \inf\bigg\{ \int_{TQ_\nu} \left[a(y)W(u) + |\nabla u|^2 \right] dy : u \in H^1(TQ_\nu),$$

$$u = q \circ h_\nu \text{ along } \partial(TQ_\nu)\Big\}. \tag{7}$$

For each fixed $T \gg 1$, by the Direct Method of the Calculus of Variations, the variational problem inside the limit has a minimizer. Such a minimizer is, perhaps, not unique, but for each T we select one and call it u_T. We discuss various properties of u_T below in Sect. 2.2.2. In light of (7), it is clear by energy comparison that

$$\sigma(\nu) \leqslant \liminf_{T\to\infty} \frac{1}{T^{N-1}} \int_{TQ_\nu} [a(y)W(q\circ h_\nu) + |\nabla(q\circ h_\nu)|^2]\,dy.$$

Toward proving the opposite bound, we introduce the function $\phi : \mathbb{R} \to \mathbb{R}$ by

$$\phi(z) := 2\int_0^z \sqrt{W(s)}\,ds.$$

This function plays a fundamental role in the Modica–Mortola analysis corresponding to $a \equiv 1$. For any $T \gg 1$, using (5) and completing squares, we find

$$\begin{aligned}
&\frac{1}{T^{N-1}}\int_{TQ_\nu}\left[a(y)W(u_T) + |\nabla u_T|^2\right]dy\\
&\quad= \frac{2}{T^{N-1}}\int_{TQ_\nu}\nabla h_\nu\cdot\sqrt{W(u_T)}\nabla u_T\,dy + \frac{1}{T^{N-1}}\int_{TQ_\nu}\left|\nabla u_T - \sqrt{W(u_T)}\nabla h_\nu\right|^2\\
&\quad\geq \frac{1}{T^{N-1}}\int_{TQ_\nu}\nabla h_\nu\cdot\nabla(\phi(u_T))\,dy\\
&\quad= \frac{1}{T^{N-1}}\int_{TQ_\nu}\nabla h_\nu\cdot\nabla(\phi(q\circ h_\nu))\,dy + \frac{1}{T^{N-1}}\int_{TQ_\nu}\nabla h_\nu\cdot\nabla\left(\phi(u_T)-\phi(q\circ h_\nu)\right)\,dy\\
&\quad= \frac{1}{T^{N-1}}\int_{TQ_\nu}|\nabla h_\nu|^2\phi'(q\circ h_\nu)q'(h_\nu)\,dy\\
&\qquad+ \frac{1}{T^{N-1}}\int_{TQ_\nu}\nabla h_\nu\cdot\nabla\left(\phi(u_T)-\phi(q\circ h_\nu)\right)\,dy\\
&\quad= \frac{1}{T^{N-1}}\int_{TQ_\nu}2a(y)W(q\circ h_\nu)\,dy + \frac{1}{T^{N-1}}\int_{TQ_\nu}\nabla h_\nu\cdot\nabla\left(\phi(u_T)-\phi(q\circ h_\nu)\right)\,dy\\
&\quad= \frac{1}{T^{N-1}}\int_{TQ_\nu}\left[a(y)W(q\circ h_\nu) + |\nabla(q\circ h_\nu)|^2\right]dy\\
&\qquad+ \frac{1}{T^{N-1}}\int_{TQ_\nu}\nabla h_\nu\cdot\nabla\left(\phi(u_T)-\phi(q\circ h_\nu)\right)\,dy,
\end{aligned} \tag{8}$$

where in the last line we used the fact that the function $q\circ h_\nu$ achieves equipartition of energy. Indeed, by the definition of h_ν, we have

$$|\nabla(q\circ h_\nu)(y)|^2 = (q'(h_\nu(y)))^2|\nabla h_\nu(y)|^2 = a(y)W(q(h_\nu(y))).$$

Defining

$$\overline{\lambda}(\nu) := \limsup_{T\to\infty} \frac{1}{T^{N-1}} \int_{TQ_\nu} \left[a(y)W(q\circ h_\nu) + |\nabla(q\circ h_\nu)|^2\right] dy,$$
$$\underline{\lambda}(\nu) := \liminf_{T\to\infty} \frac{1}{T^{N-1}} \int_{TQ_\nu} \left[a(y)W(q\circ h_\nu) + |\nabla(q\circ h_\nu)|^2\right] dy,$$

provided we can control the error term

$$\limsup_{T\to\infty} \left| \frac{1}{T^{N-1}} \int_{TQ_\nu} \nabla h_\nu(y)\cdot\nabla\left(\phi(u_T) - \phi(q\circ h_\nu)\right) dy \right| := \lambda_0(\nu),$$

we observe that the test function $q \circ h_\nu$ gives two-sided bounds on $\sigma(\nu)$. Controlling the term λ_0 is complicated by the fact that it couples a product of weakly converging sequences (on expanding domains). Indeed, rescaling using $y = Tx$ in order to work in a fixed domain Q_ν, the two weakly converging factors making up the above product are:

1. *The oscillatory factor:* by (5) and (4), the term $\{\nabla h_\nu(T\,\cdot)\}_T$, which is bounded in L^∞, converges weakly-*.
2. *The concentration factor:* The terms $\nabla\phi(u_T(T\,\cdot))$ and $\nabla\phi(q\circ h_\nu(T\,\cdot)$ converge weakly-* to measures (see Sect. 2.2.2 for precise statements).

In particular, as one of the factors converges to a measure, standard tools such as compensated compactness, used traditionally to pass to the limit in products of weakly converging sequences, are unavailable, and we must control this term "by hand." In Sect. 2.2.2 below, we obtain fine information on the concentration effects; in Sect. 2.2.3, we deduce partial results concerning the oscillatory effects. Finally, we put these together in Sect. 2.2.4 where we obtain bounds on $\lambda_0(\nu)$.

2.2.2 Structure of Minimizers of the Cell Formula

For fixed $T \gg 1$, let $u_T \in C^2(TQ_\nu)$ (by elliptic regularity) be a minimizer of the energy

$$\int_{TQ_\nu} \left[a(y)W(u) + |\nabla u|^2\right] dy,$$

among competitors that equal $q \circ h_\nu$ along the boundary $\partial(TQ_\nu)$, and set

$$v_T(x) := u_T(Tx), \qquad x \in Q_\nu.$$

Lemma 2 *The functions v_T converge in L^1 to $u_{0,\nu} : Q_\nu \to \{\pm 1\}$.*

The proof of this lemma (see [9, Lemma 3.1]) is a nice application of the convexity of the one-homogeneous extension of σ (see Remark 5), using Jensen's inequality. The argument, without any changes, holds in the complete generality of the setting of [12] on the potential (vectorial, coupled, measurable dependence on the fast variable) and does not rely on the specific structure requested in (3). Combining Lemma 2 with the results of Caffarelli–Cordoba [8], we find that the level sets of v_T, for T sufficiently large, converge uniformly to $\Sigma_\nu \cap Q_\nu$.

Restricting ourselves to the scalar setting of (3), an argument using the strong maximum principle yields that for all $T < \infty$, we have

$$-1 < u_T(y) < 1$$

(see [9, Lemma 3.2]). In particular, $w_T := \frac{1}{\sqrt{2}} \tanh^{-1} u_T$ is well-defined, finite, and smooth in TQ_ν. Further, the function w_T verifies the elliptic boundary-value problem

$$\begin{cases} \Delta w_T = \frac{4}{\sqrt{2}} \tanh w_T \big(|\nabla w_T|^2 - a(y)\big), & y \in TQ_\nu, \\ w_T(y) = h_\nu(y) & y \in \partial(TQ_\nu). \end{cases}$$

Proposition 1 *Let w_T be as above, and let $T \gg 1$. There exist universal constants α_0 and $\eta_0 > 0$ such that the following holds:*

$$\begin{cases} \sqrt{\Theta}(y \cdot \nu) - \alpha_0 \geq w_T(y) \geq \sqrt{\theta}(y \cdot \nu) - \eta_0 & \textit{if } w_T(y) > 0, \\ -\sqrt{\theta}(y \cdot \nu) + \eta_0 \geq w_T(y) \geq -\sqrt{\Theta}(y \cdot \nu) + \alpha_0 & \textit{if } w_T(y) < 0. \end{cases} \tag{9}$$

Proposition 1 asserts that, up to universal constants, the function w_T satisfies exactly the same growth rates as the function h_ν, see (6). To prove Proposition 1, consider, for instance, the lower bound in the first of the two inequalities in (9). The main observation is that the function $y \mapsto \zeta_T(y) := \frac{y \cdot \nu}{w_T(y) + \eta_0}$ satisfies an elliptic PDE that verifies a maximum principle. The remaining inequalities follow from similar arguments, and we refer the reader to [9, Proposition 3.4] for details.

2.2.3 The Planar Metric Problem

Our results on the distance function h_ν concern its large-scale behavior. The bounds on σ that we discuss in Sect. 2.2.4 below depend solely on the large-scale behavior of the distance functions h_ν for which one can readily invoke efficient numerical algorithms, for example, fast marching and sweeping methods [30].

A natural question concerns the large-scale homogenized behavior of h_ν, i.e., characterize the limit

$$\lim_{T\to\infty} \frac{h_\nu(Ty)}{T}, \qquad y \in \mathbb{R}^N,$$

in a suitable topology of functions. We fully resolve this question (see also [3]) by characterizing uniform limits of the function $h(T\cdot)/T$.

Theorem 2 *Let* $\nu \in \mathbb{S}^{N-1}$*. Then, there exists a real number* $c(\nu) \in [\sqrt{\theta}, \sqrt{\Theta}]$*, and for each* $K \subseteq \mathbb{R}^N$ *compact, we have*

$$\lim_{T\to\infty} \sup_{y\in K} \left| \frac{1}{T} h_\nu(Ty) - c(\nu)(y\cdot\nu) \right| = 0.$$

Moreover, for all compact subsets L *of* $\mathbb{R}^N \setminus \Sigma_\nu$*, we have*

$$\lim_{T\to\infty} \sup_{y\in L} \left| \frac{1}{T(y\cdot\nu)} h_\nu(Ty) - c(\nu) \right| = 0.$$

We can interpret Theorem 2 as a homogenization result for the Eikonal equation in half-spaces. Indeed, it is well known (see for example [27]) that for each fixed $\nu \in \mathbb{S}^{N-1}$, the functions $k_m(y) := T_m^{-1} h_\nu(T_m(y))$ and $\ell(y) := c(\nu)(y\cdot\nu)$ are the unique viscosity solutions to

$$\begin{cases} |\nabla k_m| = \sqrt{a(T_m y)} & \text{in } \{y\cdot\nu \geq 0\}, \\ k_m = 0 & \text{on } \Sigma_\nu, \end{cases} \qquad \text{and} \qquad \begin{cases} |\nabla \ell| = c(\nu) & \text{in } \{y\cdot\nu \geq 0\}, \\ \ell = 0 & \text{on } \Sigma_\nu. \end{cases} \tag{10}$$

In fact, small modifications of our proofs permit us to prove almost periodic homogenization theorems for convex Hamiltonians with Bohr almost periodic dependence on the fast variable and Lipschitz continuous dependence on the slow variable (see [9, Theorem 1.4] for a precise statement). Theorem 2 shows that viscosity solutions of the PDEs on the left side of converge locally uniformly to the viscosity solution of the PDE on the right. A viscous and stochastic version of these equations (termed the "planar metric problem") was introduced by Armstrong and Cardaliaguet [3] and studied by others [19, 20] in the context of stochastic homogenization of geometric flows.

2.2.4 Bounds on the Anisotropic Surface Tension

As explained in the string of inequalities (8), the function $q \circ h_\nu$ provides tight upper and lower bounds for the effective anisotropy $\sigma(\nu)$. To be precise:

Theorem 3 *Let* $\sigma : \mathbb{S}^{N-1} \to [0, \infty)$ *be the anisotropic surface energy as in* (1)*. Let* $q : \mathbb{R} \to \mathbb{R}$ *be defined by*

$$q(z) := \tanh(z), \quad z \in \mathbb{R}.$$

For $\nu \in \mathbb{S}^{N-1}$, *define*

$$\underline{\lambda}(\nu) := \liminf_{T\to\infty} \frac{1}{T^{N-1}} \int_{TQ_\nu} \left[a(y)W(q \circ h_\nu) + |\nabla(q \circ h_\nu)|^2\right] dy,$$

$$\overline{\lambda}(\nu) := \limsup_{T\to\infty} \frac{1}{T^{N-1}} \int_{TQ_\nu} \left[a(y)W(q \circ h_\nu) + |\nabla(q \circ h_\nu)|^2\right] dy.$$

There exist $\Lambda_0 > 0$ *and* $\lambda_0 : \mathbb{S}^{N-1} \to [0, \Lambda_0]$ *such that*

$$\overline{\lambda}(\nu) - \lambda_0(\nu) \leqslant \sigma(\nu) \leqslant \underline{\lambda}(\nu).$$

We do not expect these to agree when $\nu \in \mathbb{Q}^N \cap \mathbb{S}^{N-1}$ owing to finite-size effects: in such directions, h_ν is periodic, and the problem is restricted to an infinite strip, rather than all of space (see [9, Lemma 2.3]). However, generically, i.e., when ν is an irrational direction, we conjecture that $\lambda_0(\nu) = 0$, so that $\underline{\lambda}(\nu) = \overline{\lambda}(\nu)$.

2.3 *Open Problems*

The studies presented above are a good source of interesting open problems. Here we list some of them.

2.3.1 Different Scales

For $\varepsilon, \delta > 0$, consider the energy

$$\mathcal{F}_{\varepsilon,\delta}(u) := \int_\Omega \left[\frac{1}{\varepsilon} W\left(\frac{x}{\delta}, u(x)\right) + \varepsilon|\nabla u(x)|^2 \right] \mathrm{d}x,$$

defined for functions $u \in H^1(\Omega; \mathbb{R}^d)$. Here the parameter ε is related to the phase transition process, while δ describes the scale of periodicity. In the functional (1), we considered the case $\varepsilon = \delta$, namely when the two phenomena act at the same scale, but it is interesting to understand what happens when one scale is dominant with respect to the other. Heuristically, we expect the limiting energy to be the same in the green and in the blue region (see Fig. 2). In particular, when $\varepsilon \ll \delta$, we expect the limiting functional $\mathcal{F}_0^P$ to be the homogenization of a surface energy functional, while in the other case, namely when $\delta \ll \varepsilon$, we expect to obtain the limit $\mathcal{F}_0^H$ of a classical Modica–Mortola functional whose potential is the homogenization of the original potential W.

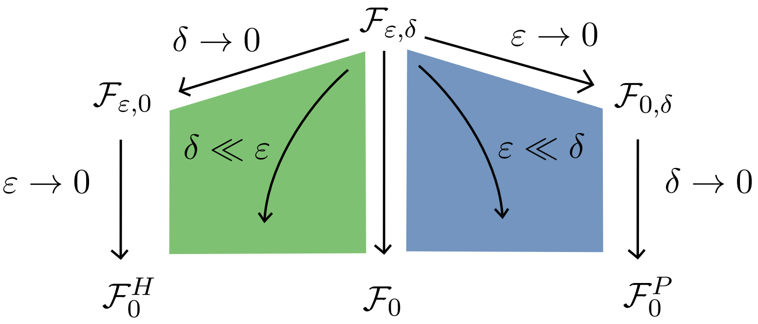


Fig. 2 The situation when phase transitions and homogenization act at possibly different scales

This latter situation was investigated in [26] under the additional assumption that the positive infinitesimal sequences $\{\varepsilon_n\}_{n\in\mathbb{N}}$ and $\{\delta_n\}_{n\in\mathbb{N}}$ satisfy

$$\lim_{n\to\infty} \frac{\varepsilon_n^{3/2}}{\delta_n} = +\infty, \tag{11}$$

and by assuming the potential W to be locally Lipschitz in the second variable, uniformly in the first one. In particular, it was proved that the limiting functional is

$$\mathcal{F}_0^H(u) := \begin{cases} K_H \mathcal{P}(\{u = z_1\}; \Omega) & \text{if } u \in BV(\Omega; \{z_1, z_2\}), \\ \\ +\infty & \text{otherwise,} \end{cases}$$

where $\mathcal{P}(\{u = z_1\}; \Omega)$ denotes the perimeter of the set $\{u = z_1\}$ in Ω, and the constant K_H is given by

$$K_H := 2\inf\left\{ \int_0^1 \sqrt{W_H(\gamma(s))}|\gamma'(s)|\mathrm{d}s : \gamma \in C^1([0,1];\mathbb{R}^d), \gamma(0) = z_1, \gamma(1) = z_2 \right\},$$

with the homogenized potential $W_H : \mathbb{R}^d \to [0, +\infty)$ given by $W_H(p) := \int_Q W(y, p)\,\mathrm{d}y$.

Some questions are still open: is this true also when $\delta \ll \varepsilon$ but without the extra assumption (11)? And what about the other regime?

2.3.2 Sharpness of Bounds and Inverse Homogenization

Various questions remain open from our discussion in 2.2. Our main contribution in that section was to relate the anisotropic surface tension σ to a purely geometric problem that had no concentration effects. Related to these bounds, we offer two open questions:

1. Examine the tightness of the bounds in Theorem 3, and closely related.
2. What does the set of effective anisotropies σ look like? In other words, which $\sigma : \mathbb{S}^{N-1} \to (0, \infty)$ with convex one-homogeneous extensions arise as a result of the homogenization procedure in [9]? Our bounds provide an approach to approximately solving this inverse homogenization question.

Acknowledgments I. Fonseca (I.F.) and R. Venkatraman (R.V.) acknowledge the Center for Nonlinear Analysis, where part of this work was carried out. The research of I.F. was partially funded by the National Science Foundation under Grant No. DMS-1906238. The research of R.V. was partially funded by the AMS-Simons Travel Award.

References

1. N. Ansini, A. Braides, V. Chiadò Piat, Interactions between homogenization and phase-transition processes. Trans. Mat. Inst. Steklova **236**, 386–398 (2002)
2. N. Ansini, A. Braides, V. Chiadò Piat, Gradient theory of phase transitions in composite media. Proc. Roy. Soc. Edinburgh Sect. A **133**, 265–296 (2003)
3. S. Armstrong, P. Cardaliaguet, Stochastic homogenization of quasilinear Hamilton-Jacobi equations and geometric motions. J. Eur. Math. Soc. **20**, 797–864 (2018)
4. S. Baldo, Minimal interface criterion for phase transitions in mixtures of Cahn-Hilliard fluids. Ann. Inst. H. Poincaré Anal. Non Linéaire **7**, 67–90 (1990)
5. A.C. Barroso, I. Fonseca, Anisotropic singular perturbations–the vectorial case. Proc. Roy. Soc. Edinburgh Sect. A **124**, 527–571 (1994)
6. K. Bhattacharya, Phase boundary propagation in a heterogeneous body. Proc. R. Soc. Lond. A **455**, 757–766 (1999)
7. A. Braides, C.I. Zeppieri, Multiscale analysis of a prototypical model for the interaction between microstructure and surface energy. Interfaces Free Bound. **11**, 61–118 (2009)
8. L.A. Caffarelli, A. Córdoba, Uniform convergence of a singular perturbation problem. Commun. Pure Appl. Math. **48**, 1–12 (1995)
9. R. Choksi, I. Fonseca, J. Lin, R. Venkatraman, Anisotropic surface tensions for phase transitions in periodic media. Calculus Var. Partial Differ. Equ. **61**(3), 1–41 (2021)
10. R. Choksi, I. Fonseca, J. Lin, R. Venkatraman, Homogenization for an Allen-Cahn equation in periodic media: a variational approach (2022, in preparation)
11. R. Cristoferi, G. Gravina, Sharp interface limit of a multi-phase transitions model under nonisothermal conditions. Calc. Var. Partial Differ. Equ. **60**, 142 (2021)
12. R. Cristoferi, I. Fonseca, A. Hagerty, C. Popovici, A homogenization result in the gradient theory of phase transitions. Interf. Free Bound. **21**, 367–408 (2019)
13. R. Cristoferi, I. Fonseca, A. Hagerty, C. Popovici, Erratum to: a homogenization result in the gradient theory of phase transitions. Interf. Free Bound. **22**, 245–250 (2020)
14. G. Dal Maso, I. Fonseca, G. Leoni, Nonlocal character of the reduced theory of thin films with higher order perturbations. Adv. Calc. Var. **3**, 287–319 (2010)
15. E. De Giorgi, Sulla convergenza di alcune successioni d'integrali del tipo dell'area. Rend. Mat. **8**, 277–294 (1975). Collection of articles dedicated to Mauro Picone on the occasion of his ninetieth birthday
16. E. De Giorgi, G. Letta, Une notion générale de convergence f faible pour des fonctions croissantes d'ensemble. Ann. Scuola Norm. Sup. Pisa Cl. Sci. **4**, 61–99 (1977)

17. N. Dirr, M. Lucia, M. Novaga, Γ-convergence of the Allen-Cahn energy with an oscillating forcing term. Interfaces Free Bound. **8**, 47–78 (2006)
18. N. Dirr, M. Lucia, M. Novaga, Gradient theory of phase transitions with a rapidly oscillating forcing term. Asymptot. Anal. **60**, 29–59 (2008)
19. W. Feldman, Mean curvature flow with positive random forcing in 2-d (2019). arXiv:1911.00488
20. W.M. Feldman, P.E. Souganidis, Homogenization and non-homogenization of certain non-convex Hamilton-Jacobi equations. J. Math. Pures Appl. **108**, 751–782 (2017)
21. R. Ferriera, I. Fonseca, R. Venkatraman, Variational homogenization: Old and new. *Proceedings of the ICM, 2022* (2021, in preparation)
22. I. Fonseca, S. Müller, Quasi-convex integrands and lower semicontinuity in L^1. SIAM J. Math. Anal. **23**, 1081–1098 (1992)
23. I. Fonseca, C. Popovici, Coupled singular perturbations for phase transitions. Asymptot. Anal. **44**, 299–325 (2005)
24. I. Fonseca, L. Tartar, The gradient theory of phase transitions for systems with two potential wells. Proc. Roy. Soc. Edinburgh Sect. A **111**, 89–102 (1989)
25. G.A. Francfort, S. Müller, Combined effects of homogenization and singular perturbations in elasticity. J. Reine Angew. Math. **454**, 1–35 (1994)
26. A. Hagerty, A note on homogenization effects on phase transition problems (2018). https://doi.org/10.48550/arxiv.1811.07357
27. C. Mantegazza, A.C. Mennucci, Hamilton-Jacobi equations and distance functions on Riemannian manifolds. Appl. Math. Optim. **47**, 1–25 (2003)
28. L. Modica, The gradient theory of phase transitions and the minimal interface criterion. Arch. Ration. Mech. Anal. **98**, 123–142 (1987)
29. S. Müller, Homogenization of nonconvex integral functionals and cellular elastic materials. Arch. Ration. Mech. Anal. **99**, 189–212 (1987)
30. J.A. Sethian, Fast marching methods. SIAM Rev. **41**, 199–235 (1999)
31. P. Sternberg, The effect of a singular perturbation on nonconvex variational problems. Arch. Ration. Mech. Anal. **101**, 209–260 (1988)

Some Recent Results on 2D Crystallization for Sticky Disc Models and Generalizations for Systems of Oriented Particles

Lucia De Luca

1 Introduction

The *crystallization* problem consists in understanding why, in appropriate physical conditions, systems of interacting particles arrange themselves into periodic lattices that assume a specific macroscopic shape [5]. Here we focus on this problem in the simplified setting of two-dimensional models at zero temperature, and we adopt a variational point of view, namely, we want to see crystallization as a phenomenon emerging from the minimization of suitable energy functionals. The results stated in this review are obtained in [11–14] that we refer to for the proofs and the main technical aspects; here we just present the basic strategy and the main ideas. We focus on pairwise interaction energies of the form

$$\mathcal{E}_{\mathscr{V}}(\mathsf{X}) := \frac{1}{2}\sum_{i\neq j}\mathscr{V}(|x_j - x_i|)\,, \tag{1}$$

where, assuming that the particles are point-like (resp., round-like), $\mathsf{X}=\{x_1,\dots,x_N\}$ ($N\in\mathbb{N}$) is the set representing the positions (resp., the positions of the barycenters) of the particles and $\mathscr{V}:[0,+\infty)\to[0,+\infty]$ is a short-range repulsive/long-range attractive interaction potential. Typically, $\mathscr{V}$ is a one-well potential such that $\lim_{r\to0^+}\mathscr{V}(r)=+\infty$, thus mimicking the constraint of non-interpenetration of bodies, and $\lim_{r\to+\infty}\mathscr{V}(r)=0$, which expresses the fact that two far away particles tend to ignore each other (see Fig. 1). Relevant examples of such a kind of potentials are the so-called Lennard–Jones $(p,2p)$ potentials defined by

L. De Luca (✉)
Istituto per le Applicazioni del Calcolo "M. Picone" IAC-CNR, Roma, Italy
e-mail: lucia.deluca@cnr.it

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_17

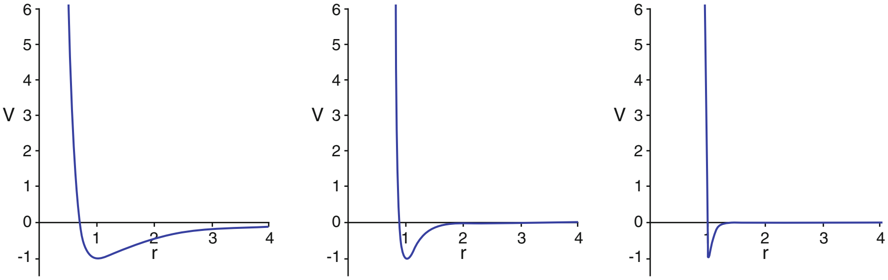


Fig. 1 The Lennard–Jones $(p, 2p)$ potentials for $p = 2$ (left), $p = 6$ (middle), $p = 20$ (right)

$$\mathscr{V}^{(p,2p)}(r) = r^{-2p} - 2r^{-p}\,, \qquad (p > 0)\,,$$

and, among them, the power $p = 6$ plays a prominent role for its chemical interpretation. Here, we have normalized the optimal interparticle distance to 1 and the associated potential energy to -1. Numerical simulations suggest 2d crystallization for the potentials $\mathscr{V}^{(p,2p)}$ [36]. Nowadays a proof of this fact is still missing; in [34], a crystallization result in the thermodynamic limit as the number of particles diverges is obtained for a one-well potential, qualitatively similar to the Lennard–Jones potential, satisfying suitably convexity/concavity assumptions. Here, we focus on the "brittle limit" as $p \to +\infty$ of the Lennard–Jones $(p, 2p)$ potentials in which the width of the well of the pair potential is compressed to zero and bonds immediately break upon increasing the interparticle distance. This procedure gives back the so-called *Heitmann–Radin sticky disc model* [23], where the interaction potential is of the form

$$\mathscr{V}_{HR}(r) = \begin{cases} +\infty & \text{if } r < 1\,, \\ -1 & \text{if } r = 1\,, \\ 0 & \text{if } r > 1\,. \end{cases} \tag{2}$$

For the above potential, Heitmann and Radin [23] proved that all the minimizers of the energy (1) are, up to rotation and translation, subsets of the regular triangular lattice

$$\mathscr{T} := \{z_1\mathbf{u_1} + z_2\mathbf{u_2} \,:\, z_1, z_2 \in \mathbb{Z}\}\,, \qquad \mathbf{u_1} = (1; 0)\,, \quad \mathbf{u_2} = (1/2; \sqrt{3}/2)\,,$$

and they provide an explicit minimizing configuration, which we will refer to as *canonical minimizer*, for every number of particles. Roughly speaking, such a canonical minimizer has the macroscopic shape of a regular hexagon with side length s if the number of particles is of the form $N = 3s^2 + 3s + 1$ and is given by an "hexagon plus a partial shell" in the other cases. The regular hexagon, namely, the Wulff shape for this problem, is actually the limit of any sequence of minimizers when the number of particles N tends to infinity and the interparticle distance is scaled by $N^{1/2}$ [2]. However, this is only the asymptotic behavior of the ground states since for a generic N the minimizers are "highly" non-unique. This fact has been proven by several authors: in [32], it has been proven that the scaling law for the fluctuation about the asymptotic Wulff shape is $CN^{3/4}$ for some $C > 0$, whereas in [10] the optimal constant C is explicitly provided (see also [30] for analogous results in the 3d cubic lattice and [8] for an interpretation of such a scaling law in terms of quantitative isoperimetric inequalities). In Sect. 3.1, we review the Heitmann–Radin crystallization result adopting the point of view of [12] and [11]. Furthermore, we determine also all the cardinalities of particles that guarantee, up to rotation and translation, the uniqueness of the minimizing configurations.

The results above deal with the behavior of minimizers for every fixed number N of particles. A natural question arising from such results concerns the behavior of quasi-minimizers, i.e., of configurations that are in the same asymptotic regime of the minimizers. In Sect. 3.2, we review the result in [13], and we show that the Heitmann–Radin energy enforces crystallization not only for minimizers, but also for low-energy configurations. But while for minimizers the orientation of the underlying lattice is constant, for almost minimizers global orientation can be disrupted, giving rise to polycrystalline structures. Moreover, we compute the Γ-limit of the energy functionals whenever the limiting orientation is constant, i.e., in the case of a single crystal. Such a result has been generalized in [20] to the very relevant case of limit polycrystals.

The results in Sect. 3 are obtained adopting a geometric approach that allows to rewrite the minimization of the sticky disc energy in terms of a suitable isoperimetric problem on the graphs generated by finite-energy configurations (see formula (3)). For such a reason in Sect. 2, we introduce the graph notions that will be needed in our analysis. We will see that the combination of graph tools and variational techniques successfully exploited in the context of the classical Heitmann–Radin model is robust enough to deal also with *vectorial crystallization problems*, namely, with crystallization for systems of oriented particles. This is the content of Sect. 4. In such a case, we focus on configurations $(X, V) = ((x_1; \dots; x_N), (v_1; \dots; v_N)) \in (\mathbb{R}^2)^N \times (\mathbb{S}^1)^N$, where x_i and v_i represent, respectively, the position and the orientation of the i-th particle; here and below $\mathbb{S}^1$ is the set of vector of $\mathbb{R}^2$ with unitary length. We consider pairwise interaction energies of the form

$$\mathcal{E}_{\mathcal{V}}(X, V) := \frac{1}{2} \sum_{i \neq j} \mathcal{V}(x_i, x_j, v_j, v_j),$$

where $\mathcal{V}$ is an interaction potential that is given by the sum of two terms, a hard core interaction depending on the mutual distance between the particles and the other one on both the mutual distance and orientation. Such a kind of interaction is governed by threshold criteria depending on a suitable parameter; tuning such a parameter, different minimal configurations are possible.

The interest toward this kind of model is motivated by the idea of building up a toy model for animal aggregations, such as fish schooling, bird flocking, or

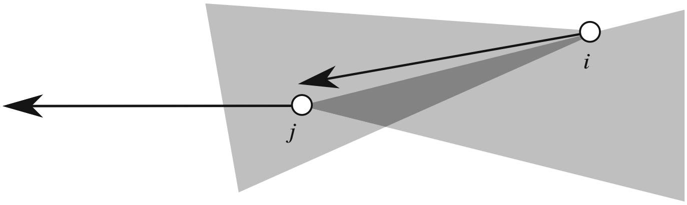


Fig. 2 The i-th fish follows the j-th fish

ducklings formation, which have been largely studied in modeling and simulation issues (see also [31] for the related theory of *boids*) but lacked of a rigorous mathematical analysis. The potential we consider encodes in a simple way the main modeling aspects [7, 15, 16, 26, 31] (Fig. 2) considered in fish schooling; in our framework, the i-th fish "follows" the j-th fish if they are at a suitable distance and if both the following conditions are satisfied:

- The j-th fish lies in the cone flap pointed at x_i, with axis parallel to v_i and angle amplitude θ_{visual}.
- The i-th fish lies in the cone flap pointed at x_j, with axis parallel to $-v_j$ and angle amplitude θ_{wake}.

As a further simplification, we assume $\theta_{\mathrm{visual}} = \theta_{\mathrm{wake}} =: \theta$ and $\theta \in [0, \pi]$. Therefore, renormalizing the hard core distance to 1, the model depends only on the parameter θ. We underline that the toy model described above is able to predict, rather than assume, alignment of orientations [9, 35]. When θ is small, the optimal configurations tend to be lines or cycles (for large enough particles), whereas for θ close to π, the minimizers tend to coincide with the ones of the Heitmann–Radin energy. The case $\theta = \frac{\pi}{3}$ provides the so-called *diamond formation* that is the one expected in fish schooling [6, 26, 27, 37]. Summarizing, we have shown that, in the Heitmann–Radin model, crystallization in the regular triangular lattice occurs for minimizers (unique global orientation) and for quasi-minimizers (polycrystals with possible different orientations), whereas considering a sticky disc potential, depending also on the orientation of particles, different structures (triangular lattice, diamond formation, one-dimensional configurations) can appear.

We remark that also models depending only on the particle positions can provide different lattice structures; by incorporating a three-body potential in the model, it is possible to obtain the honeycomb lattice [28] or the square lattice [29]. The latter can be obtained also, at least in terms of the energy per particle, by using only a Lennard–Jones type potential with large well (in order to include next-to-nearest neighbor interactions) [4]. Finally, all the results we have presented are only in dimension two. The situation in dimension one is nowadays quite well understood; indeed, it is immediate to check that the equi-spaced configurations with nearest-neighboring distance equal to 1 provide all the minimizers of the Heitmann–Radin energy (2) among the configurations of particles sitting on a straight line. The same result holds true, in terms of the energy density, also for the Lennard–Jones potential

$\mathscr{V}^{(6,12)}$ [21] (see also [24] for a recent extension of such a result to slightly more general potentials and [24, 25] for an asymptotic analysis of energy minimizers at low temperature). However, a complete crystallization result for *stable* one-well potentials exhibiting the "right behavior" at 0 and ∞ is still missing also in the one-dimensional case. As for the three-dimensional case, the situation is much less clear. The rough idea is that ground states of the Heitmann–Radin model should be given by a suitable superposition of parallel triangular lattice sheets; however, no proof of this fact is at disposal. The candidate lattices should then be the FCC lattice and the HCP lattice. In [17, 18], the authors have proven that the FCC lattice is a ground state in the limit as the number of particles diverges when adding a suitable three-body interaction to a Lennard–Jones type potential. It would be interesting to incorporate in the 2d and 3d models mentioned above a dependence on the orientation of particles as the one in Sect. 4 in order to study, at least numerically, how this affects the behavior of ground states. Moreover, our purely variational and static approach to vectorial crystallization could be compared with richer models [3] accounting for topological rather than metric distances, fluctuations, and dynamical aspects [9, 35].

2 Preliminaries on Planar Graphs

Here we collect some notions and notation on planar graphs that will be adopted in this chapter. We refer to [14, Section 1] for a complete analysis.

Let X be a finite subset of $\mathbb{R}^2$, and let Ed be a given subset of $\mathsf{E}(\mathsf{X})$, where

$$\mathsf{E}(\mathsf{X}) := \{\{x, y\} \subset \mathbb{R}^2 \,:\, x, y \in \mathsf{X}\,,\ x \neq y\}\,.$$

The pair $\mathsf{G} = (\mathsf{X}, \mathsf{Ed})$ is called *graph*; X is called the set of *vertices* of G, and Ed is called the set of *edges* (or *bonds*) of G .

Given $\mathsf{X}' \subset \mathsf{X}$, we denote by $\mathsf{G}_{\mathsf{X}'}$ the *subgraph* (or *restriction*) of G generated by X' , defined by $\mathsf{G}_{\mathsf{X}'} = (\mathsf{X}', \mathsf{Ed}')$, where $\mathsf{Ed}' := \{\{x', y'\} \in \mathsf{Ed} \,:\, x', y' \in \mathsf{X}'\}$.

Definition 1 We say that two points $x, z \in \mathsf{X}$ are connected, and we write $x \sim z$ if there exist $M \in \mathbb{N}$ and a *path* $x = y_0, \dots, y_M = z$ such that $\{y_{m-1}, y_m\} \in \mathsf{Ed}$ for every $m = 1, \dots, M-1$. We say that $\mathsf{G}_{\mathsf{X}_1}, \dots, \mathsf{G}_{\mathsf{X}_K}$ with $K \in \mathbb{N}$ are the *connected components* of G if $\{\mathsf{X}_1, \dots, \mathsf{X}_K\}$ is a partition of X, and for every $k, k' \in \{1, \dots, K\}$ with $k \neq k'$, it holds

$$\begin{aligned} &x_k \sim y_k && \text{for every } x_k, y_k \in \mathsf{X}_k\,,\\ &x_k \not\sim x_{k'} && \text{for every } x_k \in \mathsf{X}_k\,, x_{k'} \in \mathsf{X}_{k'}\,. \end{aligned}$$

If G has only one connected component, we say that G is *connected.*

We say that G is planar if for every pair of (distinct) bonds $\{x_1, x_2\}, \{y_1, y_2\} \in \mathsf{Ed}$, the (open) segments (x_1, x_2) and (y_1, y_2) have empty intersection.

From now on, we assume that $\mathsf{G} = (\mathsf{X}, \mathsf{Ed})$ is planar, so that we can introduce the notion of face (see also [12]).

By a face f of G, we mean any open, bounded, connected component of $\mathbb{R}^2 \setminus \big(\mathsf{X} \cup \bigcup_{\{x,y\}\in\mathsf{Ed}}[x, y]\big)$, which is also simply connected; here $[x, y]$ is the closed segment with extreme points x and y. We denote by $\mathsf{F}(\mathsf{G})$, the set of faces of G, and we set

$$O(\mathsf{G}) := \bigcup_{f\in\mathsf{F}(\mathsf{G})} \operatorname{clos}(f)\,.$$

We define the Euler characteristic of G as

$$\chi(\mathsf{G}) = \sharp\mathsf{X} - \sharp\mathsf{Ed} + \sharp\mathsf{F}(\mathsf{G})\,,$$

and we warn the reader that this may differ from the standard Euler characteristic in graph theory (see [14, Lemma 1.2 & Remark 1.3]). We just remark that if $\chi(\mathsf{G}) = 1$, then G is connected.

With a little abuse of language, we will say that an edge $\{x, y\}$ lies on a set $E \subset \mathbb{R}^2$ if the segment $[x, y]$ is contained in E. We classify the edges in Ed in the following subclasses:

- $\mathsf{Ed}^{\mathrm{int}}$ is the set of *interior edges*, i.e., of edges lying on the boundary of two (distinct) faces.
- $\mathsf{Ed}^{\mathrm{wire,ext}}$ is the set of *exterior wire edges*, i.e., of edges that do not lie on the boundary of any face.
- $\mathsf{Ed}^{\mathrm{wire,int}}$ is the set of *interior wire edges*, i.e., of edges lying on the boundary of precisely one face but not on the boundary of its closure (or, equivalently, of $O(\mathsf{G})$).
- Ed^{∂} is the set of *boundary edges*, i.e., of edges lying on $\partial O(\mathsf{G})$.

Analogously, for every face $f \in \mathsf{F}(\mathsf{G})$, one can define the following subclasses of edges delimiting f:

- $\mathsf{Ed}^{\mathrm{wire,int}}(f)$ is the set of edges lying on the boundary of f but not on the boundary of the closure of f.
- $\mathsf{Ed}^{\partial}(f)$ is the set of edges lying on the boundary of the closure of f.

We define the *graph perimeter* of G as

$$\mathrm{Per}_{\mathsf{gr}}(\mathsf{G}) := \sharp\mathsf{Ed}^{\partial} + 2\sharp\mathsf{Ed}^{\mathrm{wire,ext}}\,.$$

Analogously, the *graph perimeter* of a face f is defined by

$$\mathrm{Per}_{\mathsf{gr}}(f) := \sharp\mathsf{Ed}^{\partial}(f) + 2\sharp\mathsf{Ed}^{\mathrm{wire,int}}(f).$$

3 The Sticky Disc Model: Minimizers and Quasi-minimizers

Here we briefly review the crystallization result in [23] following the approach in [12]. For every $N \in \mathbb{N}$, we denote by $\mathcal{A}_N$ the set of N-particle configurations with finite energy, i.e., $\mathcal{A}_N := \{\mathsf{X} \subset \mathbb{R}^2 : \sharp\mathsf{X} = N\,, \mathcal{E}_{\mathcal{V}_{HR}}(\mathsf{X}) < +\infty\}$, and we set $\mathcal{A} := \bigcup_{N\in\mathbb{N}} \mathcal{A}_N\,.$

For every $\mathsf{X} \in \mathcal{A}$, we denote by $\mathsf{G}(\mathsf{X})$ the *graph generated by* X, i.e., $\mathsf{G}(\mathsf{X}) = (\mathsf{X}, \mathsf{Ed}(\mathsf{X}))$, where $\mathsf{Ed}(\mathsf{X}) := \{\{x, y\} : x, y \in \mathsf{X}\,, |x - y| = 1\}$, and we notice that

$$\mathcal{E}_{\mathcal{V}_{HR}}(\mathsf{X}) = -\sharp\mathsf{Ed}(\mathsf{X})\,.$$

Notice that the finiteness of $\mathcal{E}_{\mathcal{V}_{HR}}(\mathsf{X})$, implies that $\mathsf{G}(\mathsf{X})$ is a planar graph and that, denoting by $\mathsf{F}^{\triangle}(\mathsf{G}(\mathsf{X}))$ the set of the triangular faces of $\mathsf{G}(\mathsf{X})$, the elements of $\mathsf{F}^{\triangle}(\mathsf{G}(\mathsf{X}))$ are equilateral triangles with unitary side length.

3.1 *Minimizers of the Heitmann–Radin Sticky Disc Model: Single Crystals*

The crystallization result for the Heitmann–Radin model in our notations reads as follows.

Theorem 1 *Let $N \in \mathbb{N}$, and let X_N be a minimizer of $\mathcal{E}_{\mathcal{V}_{HR}}$ in $\mathcal{A}_N$. Then $G(X_N)$ is connected and, up to rotation and translation, $X_N \subset \mathcal{T}$. Moreover, if $N \geq 3$, then $O(G(X_N))$ has simply closed polygonal boundary, $F(G(X_N)) = F^{\triangle}(G(X_N))$ and $\sharp Ed^{\text{wire,ext}}(G(X_N)) = \emptyset$.*

Theorem 1 has been proved by Heitmann and Radin in [23], following an ansatz on the numerical value of minimal energy previously found by Harborth [22]. In [12], a slightly different proof of Theorem 1 has been found, based on a geometric interpretation of the Heitmann–Radin energy rather than on the result in [22]. The rough idea of the proof in [12] is the following: Using Euler characteristic formula, one can prove that (see [12, Theorem 3.1]) the energy of any configuration $\mathsf{X} \in \mathcal{A}$ can be decomposed into the sum of a volume term that is negative and depend only on the number of particles and a perimeter/defect term that counts the number of points having the “wrong” number of nearest neighbors; more precisely,

$$\mathcal{E}_{\mathcal{V}_{HR}}(\mathsf{X}) = -3\sharp\mathsf{X} + \mathrm{Per}_{\mathrm{gr}}(\mathsf{G}(\mathsf{X})) + \mathcal{E}^{\mathrm{def}}_{\mathcal{V}_{HR}}(\mathsf{X}) + 3\chi(\mathsf{G}(\mathsf{X}))\,, \tag{3}$$

where $\mathcal{E}^{\mathrm{def}}_{\mathcal{V}_{HR}}(\mathsf{X}) := \sum_{f\notin\mathsf{F}^{\triangle}(\mathsf{G}(\mathsf{X}))}(\mathrm{Per}\mathsf{gr}(f) - 3)\,.$

With decomposition (3) in hand, noticing immediately that for minimizers X the graph $\mathsf{G}(\mathsf{X})$ should be connected and cannot have wire edges, finding the ground states of $\mathcal{E}_{\mathcal{V}_{HR}}$ in $\mathcal{A}_N$ (for some $N \in \mathbb{N}$) coincides with minimizing the perimeter-

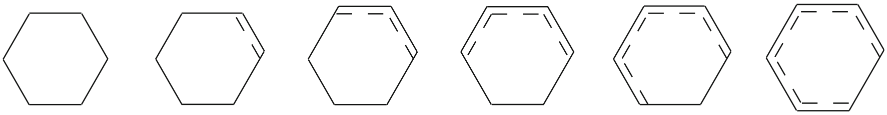


Fig. 3 The macroscopic shapes of minimizers in the cases of uniqueness. Starting from the left: $N = 3s^2+3s+1$, $N = 3s^2+3s+1+s$, $N = 3s^2+3s+1+2s+1$, $N = 3s^2+3s+1+3s+2$, $N = 3s^2 + 3s + 1 + 4s + 3$, $N = 3s^2 + 3s + 1 + 5s + 4$. The unique minimizer is a regular hexagon with side length s in the first case, whereas in the other cases is a regular hexagon plus "a suitable partial shell" around it (the sides of the regular hexagon contained in the Wulff shape are represented by the dashed lines)

like term

$$\mathcal{F}(\mathsf{X}) := \mathrm{Per}(O(\mathsf{X})) + \mathcal{E}^{\mathrm{def}}_{\mathcal{V}_{HR}}(\mathsf{X}) . \tag{4}$$

Now, the proof of Theorem 1 is just a consequence of the following elementary facts:

for every $\mathsf{X} \in \mathcal{A}$ it holds $\mathcal{F}(\mathsf{X}) \geq \mathcal{F}(\mathsf{X} \setminus \partial\mathsf{X}) + 6$ where $\partial\mathsf{X}$ denotes the set of the *boundary particles* of X, i.e., of the particles lying only on boundary edges; (a)

$\mathcal{F}(\mathsf{X}) = \mathcal{F}(\mathsf{X} \setminus \partial\mathsf{X}) + 6$ only if the graph generated by $\mathsf{X} \setminus \partial\mathsf{X}$ has only triangular faces. (b)

By adopting an induction procedure and using (a) and (b) (see [12, Section 5]), it is possible to show that the minimum of $\mathcal{F}$ is achieved when $\mathcal{E}^{\mathrm{def}}_{\mathcal{V}_{HR}} \equiv 0$, i.e., for subsets of the regular triangular lattice.

Using the procedure described above, in [11], the following result on the uniqueness of minimizers has been proven (Fig. 3).

Theorem 2 *Let $N \in \mathbb{N}$. The minimizers of $\mathcal{E}_{\mathcal{V}_{HR}}$ in $\mathcal{A}_N$ are unique, up to rotation and translation, if and only if either*

$$N = 3s^2 + 3s + 1 \qquad \textit{for some } s \in \mathbb{N} \cup \{0\} , \tag{5}$$

or

$$N=3s^2+3s+1+(s+1)k+s \quad \textit{for some } s \in \mathbb{N} \cup \{0\} , k \in \mathbb{N} \cup \{0\} , \textit{ with } 0 \leq k \leq 4 .$$

We remark that the uniqueness in the case (5) was already observed in [32, Proposition 2.7], using a discrete-to-continuum approach that exploits the results [19, 33] on the uniqueness of the Wulff shape for crystalline perimeters (see Sect. 3.2). Our approach is, instead, purely discrete, and it allows to get uniqueness also in the case that the macroscopic shape of minimizers is no longer a regular hexagon, and hence, the associated continuum sets introduced in [32] do not minimize the corresponding continuum crystalline perimeter.

3.2 *Quasi-minimizers of the Heitmann–Radin Model: Polycrystalline Structures*

We aim at determining the asymptotic behavior, as $N \to +\infty$ of the configurations $X_N \in \mathcal{A}_N$ in the scaling perimeter regime, i.e., such that

$$\mathcal{E}_{\mathcal{V}_{HR}}(\mathsf{X}_N) + 3N \leq CN^{1/2}\,. \tag{6}$$

To this purpose, to every $\mathsf{X} \in \mathcal{A}$, we associate the empirical measure $\mu = \mu^{\mathsf{X}} := \sum_{x\in\mathsf{X}} \delta_x$. Moreover, one can easily check that for configurations satisfying the energy bound (6), "almost all" the faces are triangles, so that also in the case of quasi-minimizers, triangles will play a crucial role. For such a reason, we introduce a notion of orientation of triangles by associating at each triangle $f \in \mathsf{F}(\mathsf{G}(\mathsf{X}))$ the quantity $\theta(f) \in (\frac{\pi}{3}, \frac{2}{3}\pi]$ given by the angle between e_1 and one of the medians of f (see [13, Subsection 2.3]).

For every $N \in \mathbb{N}$ and for every $\mathsf{X}_N \in \mathcal{A}_N$, we define the *orientation map* as

$$\theta_N(\mathsf{X}_N) := \sum_{f\in F_\varepsilon^\triangle(\mathsf{X}_N)} \theta(f)\mathbb{1}_{\frac{f}{\sqrt{N}}}\,.$$

The following result is [13, Theorem 3.1] . We refer to [1] for the definitions of finite perimeter sets and of Caccioppoli partitions.

Theorem 3 *Let $\{X_N\}_{N\in\mathbb{N}}$ be a sequence of sets of points such that X_N belongs to $\mathcal{A}_N$ for every $N \in \mathbb{N}$ and satisfies (6) for some universal constant C (independent of N). Then, up to a subsequence:*

(i) $N\frac{\sqrt{3}}{2}\mu^{\mathsf{X}_N} \overset{*}{\rightharpoonup} \mathbb{1}_\Omega\,\mathrm{d}x$ *(as $N \to +\infty$), for some set $\Omega \subset \mathbb{R}^2$ with finite perimeter.*

(ii) $\theta_N(X_N) \rightharpoonup \theta$ *in $SBV_{\mathrm{loc}}(\mathbb{R}^2)$ (as $N \to +\infty$), for some $\theta = \sum_{j\in J} \theta_j \mathbb{1}_{\omega_j}$ in $SBV(\mathbb{R}^2)$, where $J \subseteq \mathbb{N}$, $\{\omega_j\}_{j\in J}$ is a Caccioppoli partition of Ω, and $\{\theta_j\}_{j\in J} \subset (\frac{\pi}{3}, \frac{2}{3}\pi]$.*

Moreover, we computed the effective energy of $\mathcal{E}_{\mathcal{V}_{HR}}$ when the number of particles diverges in the case of a unique limit orientation $\bar{\theta}$. This is given by the crystalline perimeter $\mathrm{Per}_{\varphi_{\bar{\theta}}}$ whose ball is the unitary hexagon "oriented according with $\bar{\theta}$" (see [13, Subsection 2.4] for its precise definition). This is the content of [13, Theorem 3.2] that is stated below.

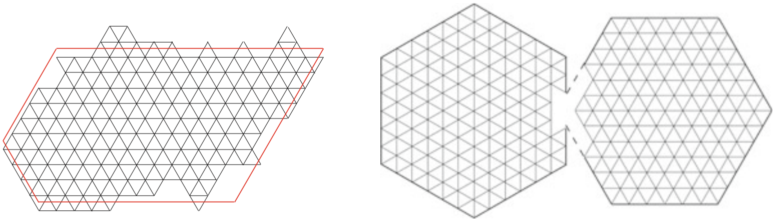


Fig. 4 Left: If Ω is a polygon with side parallel to the Wulff shape associated to $\varphi_{\bar{\theta}}$ for some $\bar{\theta} \in (\frac{\pi}{3}, \frac{2}{3}\pi]$, then the limit orientation is $\bar{\theta}$, i.e., $\theta_N(\mathsf{X}_N) \rightharpoonup \bar{\theta}\mathbb{1}_\Omega$. Right: If Ω is given by a small enough intersection of two regular hexagons with different orientations, the limit orientation is made by at most two grains, giving rise to a polycrystal

Theorem 4 *The following* Γ*-convergence result holds true:*

(i) (Γ*-liminf inequality) Let* $\{\mathsf{X}_N\}_{N\in\mathbb{N}}$ *satisfy (i) and (ii) of Theorem 3 with* $\theta = \bar{\theta}\mathbb{1}_\Omega$ *for some* $\bar{\theta} \in (\frac{\pi}{3}, \frac{2}{3}\pi]$*. Then*

$$\liminf_{N\to+\infty} \frac{1}{N^{1/2}}(\mathcal{E}_{\mathcal{V}_{HR}}(\mathsf{X}_N) + 3N) \geq \mathrm{Per}_{\varphi_{\bar{\theta}}}(\Omega)\,.$$

(ii) (Γ*-limsup inequality) For every set* $\Omega \subset \mathbb{R}^2$ *of finite perimeter and for every* $\bar{\theta} \in (\frac{\pi}{3}, \frac{2}{3}\pi]$*, there exists a sequence* $\{X_N\}_{N\in\mathbb{N}}$ *satisfying (i) and (ii) of Theorem 3 with* $\theta = \bar{\theta}\mathbb{1}_\Omega$ *such that:*

$$\limsup_{N\to+\infty} \frac{1}{N^{1/2}}(\mathcal{E}_{\mathcal{V}_{HR}}(\mathsf{X}_N) + 3N) \leq \mathrm{Per}_{\varphi_{\bar{\theta}}}(\Omega)\,.$$

In the case of limit polycrystal, [13] do not provide a full result (see Fig. 4). However, in [13, Subsection 4.2]), it is shown that depending on the shape of Ω, single crystals and polycrystals could be preferable when considering the minimum problem

$$\inf_{N\frac{\sqrt{3}}{2}\mu_{\mathsf{X}_N} \overset{*}{\rightharpoonup} \mathbb{1}_\Omega\,\mathrm{d}x} \ \liminf_{N\to+\infty} \frac{1}{N^{1/2}}(\mathcal{E}_{\mathcal{V}_{HR}}(\mathsf{X}_N) + 3N)\,.$$

The general case of limit polycrystal has been treated in [20] where the precise line tension energy at grain boundaries is provided.

4 Vectorial Crystallization and Collective Behavior

Here we discuss a sticky disc type model for collective behavior of oriented particles, following the approach in [14]. The modeling aspects of such a theory are described in the introduction; here we focus on qualitative properties of minimizers for different values of the angle parameter θ. In order to have notation coherent with [14], the different behaviors of minimizers in our analysis will depend on the parameter $\gamma := \cos\frac{\theta}{2}$.

For every $\gamma \in [0, 1]$, let $\mathscr{V}^\gamma : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{S}^1 \times \mathbb{S}^1 \to [0, +\infty)$ be defined as

$$\mathscr{V}^\gamma(y_1, y_2, w_1, w_2) := \begin{cases} +\infty \text{ if } |y_1 - y_2| < 1\,, \\ -1 \quad \text{if } |y_1 - y_2| = 1 \text{ and either } \langle y_1 - y_2, w_k\rangle \geq \gamma \text{ for } k = 1, 2 \\ \qquad \text{or } \langle y_2 - y_1, w_k\rangle \geq \gamma \text{ for } k = 1, 2\,, \\ 0 \qquad \text{elsewhere.} \end{cases}$$

For every $N \in \mathbb{N}$, let $C_N := \{(X, V) : X = (x_1, \ldots, x_N) \in (\mathbb{R}^2)^N, V = (v_1, \ldots, v_N) \in (\mathbb{S}^1)^N\}$ and set $C := \bigcup_{N\in\mathbb{N}} C_N$. We define the energy $\mathcal{E}_{V^\gamma} : C \to [0, +\infty]$ as

$$\mathcal{E}_{\mathscr{V}^\gamma}(X, V) := \frac{1}{2}\sum_{i\neq j} \mathscr{V}^\gamma(x_i, x_j, v_i, v_j),$$

and we aim at establishing qualitative properties of the minimizers of $\mathcal{E}_\gamma$ for different values of γ. Following the approach in Sect. 3, we introduce a graph structure on configurations $(X, V) \in C$, by defining the set

$$\mathsf{Ed}^\gamma(X, V) := \{\{x_i, x_j\} : x_i, x_j \in \mathsf{X},\ \mathscr{V}^\gamma(x_i, x_j, v_i, v_j) = -1\}\ ,$$

where X denotes the set of particle positions, i.e., $\mathsf{X} = \{x_1, \ldots, x_N\}$.

In such a way, we have that configurations (X, V) having finite energy generate a planar graph $\mathsf{G}(X, V) = (\mathsf{X}, \mathsf{Ed}^\gamma(X, V))$ and that $\mathcal{E}_{\mathscr{V}^\gamma}(X, V) = -\sharp\mathsf{Ed}^\gamma(X, V)$. Now we discuss the behavior of minimizers for different values of $\gamma \in [0, 1]$.

We notice that the case $\gamma = 0$ is a slight generalization of the Heitmann–Radin model, being the dependence on the orientations only fictitious. In particular, for configurations with constant orientation, i.e., with $v_i \equiv v$ for some $v \in \mathbb{S}^1$, the class of minimizers (resp. almost minimizers) of $\mathcal{E}_{\mathscr{V}^0}$ coincides with the class of minimizers (resp. almost minimizers) of $\mathcal{E}_{\mathscr{V}_{HR}}$ analyzed in Sect. 3.

Now we focus on the case $0 < \gamma \leq \frac{1}{2}$. In such a range, for suitably chosen constant orientations, the ground states are the same as in the Heitmann–Radin model; in particular, they are subsets of the unitary triangular lattice $\mathcal{T}$ (see [14, Subsection 2.1]). However, not all constant orientations do the job as shown in Fig. 5 below.

If $\frac{1}{2} < \gamma \leq \frac{\sqrt{3}}{2}$, then the maximal number of nearest neighbors passes from 6 to 4, so that the asymptotic (as $N \to +\infty$) energy per particle for minimizers equals to -2; more precisely,

$$-2N \leq \min_{(X,V)\in\mathcal{AC}_N} \mathcal{E}^\gamma(X, V) \leq -2N + CN^{\frac{1}{2}}\,, \tag{7}$$

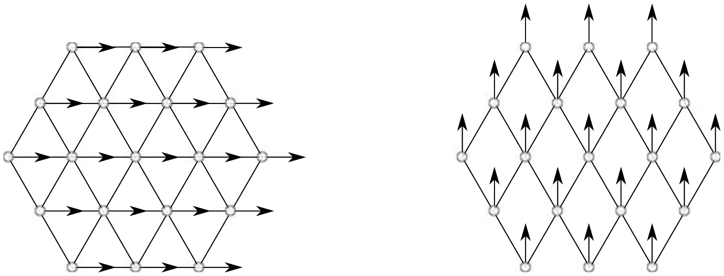


Fig. 5 The unique, up to rotation, minimizer X of the Heitmann–Radin energy for $N = 19$. On the left: the configuration with constant orientation $v_i \equiv (1; 0)$, for which the bonds coincide with the ones in the Heitmann–Radin case. On the right: the configuration with constant orientation $v_i \equiv (0; 1)$ that neglects the horizontal bonds in the Heitmann–Radin model

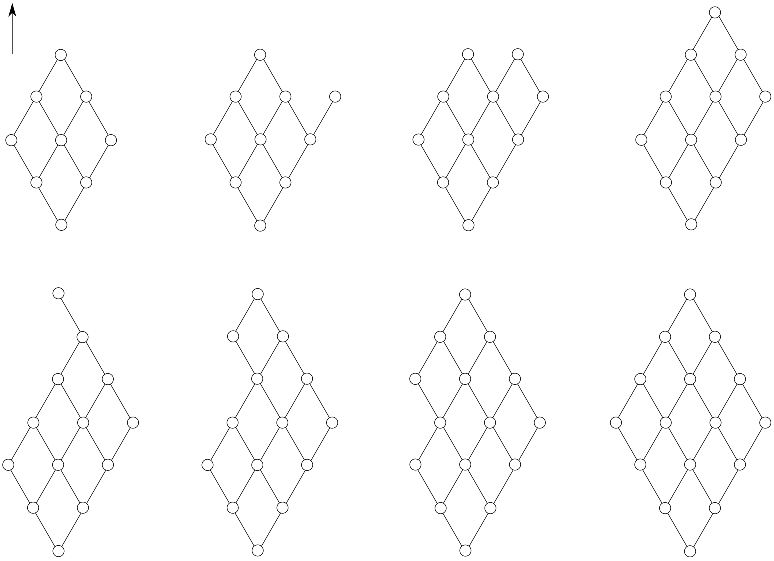


Fig. 6 The graphs $\mathsf{G}(Y_N, W_N)$ of the configurations (Y_N, W_N) for $N = 9, \ldots, 16$

for some $C > 0$ independent of N. In order to construct a configuration that provides the second inequality in (7), it is enough to take $v_i \equiv (0; 1)$ and $\mathsf{X} \subset \mathcal{T}$ in such a way that if $N = (l+1)^2$, then $O(\mathsf{X})$ is a rhombus with side length l, whereas for other values of N $O(\mathsf{X})$ is a rhombus plus an "incomplete shell" around it. We will denote such configurations by (Y_N, W_N) (see Fig. 6).

We notice that, for $\frac{1}{2} < \gamma < \frac{\sqrt{3}}{2}$, small perturbations of the configuration (Y_N, W_N) constructed above still yield almost minimizers for $\mathcal{E}_{\mathcal{V}\gamma}$ satisfying (7). By construction, such perturbations should map the equilateral triangular lattice into a suitable monoclinic lattice. Moreover, one can easily see that energy is invariant also under small perturbations of the orientation field W_N.

The same behavior does not appear in the case $\gamma = \frac{\sqrt{3}}{2}$ that is actually more rigid. Indeed, in such a case, a point having the maximal number of nearest neighbors should lie on 4 bonds forming alternate angles equal to $\frac{\pi}{3}$ and $\frac{2}{3}\pi$. In such a case the orientation associated to this point should be parallel to the bisector of the $\frac{\pi}{3}$-angles and hence orthogonal to the one of the $\frac{2}{3}\pi$-angles (see [14, Lemma 3.1]).

Using this kind of geometric considerations, it is possible to prove that the configurations (Y_N, W_N) are not only asymptotic minimizers of $\mathcal{E}_{\mathscr{V}\frac{\sqrt{3}}{2}}$ but actually minimizers in $\mathcal{C}_N$ for every $N \in \mathbb{N}$, i.e., the following result (see [14, Theorem 3.7]).

Theorem 5 *For every $N \in \mathbb{N}$, it holds*

$$\mathcal{E}_{\mathscr{V}\frac{\sqrt{3}}{2}}(Y_N, W_N) = \min_{(X,V)\in\mathcal{C}_N} \mathcal{E}_{\mathscr{V}\frac{\sqrt{3}}{2}}(X, V)\,.$$

The proof of Theorem 5 uses an approach similar to that of Theorem 1. The starting point is also in this case an energy decomposition into a negative volume part and a (positive) surface-like term; such a decomposition is proven in [14, Proposition 3.4] and reads as

$$\mathcal{E}_{\mathscr{V}\frac{\sqrt{3}}{2}}(X, V) := -2\sharp\mathsf{X} + \frac{1}{2}\mathrm{Per}_{\mathsf{gr}}(\mathsf{G}(X, V)) + \frac{1}{2}\mathcal{E}^{\mathrm{def}}_{\mathscr{V}\frac{\sqrt{3}}{2}}((X, V)) + 2\chi(\mathsf{G}(X, V))\,,$$

where the defect energy $\mathcal{E}^{\mathrm{def}}_{\mathscr{V}\frac{\sqrt{3}}{2}}$ in this case penalizes the non-rhombic faces. Moreover, also in such a case, we have that a finite-energy configuration (X, V) satisfies the bound

$$\begin{aligned}\frac{1}{2}\mathrm{Per}_{\mathsf{gr}}(\mathsf{G}(X, V)) + \frac{1}{2}\mathcal{E}^{\mathrm{def}}_{\mathscr{V}\frac{\sqrt{3}}{2}}((X, V)) \geq{}& \frac{1}{2}\mathrm{Per}_{\mathsf{gr}}(\mathsf{G}(X \setminus \partial X, V \setminus \partial V)) \\ &+ \frac{1}{2}\mathcal{E}^{\mathrm{def}}_{\mathscr{V}\frac{\sqrt{3}}{2}}((X \setminus \partial X, V \setminus \partial V)) + 4\,,\end{aligned} \tag{8}$$

where $\mathsf{G}(X \setminus \partial X, V \setminus \partial V)$ is obtained by (X, V) once removing the boundary particles together with the corresponding orientations. Notice that (8) corresponds to (a) in Sect. 3.1, and this is enough, in order to show Theorem 5 (which is actually weaker than Theorem 1). In order to get a general rhombic crystallization result, it would be enough to have also property (b) in our case. Such a property does not hold in this framework, and actually a general rhombic crystallization result for all the minimizers of $\mathcal{E}_{\mathscr{V}\frac{\sqrt{3}}{2}}$ fails to be true (see [14, Remark 3.8]), as shown by Fig. 7 below. Nevertheless, although at the present, we do not have a proof, we believe that all minimizing configurations are subsets of the rhombic lattice up to a finite number of "boundary defects"; more precisely, we think that all faces are rhombic up to (at most) one pentagonal face touching the boundary, exactly as in Fig. 7.

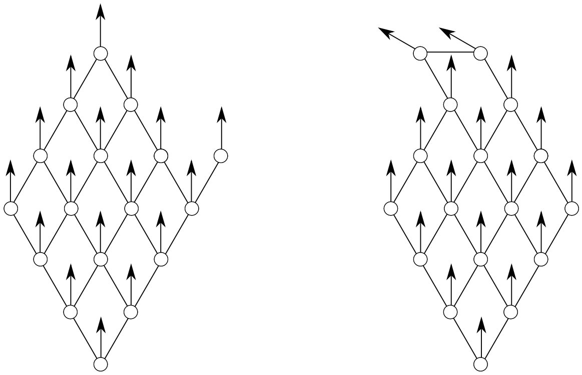


Fig. 7 Two minimizers of $\mathcal{E}_{\mathcal{V}_{\frac{\sqrt{3}}{2}}}$ in C_{17}. On the left: the minimizer (Y_N, W_N). On the right: a minimizer having a non-rhombic face and non-constant orientation

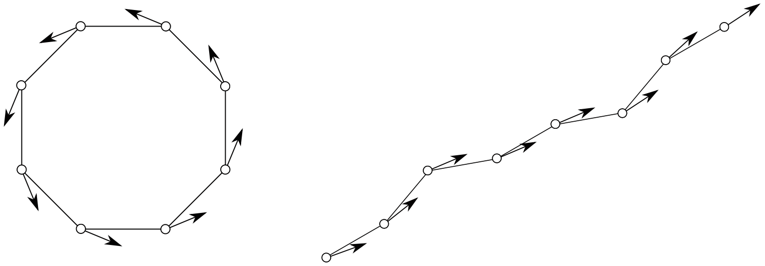


Fig. 8 On the left: a minimizer for $N \geq N_\gamma$. On the right: a minimizer for $N < N_\gamma$

Furthermore, also in this case, it is possible to prove a compactness result in the spirit of Theorem 3. Indeed, in [14, Theorem 3.9], it has been proved that, in the limit as the number of particles diverges, all configurations with perimeter-like energy bound consist of patched configurations of diamond formations with bounded perimeter.

To conclude, we discuss the case $\frac{\sqrt{3}}{2} < \gamma \leq 1$. For such a range of parameters, the maximal number of nearest neighbors appears equal to 2, and the minimizers satisfy

$$-N \leq \min_{(X,V)\in\mathcal{C}_N} \mathcal{E}_{\mathcal{V}_\gamma}(X, V) \leq -N + 1,$$

where the second inequality follows by considering the competitor $(\bar{X}, \bar{V}) \in C_N$ with $\bar{X} = ((0; 0), (1; 0), \dots, (N; 0))$ and $\bar{V} = \{(1; 0)\}^N$. For $\frac{\sqrt{3}}{2} < \gamma < 1$, we notice that setting $N_\gamma := \lceil \frac{\pi}{\arccos \gamma} \rceil$, the minimal energy is $-N$ for $N \geq N_\gamma$ and $-N + 1$ for $N < N_\gamma$. In the former case, any minimizer is made by a finite union of simple and closed polygonal curves (with suitable angles), whereas in the latter the minimizer is an open polygonal curve (see Fig. 8).

Notice that for $\gamma = 1$, the ground states of $\mathcal{E}_{\mathscr{V}^1}$ in C_N are made of N aligned points forming a segment with constant tangent orientation, while the corresponding minimal energy is equal to $-N + 1$.

Acknowledgments The author is member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

1. L. Ambrosio, N. Fusco, D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems* (Oxford University Press, Oxford, 2000)
2. Y. Au Yeung, G. Friesecke, B. Schmidt, Minimizing atomic configurations of short range pair potentials in two dimensions: crystallization in the Wulff shape. Calc. Var. Partial Differ. Equ. **44**, 81–100 (2012)
3. M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, V. Zdravkovic, Interaction ruling animal collective behavior depends on topological rather than metric distance: evidence from a field study. Proc. Nat. Acad. Sci. U.S.A. **105**, 1232–1237 (2008)
4. L. Bétermin, L. De Luca, M. Petrache, Crystallization to the square lattice for a two-body potential. Arch. Rational Mech. Anal. **240**, 987–1053 (2021)
5. X. Blanc, M. Lewin, The crystallization conjecture: a review. EMS Surv. Math. Sci. **2**, 255–306 (2015)
6. C.M. Breder, Vortices and fish schools. Zoologica **50**, 97–114 (1965)
7. C.M. Breder, Fish schools as operational structures. Fish. Bull. **74**, 471–502 (1976)
8. M. Cicalese, G. Leonardi, Maximal fluctuations on periodic lattices: an approach via quantitative wulff inequalities. Commun. Math. Phys. **375**, 1931–1944 (2020)
9. F. Cucker, S. Smale, Emergent behavior in flocks. IEEE Trans. Autom. Control **52**, 852–862 (2007)
10. E. Davoli, P. Piovano, U. Stefanelli, Sharp $N^{3/4}$ law for the minimizers of the edge-isoperimetric problem on the triangular lattice. J. Nonlinear Sci. **27**, 627–660 (2017)
11. L. De Luca, G. Friesecke, Classification of particle numbers with unique Heitmann-Radin minimizer. J. Stat. Phys. **167**, 1586–1592 (2017)
12. L. De Luca, G. Friesecke, Crystallization in two dimensions and a discrete Gauss-Bonnet theorem. J. Nonlinear Sci. **28**, 69–90 (2018)
13. L. De Luca, M. Novaga, M. Ponsiglione, Γ-convergence of the Heitmann-Radin sticky disc energy to the crystalline perimeter. J. Nonlinear Sci. **29**, 1273–1299 (2019)
14. L. De Luca, A. Ninno, M. Ponsiglione, Vectorial crystallization and collective behavior. J. Math. Biol. **84**, art. n. 6 (2022)
15. F.E. Fish, Kinematics of ducklings swimming in formation: consequence of position. J. Exper. Zool. **273**, 1–11 (1995)
16. F.E. Fish, Energetics of swimming and flying in formation. Comments Theor. Biol. **5**, 283–304 (1999)
17. L.C. Flatley, F. Theil, Face-centered cubic crystallization of atomistic configurations. Arch. Rational Mech. Anal. **218**, 363–416 (2015)
18. L.C. Flatley, A. Tarasov, M. Taylor, F. Theil, Packing twelve spherical caps to maximize tangencies. J. Comput. Appl. Math. **254**, 220–225 (2013)
19. I. Fonseca, S. Müller, A uniqueness proof for the Wulff theorem. Proc. R. Soc. Edinburgh Sect. A **119**, 125–136 (1991)

20. M. Friedrich, L. Kreutz, B. Schmidt, Emergence of rigid polycrystals from atomistic systems with Heitmann-Radin sticky disk energy. Arch. Rational Mech. Anal. **240**, 627–698 (2021)
21. C.S. Gardner, C. Radin, The infinite-volume ground state of the Lennard-Jones potential. J. Stat. Phys. **20**, 719–724 (1979)
22. H. Harborth, Lösung zu problem 664A. Elem. Math. **29**, 14–15 (1974)
23. R.C. Heitmann, C. Radin, The ground state for sticky disks. J. Stat. Phys. **22**, 281–287 (1980)
24. S. Jansen, W. König, B. Schmidt, F. Theil, Surface energy and boundary layers for a chain of atoms at low temperature. Arch. Rational Mech. Anal. **239**, 915–980 (2021)
25. S. Jansen, W. König, B. Schmidt, F. Theil, Distribution of cracks in a chain of atoms at low temperature. Ann. Henri Poincaré **22**, 4131–4172 (2021)
26. A.O. Kasumyan, D. Pavlov, Patterns and mechanisms of schooling behavior in fish: a review. J. Ichthyol. **40**(suppl. 2), S163–S231 (2000)
27. J.C. Liao, A review of fish swimming mechanics and behaviour in altered flows. Philos. Trans. R. Soc. B **362**, 1973–1993 (2007)
28. E. Mainini, U. Stefanelli, Crystallization in carbon nanostructures. Commun. Math. Phys. **328**, 545–571 (2014)
29. E. Mainini, P. Piovano, U. Stefanelli, Finite crystallization in the square lattice. Nonlinearity **27**, 717–737 (2014)
30. E. Mainini, P. Piovano, B. Schmidt, U. Stefanelli, $N^{3/4}$ law in the cubic lattice. J. Stat. Phys. **176**, 1480–1499 (2019)
31. C.W. Reynolds, Flocks, herds, and schools: a distributed behavioral model. Comp. Graph. **21**, 25–33 (1987)
32. B. Schmidt, Ground states of the 2D sticky disc model: fine properties and $N^{3/4}$ law for the deviation from the asymptotic Wulff shape. J. Stat. Phys. **153**, 727–738 (2013)
33. J.E. Taylor, Unique structure of solutions to a class of nonelliptic variational problems, in *Differential Geometry, Part 1*. Proceedings of Symposia in Pure Mathematics, vol. 27 (AMS, Providence, 1975), pp. 419–427
34. F. Theil, A proof of crystallization in two dimensions. Commun. Math. Phys. **262**, 209–236 (2006)
35. T. Vicsek, A. Czirók, E. Ben-Jacob, O. Shochet, Novel type of phase transition in a system of self-driven particles. Phys. Rev. Lett. **75**, 1226–1229 (1995)
36. D.J. Wales, Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110. Atoms. J. Phys. Chem. A **101**, 5111–5116 (1997)
37. D. Weihs, Hydromechanics of fish schooling. Nature **241**, 290–291 (1973)

Pattern Formation for Nematic Liquid Crystals—Modelling, Analysis, and Applications

Yucen Han and Apala Majumdar

1 Introduction

Liquid crystals are partially ordered materials, intermediate between conventional solid and liquid phases. They typically combine fluidity with some translational or orientational order characteristic of a solid. Liquid crystals were accidentally discovered by Friedrich Reinitzer, an Austrian plant physiologist, when he was experimenting with cholesteryl benzoate [43]. The unique physical, optical, and rheological properties of liquid crystals were gradually unveiled with time and today, and we know that liquid crystals are ubiquitous in daily life, e.g., in some clays, soap, the human DNA, cell membranes, polymers, elastomers, and the list keeps growing [26]. Liquid crystals can be classified as: nematic liquid crystals, cholesteric liquid crystals, and smectic liquid crystals [11]. The simplest phase is the nematic liquid crystal (NLC) phase, for which the constituent rod-like molecules exhibit long-range orientational order, with no positional order. In the cholesteric phase, the molecules naturally twist following a helical pattern, and one can view the NLC phase as a special cholesteric with no twist. In the smectic phase, the molecules arrange themselves in layers, there is orientational order within the layers, and the layers can slide past each other. In this review, we focus on the mathematical modelling, analysis, and simulations of NLCs in confinement, illustrating the plethora of exotic possibilities and how mathematics can be used to predict, tune, and select the properties of confined NLC systems.

NLCs are the most widely used liquid crystals perhaps because of their relative simplicity. NLC molecules are typically asymmetric in shape, e.g., rod-shaped, disc-shaped, banana-shaped, or bent core [13]. These asymmetric NLC molecules

Y. Han · A. Majumdar (✉)
Department of Mathematics and Statistics, University of Strathclyde, Strathclyde, UK
e-mail: yucen.han@strath.ac.uk; apala.majumdar@strath.ac.uk

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_18

move freely but tend to align along certain locally preferred directions, referred to as *nematic directors* in the literature [13]. Consequently, NLC phases exhibit long-range orientational order and naturally have direction-dependent responses to incident light, external electric field or magnetic fields, temperature, and mechanical stresses. Consequently, NLCs are anisotropic with directional physical properties such as the NLC dielectric anisotropy, magnetic susceptibility, and the optical refractive indices [42]. In particular, the anisotropic NLC responses to light and electric fields have made NLCs the working material of choice for the multi-billion dollar liquid crystal display (LCDs) industry [1]. NLCs have been widely used in electric billboards, TVs, laptops, calculators, and watches and a range of opto-electric devices. In recent years, there has been unprecedented interest in using NLCs for the design of new meta-materials, bio-materials, composite materials, all of which render new possibilities for sensors, photonics, actuators, artificial intelligence, and diagnostics [20, 27].

Mathematics can play a crucial role in predicting, manipulating, and even designing tailor-made NLC systems. NLC systems can be mathematically modelled at different levels, ranging from fully molecular approaches to mean-field approaches such as Onsager theory, Maier–Saupe theory [33, 40] to fully continuum approaches such as the Oseen–Frank theory, the Ericksen–Leslie theory, and the celebrated Landau–de Gennes (LdG) theory [8, 13, 13, 41]. We focus on continuum approaches wherein we do not focus on microscopic details or microscopic interactions, assuming that macroscopic properties of interest vary slowly on microscopic length scales. In the continuum approach, the NLC state is described by a macroscopic order parameter that is an averaged measure of the degree of nematic orientational order. The physically observable states are modelled by local or global minimizers of an appropriately defined free NLC energy, which typically depends on the NLC order parameter, its gradient, and various material-dependent and temperature-dependent phenomenological constants. Mathematically, this naturally raises highly non-trivial questions in the calculus of variations, singular perturbation theory, homogenization theory, and algebraic topology. The critical points of the NLC free energy are solutions (in an appropriately defined sense) of a system of nonlinear, coupled partial differential equations—the Euler–Lagrange equations, with different types of boundary conditions—Dirichlet, Neumann, Robin, etc., for the NLC order parameter. Of particular interest are the multiplicity and regularity of solutions, and how this depends on the structure of the model, the phenomenological model parameters, the symmetry of the domain, and the boundary frustration. Given that the variational problems are typically nonlinear and non-convex, there are multiple solutions of the Euler–Lagrange solutions and some energy-minimizing and some non-energy-minimizing solutions [37]. The non-minimizing solutions play a crucial role in the selection of energy minimizers and switching mechanisms in NLC systems with multiple energy minimizers. Regarding regularity, NLC defects are interpreted as a localized region of reduced NLC orientational order, which could be induced by temperature changes or by discontinuities in the nematic directors [18, 36, 37]. NLC defects are a fundamental optical signature of NLCs in confinement [23, 26]. NLC defects play a crucial role in multiplicity of solutions, the solution properties, and ultimately, structural transitions often proceed via the creation and annihilation of defects [25]. There are open mathematical questions regarding the mathematical definition of a defect and how the defect set depends on the nature of the partial order, the mathematical model, and the physical variables. Collectively, liquid crystals are a fascinating playground for mechanics, geometry, modelling, and analysis to drive a new revolution in mathematics-driven interactive materials science, for sweeping interdisciplinary and practical advances.

In this review, we focus on multistable two-dimensional NLC systems, driven by recent advances in new generations of bistable LCDs [21, 47], micropatterned surfaces [22], and also in 3D printing [12]. Multistable systems can support multiple stable nematic equilibria without any external applied fields, ideally with distinct optical and physical properties, offering multiple modes of functionality. For example, in a bistable LCD, the bright (transparent) state and the dark (opaque) states are stable without any external electric fields, so that power is only needed to switch between states or to refresh the image, but not to maintain a static image [46]. Therefore, bistable LCDs are efficient, low-cost displays with enhanced optical properties. Some of the results reviewed in this chapter are motivated by the planar bistable LC device reported in [47]. This planar device comprises a periodic array of square or rectangular NLC-filled wells, typically on the micron-scale such that the well height is much smaller than the cross-section dimensions. Hence, it is reasonable to assume that the NLC structural profile is invariant along the height of the well, and it suffices to model planar profiles in the well square cross-section. The well surfaces are treated to induce tangential or planar anchoring, so that the NLC molecules lie in the plane of the well surfaces and tangent to the well edges. There is a natural mismatch in the nematic directors at the square vertices, leading to interesting and multiple possibilities for stable NLC configurations.

Indeed, this relatively simple geometry is actually experimentally reported to be bistable [47]. There are at least two experimentally reported stable states—the *diagonal* state for the nematic director is roughly along a square diagonal, and the *rotated* state for which the nematic director rotates by 180° between a pair of opposite edges [12]. Both states have long-term stability and somewhat contrasting optical properties, without external electric fields. We use this example of a multistable system as a benchmark example, and in this review, we address natural questions such as—what happens if we replace the square with a regular or asymmetric 2D polygon, what are the effects of material anisotropy on multistability, and crucially, can we mathematically model solution landscapes in reduced 2D frameworks and study the connectivity of non-energy-minimizing solutions to energy-minimizing solutions? The transition pathways between the diagonal and rotated states have been studied in [25], but a systematic study of the non-energy-minimizing critical points is largely open.

The review paper is organized as follows. The LdG theory is reviewed in Sect. 2. In Sect. 3, we review some known results on NLC solution landscapes for square domains as a benchmark example. In Sect. 4, we summarize the results in [16] to illustrate multistability for NLCs in 2D polygons and the effects of geometry. In Sects. 5 and 6, we summarize the results reported in [10] and [18] to elucidate the effects of geometrical asymmetry (by taking the rectangle as an example) and the effects of elastic anisotropy on NLC solution landscapes on square domains. The last leg of the review concerns unstable saddle points (non-minimizing solutions) and transition pathways for NLCs, as reported in [19] in Sect. 7. Some conclusions are discussed in Sect. 8. In the supplement, Sect. 9, we summarize the numerical methods used for solving the complex system of LdG Euler–Lagrange equations and for computing the non-trivial NLC solution landscapes.

2 The Landau–de Gennes Theory

The Landau–de Gennes (LdG) theory is perhaps the most powerful continuum theory for NLCs [13, 30, 48]. In 1991, Pierre-Gilles de Gennes was awarded the Nobel prize in Physics for discovering that "methods developed for studying order phenomena in simple systems can be generalized to more complex forms of matter, in particular to liquid crystals and polymers." The LdG model describes the NLC state by a macroscopic order parameter—the LdG **Q**-tensor, which is a macroscopic measure of NLC orientational order, i.e., the deviation of the ordered nematic phase from the isotropic disordered phase. Mathematically, the **Q**-tensor is a symmetric traceless 3×3 matrix. The **Q**-tensor has five degrees of freedom and can be written as [39]

$$\mathbf{Q} = \lambda_1 \mathbf{n} \otimes \mathbf{n} + \lambda_2 \mathbf{m} \otimes \mathbf{m} + \lambda_3 \mathbf{p} \otimes \mathbf{p}, \tag{1}$$

where **n**, **m**, and **p** are eigenvectors of **Q** that model the nematic directors, and λ_i, $i = 1, 2, 3$, are the corresponding eigenvalues, which measure the degree of orientational order about these directors. In particular, $\sum_{i=1}^{3} \lambda_i = 0$, from the tracelessness constraint. A **Q**-tensor is said to be (i) isotropic if $\mathbf{Q} = 0$, i.e., $(\lambda_1, \lambda_2, \lambda_3) = (0, 0, 0)$, (ii) uniaxial if **Q** has a pair of degenerate non-zero eigenvalues, $(\lambda, \lambda, -2\lambda)$, and (iii) biaxial if **Q** has three distinct eigenvalues [13]. A uniaxial **Q**-tensor can be written as

$$\mathbf{Q} = s\,(\mathbf{n} \otimes \mathbf{n} - \mathbf{I}/3) \tag{2}$$

with **I** being the 3×3 identity matrix, $s = -3\lambda$ is real, and $\mathbf{n} \in \mathbb{S}^2$, a unit vector. The vector, **n**, is the eigenvector with the non-degenerate eigenvalue, known as the "director" and models the single preferred direction of uniaxial nematic alignment at every point in space [13, 48]. The scalar, s, is the scalar order parameter, which measures the degree of orientational order about **n**. In the biaxial case, there are primary and secondary nematic directors, with two scalar order parameters.

In the absence of surface energies, the LdG energy is given by

$$I_{LdG}[\mathbf{Q}] := \int f_{el}(\mathbf{Q}, \nabla\mathbf{Q}) + f_b\,(\mathbf{Q})\,\mathrm{d}\mathbf{x}, \tag{3}$$

where f_{el} and f_b are the elastic and thermotropic bulk energy densities, respectively. The elastic energy density is typically quadratic and convex in $\nabla\mathbf{Q}$ and penalises spatial inhomogeneities. In general, the elastic energy density has different contributions from different deformation modes, e.g., splay, twist, and bend [13]. A commonly used version is

$$f_{el}(\mathbf{Q}) = \frac{L_1}{2} Q_{ij,k} Q_{ij,k} + \frac{L_2}{2} Q_{ij,j} Q_{ik,k} + \frac{L_3}{2} Q_{ik,j} Q_{ij,k}, \tag{4}$$

where L_1, L_2, L_3 are the material elastic constants, subject to certain constraints to ensure $f_{el}(\mathbf{Q}) \geq 0$. Since

$$Q_{ij,j} Q_{ik,k} - Q_{ik,j} Q_{ij,k} = (Q_{ij} Q_{ik,k})_{,j} - (Q_{ij} Q_{ik,j})_{,k} \tag{5}$$

is a null Lagrangian, we can ignore the L_3-term with Dirichlet boundary conditions. Hence, the elastic energy density in (4) is reduced to a two-term elastic energy density, as shown below

$$f_{el}(\mathbf{Q}) = \frac{L}{2} \left(|\nabla\mathbf{Q}|^2 + \hat{L}_2 (\mathrm{div}\mathbf{Q})^2 \right), \tag{6}$$

where $\hat{L}_2 \in (-1, \infty)$ is the "elastic anisotropy" parameter. The elastic anisotropy can be strong for polymeric materials [50]. In Sect. 6, we study the effects of elastic anisotropy on NLC solution landscapes on square domains.

In Sects. 4, 5, and 7, we use the one-constant approximation, for which, $L_2 = L_3 = 0$ in (4), i.e., $\hat{L}_2 = 0$ in (6), so that the elastic energy density simply reduces to the Dirichlet energy density $|\nabla\mathbf{Q}|^2$. The one-constant approximation assumes that all deformation modes have comparable energetic penalties, i.e., equal elastic constants, and this is a good approximation for some characteristic NLC materials such as MBBA [13, 48], which makes the mathematical analysis more tractable.

The bulk energy density f_b is a polynomial of the eigenvalues of order parameter $\mathbf{Q}$ and drives the isotropic–nematic phase transition as a function of the temperature [13, 39]. We work with the simplest form of f_b, a quartic polynomial of eigenvalues of $\mathbf{Q}$-tensor:

$$f_b(\mathbf{Q}) := \frac{A}{2} \mathrm{tr}\mathbf{Q}^2 - \frac{B}{3} \mathrm{tr}\mathbf{Q}^3 + \frac{C}{4} (\mathrm{tr}\mathbf{Q}^2)^2, \tag{7}$$

where $\mathrm{tr}\mathbf{Q}^2 = Q_{ij} Q_{ij} = \lambda_i^2$, and $\mathrm{tr}\mathbf{Q}^3 = Q_{ij} Q_{jk} Q_{ki} = \lambda_i^3$, for $i, j, k = 1, 2, 3$. The variable $A = \alpha(T - T^*)$ is a rescaled temperature, α, L, B, $C > 0$ are material-dependent constants, and T^* is the characteristic nematic supercooling temperature. The rescaled temperature A has three characteristic values: (i) $A = 0$, below which the isotropic phase $\mathbf{Q} = 0$ loses stability, (ii) the nematic–isotropic transition temperature, $A = B^2/27C$, at which f_b is minimized by the isotropic phase and a continuum of uniaxial states with $s = s_+ = B/3C$ and $\mathbf{n}$ arbitrary in

(2), and (iii) the nematic superheating temperature, $A = B^2/24C$, above which the isotropic state is the unique critical point of f_b.

As Proposition 1 in [36], for a given $A < 0$, the set of minima of the bulk potential is

$$\mathscr{N} := \{\mathbf{Q} \in S_0 : \mathbf{Q} = s_+ (\mathbf{n} \otimes \mathbf{n} - \mathbf{I}/3)\}, \tag{8}$$

where

$$s_+ := \frac{B + \sqrt{B^2 + 24|A|C}}{4C}$$

and $\mathbf{n} \in S^2$ arbitrary. In particular, this set is relevant to our choice of Dirichlet conditions for boundary-value problems in subsequent sections. The size of defect cores is typically inversely proportional to s_+ for low temperatures $A < 0$. Following [51], we use MBBA as a representative NLC material and use its reported values for B and C to fix $B = 0.64 \times 10^4\,\mathrm{N/m^2}$ and $C = 0.35 \times 10^4\,\mathrm{N/m^2}$ throughout this review. We also frequently use the fixed temperature, $A = -B^2/3C$ for numerical simulations, although the qualitative conclusions remain unchanged for $A < 0$.

Boundary effects are a crucial consideration for NLCs in confinement and dictate multistability to some extent. There are multiple mathematical choices for the boundary conditions. The simplest approach is Dirichlet boundary conditions or fixed boundary conditions for the LdG **Q**-tensor order parameter. This fixes the nematic directors and the scalar order parameters on the boundary. Typically, we impose tangential or homeotropic boundary condition, which means the nematic director is tangent or normal to the domain boundary. On domains with sharp corners, some care is needed to deal with the mismatch in the nematic director at the corners. This could involve truncating the geometry or imposing a low-order point at the sharp corners. Dirichlet conditions are mathematically more tractable, but weak anchoring is more realistic, with surface energies, and the resulting boundary conditions typically involve the normal derivatives of **Q** on the boundary. A popular surface energy, known as the Rapini–Papoular energy, is [48]

$$E_s[\mathbf{Q}] = \int_{\partial} W \mathrm{tr}(\mathbf{Q} - \mathbf{Q}_s)^2 dA, \tag{9}$$

where W is the surface anchoring strength and $\mathbf{Q}_s$ is the preferred LdG **Q**-tensor on the boundary. As $W \to \infty$, we qualitatively recover the Dirichlet condition $\mathbf{Q} = \mathbf{Q}_s$ on the boundary. Interested readers are referred to [32].

We model nematic profiles inside three-dimensional wells

$$\mathscr{B} = \Omega \times [0, h], \tag{10}$$

whose cross-section is a two-dimensional polygon Ω and h is the well height. The two-dimensional working domain Ω is any regular polygon in Sect. 4, a rectangle in Sect. 5, a square in Sect. 6, and a regular hexagon in Sect. 7. In the thin-film limit, i.e., $h \to 0$ limit and imposing surface energies, f_s, on the top and bottom surfaces, which favour planar degenerate boundary conditions or equivalently constrain the nematic directors to be in the plane of the cross-section without a fixed direction and require $\mathbf{z} = (0, 0, 1)$ to be a fixed eigenvector of the corresponding $\mathbf{Q}$-tensor, we can rigorously justify the reduction from the three-dimensional domain $\mathscr{B}$ to the two-dimensional domain Ω [14]. If we impose a Dirichlet boundary condition, $\mathbf{Q}_b$, which has the unit vector, $\mathbf{z} = (0, 0, 1)$ as a fixed eigenvector, on the lateral surfaces, $\partial\Omega \times [0, h]$, then one can show that in the $\frac{h}{\lambda} \to 0$ limit, where λ^2 is a measure of the cross-section size, minima of the LdG energy (3) converge (weakly in H^1) to minima of the reduced functional

$$F_0[\mathbf{Q}] := \int_\Omega \frac{1}{2}\left(\left|\nabla_{x,y}\mathbf{Q}\right|^2 + \hat{L}_2\left(\mathrm{div}_{x,y}\mathbf{Q}\right)^2\right) + \frac{\lambda^2}{L} f_b(\mathbf{Q})\,\mathrm{dA} \tag{11}$$

subject to the boundary condition $\mathbf{Q} = \mathbf{Q}_b$ on $\partial\Omega$ and to the constraint that $\mathbf{z}$ is an eigenvector of $\mathbf{Q}(x,\ y)$ for any $(x,\ y) \in \Omega$. Using the reasoning above, we restrict ourselves to $\mathbf{Q}$-tensors with $\mathbf{z}$ as a fixed eigenvector and study critical points or minima of (11) with three degrees of freedom as

$$\begin{aligned} \mathbf{Q}(x, y) = {} & q_1(x, y)\left(\hat{\mathbf{x}} \otimes \hat{\mathbf{x}} - \hat{\mathbf{y}} \otimes \hat{\mathbf{y}}\right) + q_2(x, y)\left(\hat{\mathbf{x}} \otimes \hat{\mathbf{y}} + \hat{\mathbf{y}} \otimes \hat{\mathbf{x}}\right) \\ & + q_3(x, y)\left(2\hat{\mathbf{z}} \otimes \hat{\mathbf{z}} - \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} - \hat{\mathbf{y}} \otimes \hat{\mathbf{y}}\right), \end{aligned} \tag{12}$$

where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are the unit coordinate vectors in the x, y, and z directions, respectively. Informally speaking, q_1 and q_2 measure the degree of "in-plane" order, q_3 measures the "out-of-plane" order, and $\mathbf{Q}$ is invariant in the z direction. This constraint naturally excludes certain solutions such as the stable escaped radial with ring defect solution in a cylinder with large radius in [15], for which the z-invariance does not hold or critical points that exhibit "escape into the third dimension" [45], for which $\hat{\mathbf{z}}$ is not a fixed eigenvector for $\mathbf{Q}$. While we present our results in a 2D framework in the case studies, these reduced critical points survive for all $h > 0$ (beyond the thin-film limit) although they may not be physically relevant or energy-minimizing outside the thin-film limit ([6] and [49]).

3 Benchmark Example

The square domain is a very well-studied domain, and we review some classical results in this section. In [24] and [6], the authors report the Well Order Reconstruction Solution ($WORS$) on a square domain, for all square edge lengths $\lambda > 0$, without elastic anisotropy, for Dirichlet tangent boundary conditions. The $WORS$

has a constant set of eigenvectors, $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$, which are the coordinate unit vectors. The $WORS$ is further distinguished by a uniaxial cross, with negative scalar order parameter, along the square diagonals. Physically, this implies that there is a planar defect cross along the square diagonals, and the nematic molecules are disordered along the square diagonals. This defect cross partitions the square domain into four quadrants, and the nematic director is constant in each quadrant. The defect cross is an interesting example of a negatively ordered uniaxial interface that separates distinct polydomains. In [6], the authors analyze this system at a fixed temperature $A = -B^2/3C$ and show that the $WORS$ is a classical solution of the associated Euler–Lagrange (EL) equations for the LdG free energy, of the form:

$$\mathbf{Q}_{WORS}(x, y) = q(\hat{\mathbf{x}} \otimes \hat{\mathbf{x}} - \hat{\mathbf{y}} \otimes \hat{\mathbf{y}}) - \frac{B}{6C}(2\hat{\mathbf{z}} \otimes \hat{\mathbf{z}} - \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} - \hat{\mathbf{y}} \otimes \hat{\mathbf{y}}). \tag{13}$$

There is a single degree of freedom, $q : \Omega \to \mathbb{R}$, which satisfies the Allen–Cahn equation

$$\Delta q = \frac{\lambda^2}{L}(2Cq^3 - \frac{B^2}{2C}q) \tag{14}$$

and exhibits the following symmetry properties:

$$q = 0 \quad \text{on} \quad \{y = x\} \cup \{y = -x\}, \qquad (y^2 - x^2)q(x, y) \geq 0. \tag{15}$$

Mathematically speaking, this implies that the $\mathbf{Q}_{WORS}$ is strictly uniaxial with negative order parameter along the square diagonals that would manifest as a pair of orthogonal defect lines in experiments. They also prove that the $WORS$ is globally stable for λ small enough, i.e., nano-scale domains, and becomes unstable as λ increases, demonstrating a pitchfork bifurcation in a scalar setting. Numerical experiments suggest that the $WORS$ acts as a transition state between energy minimizers for large λ. For large square domains (on micron-scale or larger), there are two competing stable physically observable states: the largely uniaxial diagonal states (D), for which the nematic director (in the plane) is aligned along one of the square diagonals, and the rotated states (R) for which the director rotates by π radians between a pair of opposite square edges. On a square domain, there are 2 rotationally equivalent D states and 4 rotationally equivalent R states [25, 29]. We note that the D and R states have non-zero q_2 in (12), while the $WORS$ has $q_2 = 0$ everywhere. In other words, the $WORS$ solution has constant eigenvectors everywhere, whereas the D and R solutions have varying eigenvectors in the plane of the square domain.

The bifurcation diagram for this model problem has been documented in [44] (see Fig. 1). For $\lambda < \lambda^*$, there is the unique $WORS$. For $\lambda = \lambda^*$, the stable $WORS$ bifurcates into an unstable $WORS$ and two stable D solutions. The WORS exists for all values of λ. When $\lambda = \lambda^{**} > \lambda^*$, the unstable $WORS$ bifurcates into two unstable BD solutions, which are featured by defect lines localized near a

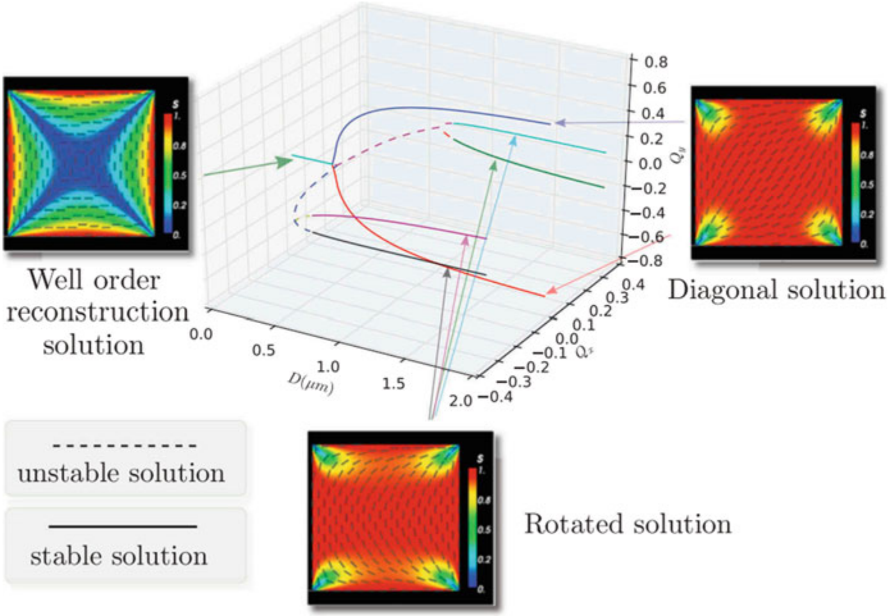


Fig. 1 The bifurcation diagram on square without elastic anisotropy. Reproduced from [44] with permission from Taylor&Francis

pair of opposite square edges. The two BD solution branches are represented by the dashed lines in Fig. 1. Each unstable BD solution further bifurcates into two unstable R solutions, which gain stability as λ further increases. The $WORS$ has the highest energy among the numerically computed solutions, for all λ. For the numerical results presented in [44], the authors have $D^* \approx 0.36$ and $D^{**} \approx 0.38$ in Fig. 1, and correspondingly, we have $2C\lambda^{*2}/L \approx 7$ and $2C\lambda^{**2}/L \approx 10$.

4 Nematic Equilibria on 2D Polygons

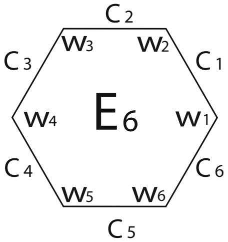
This section reviews results from a recent paper [16], where the authors study multistability for NLCs in regular 2D polygons, with tangent boundary conditions, with emphasis on the effects of geometry captured by the polygon edge length, λ.

The working domain, Ω, is a regular rescaled polygon, E_K, with K edges, centred at the origin with vertices

$$w_k = \left(\cos\left(2\pi\left(k-1\right)/K\right), \sin\left(2\pi\left(k-1\right)/K\right)\right), \; k = 1, \ldots, K.$$

We label the edges counterclockwise as $C_1, \ldots, C_K$, starting from $(1, 0)$. For example, E_6 is a regular hexagon shown in Fig. 2 and E_4 is a square.

Fig. 2 The regular rescaled hexagon domain E_6. Reproduced from [16] with permission from Society for Industrial and Applied Mathematics



As elaborated in Sect. 2, it is reasonable to work with in a reduced LdG framework, with **Q**-tensors of the form in (12), on 2D polygons. From [7], for the special temperature $A = -B^2/3C$, we necessarily have $q_3 = -\frac{B}{6C}$, for all $\lambda > 0$. For arbitrary $A < 0$, we would have non-constant q_3 profiles, and while we conjecture that some qualitative solution properties are universal for $A < 0$, a non-constant q_3 profile would introduce new technical difficulties. For $A = -B^2/3C$ and with constant q_3, the **Q**-tensor in (12) reduces to a symmetric, traceless 2×2 matrix, **P**, as given below

$$\mathbf{P} = \begin{pmatrix} P_{11} & P_{12} \\ P_{12} & -P_{11} \end{pmatrix}.$$

The relation between the LdG order parameter **Q**-tensor and the reduced **P**-tensor is

$$\mathbf{Q} = \left(\begin{array}{cc|c} \multicolumn{2}{c|}{\multirow{2}{*}{$\mathbf{P}(\mathbf{r}) + \frac{B}{6C}\mathbf{I}_2$}} & 0 \\ & & 0 \\ \hline 0 & 0 & -B/3C \end{array}\right). \tag{16}$$

Therefore, the energy in (11) is reduced to

$$F[\mathbf{P}] := \int_\Omega \frac{1}{2}|\nabla\mathbf{P}|^2 + \frac{\lambda^2}{L}\left(-\frac{B^2}{4C}tr\mathbf{P}^2 + \frac{C}{4}\left(tr\mathbf{P}^2\right)^2\right)\mathrm{dA}, \tag{17}$$

and the corresponding EL equations are

$$\begin{aligned} \Delta P_{11} &= \frac{2C\lambda^2}{L}\left(P_{11}^2 + P_{12}^2 - \frac{B^2}{4C^2}\right)P_{11}, \\ \Delta P_{12} &= \frac{2C\lambda^2}{L}\left(P_{11}^2 + P_{12}^2 - \frac{B^2}{4C^2}\right)P_{12}. \end{aligned} \tag{18}$$

We can also write **P** in terms of an order parameter s and an angle γ as shown below

$$\mathbf{P} = 2s\left(\mathbf{n}\otimes\mathbf{n} - \frac{1}{2}\mathbf{I}_2\right); \quad \mathbf{n} = (\cos\gamma, \sin\gamma)^T, \tag{19}$$

where $\mathbf{I}_2$ is the 2×2 identity matrix, so that $P_{11} = s\cos(2\gamma)$, $P_{12} = s\sin(2\gamma)$. The nodal set, defined by the zeroes of $\mathbf{P}$, models the planar defects in Ω, i.e., when $\mathbf{P} = 0$, $s = 0$ in (19) so that there is no nematic order in the plane of Ω, and the eigenvalues of the corresponding $\mathbf{Q}$ are $(B/6C, B/6C, -B/3C)$. In other words, the nodal set of $\mathbf{P}$ defines a uniaxial set of $\mathbf{Q}$ with negative order parameter and will have a distinct optical signature in experiments.

Next, we specify Dirichlet tangent boundary conditions for $\mathbf{P}$ on ∂E_K, labelled by $\mathbf{P}_b$. The tangent boundary conditions require $\mathbf{n}$ in (19) to be tangent to the edges of E_K, and $s = s_+ = B/3C$. However, there is a necessary mismatch at the vertices, so that we fix the value of $\mathbf{P}$ at the vertex, to be the average of the two constant values, on the two intersecting edges. On a $d \ll \frac{1}{2}$-neighbourhood of the vertices, we linearly interpolate between the constant values on the edge and the average value at the vertex. For d sufficiently small, the choice of the interpolation does not change the qualitative solution profiles. This means that the tangent conditions are not necessarily respected in the d-neighbourhood of vertices.

In what follows, we study the minima of (17) in two distinguished limits analytically—the $\lambda \to 0$ limit is relevant for nano-scale domains and the $\lambda \to \infty$ limit, which is the macroscopic limit relevant for micron-scale or larger cross-sections, Ω. We present rigorous results for limiting problems below, but our numerical simulations show that the limiting results are valid for non-zero but sufficiently small λ (or even experimentally accessible nano-scale geometries depending on parameter values) and sufficiently large but finite λ too. In other words, these limiting results are of potential practical value too.

In the $\lambda \to 0$ limit, using methods from [4] and from Proposition 3.1 of [10], we can show that minima of (17), subject to the Dirichlet tangent boundary conditions (for d sufficiently small), converge uniformly to the unique solution of the following limiting problem for $\lambda = 0$,

$$\begin{aligned} &\Delta P_{11}^0 = 0,\ \Delta P_{12}^0 = 0, \mathit{on}\ E_K, \\ &P_{11}^0 = P_{11b},\ P_{12}^0 = P_{12b},\ \mathit{on}\ \partial E_K. \end{aligned} \tag{20}$$

The solution of Laplace equation on disc can be explicitly solved. Our strategy to solve the Dirichlet boundary-value problem (20) on polygon E_K is to map it to an associated Dirichlet boundary-value problem on the unit disc in Fig. 3, by using the Schwarz–Christoffel mapping [5]. The SC mapping from a unit disc to a regular polygon E_K is

$$f(z) = C_1(K) \int_0^z \frac{1}{\left(1 - x^K\right)^{2/K}} \mathrm{d}x$$

with

$$C_1(K) = \frac{\Gamma(1 - 1/K)}{\Gamma(1 + 1/K)\,\Gamma(1 - 2/K)}.$$

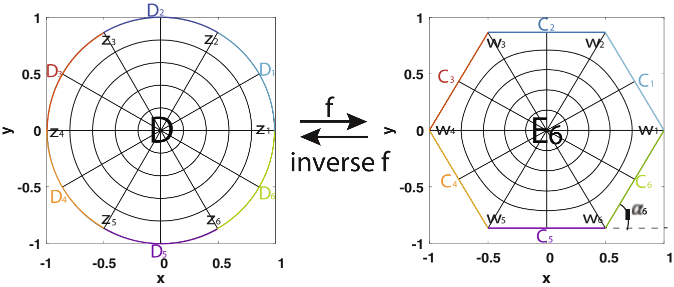


Fig. 3 Schwarz–Christoffel mapping f from a unit disc to a regular hexagon and inverse mapping f^{-1} from a regular hexagon to a unit disc. Reproduced from [16] with permission from Society for Industrial and Applied Mathematics

Using the symmetries of the boundary condition and the regular polygon, we can prove the symmetry properties of the limiting solution of (20) accompanied by rigorous results for the corresponding nodal set, as given below (from [16]).

Proposition 1 *Let* (P_{11}, P_{12}) *be the unique solution of (20), and let*

$$G_K := \{S \in O(2) : SE_K \subseteq E_K\} \tag{21}$$

be a set of symmetries consisting of K *rotations by angles* $2\pi k/K$ *for* $k = 1, \ldots, K$ *and* K *reflections about the symmetry axes* ($\phi = \pi k/K$, $k = 1, \ldots, K$) *of the polygon* E_K*.* $P_{11}^2 + P_{12}^2$ *is invariant under* G_K*. If* $(P_{11}, P_{12}) \neq (0, 0)$*, then* $\frac{(P_{11}, P_{12})}{\sqrt{P_{11}^2+P_{12}^2}}$ *undergoes a reflection about the symmetry axes of the polygon and rotates by* $4\pi k/K$ *under rotations of angle* $2\pi k/K$ *for* $k = 1, \ldots, K$*.*

Proposition 2 *Let* $\mathbf{P}_R = (P_{11}, P_{12})$ *be the unique solution of the boundary-value problem (20). Then* $P_{11}(0, 0) = 0$, $P_{12}(0, 0) = 0$ *at the centre of all regular polygons,* E_K*. However,* $\mathbf{P}_R(x, y) \neq (0, 0)$ *for* $(x, y) \neq (0, 0)$*, for all* E_K *with* $K \neq 4$*, i.e., the* $WORS$ *is a special case of* $\mathbf{P}_R$ *on* E_4 *such that* $\mathbf{P}_R = (0, 0)$ *on the square diagonals.*

For $K \neq 4$, $\mathbf{P}_R$ has a unique isotropic point at the origin and is referred to as the $Ring$ solution, since for $K > 4$, the director profile (the profile of the leading eigenvector of $\mathbf{P}_R$ with the largest positive eigenvalue) exhibits a $+1$-vortex at the centre of the polygon. In Fig. 4, we numerically plot the ring configuration for a triangle, pentagon, hexagon, and a disc ($K \to \infty$) and the $WORS$ for the square. For $K = 3$, the isotropic point at the centre of the equilateral triangle resembles a $-1/2$ point defect. This is a very interesting example of the effect of geometry on solutions and their defect sets.

The $\lambda \to \infty$ limit is analogous to the "Oseen–Frank limit" in [37]. Let $\mathbf{P}^\lambda$ be a global minimizer of (17), subject to a fixed boundary condition $\mathbf{P}_b = (P_{11b}, P_{12b})$ on ∂E_K. As $\lambda \to \infty$, the minima, $\mathbf{P}^\lambda$, are well approximated by $\mathbf{P}^\infty$, at least

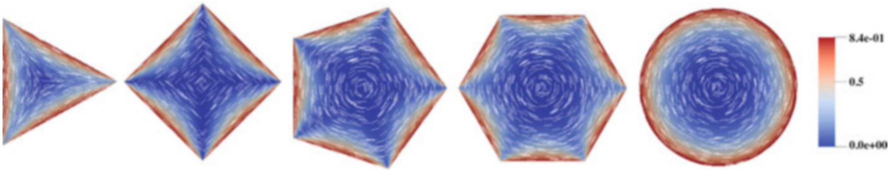


Fig. 4 Solutions $\left(P^0_{11}, P^0_{12}\right)$ of (20) when $K = 3, 4, 5, 6$ in regular triangle, square, pentagon, hexagon domain and $K \to \infty$ in disc domain. The vector $\left(\cos\left(\arctan\left(P^0_{12}/P^0_{11}\right)/2\right), \sin\left(\arctan\left(P^0_{12}/P^0_{11}\right)/2\right)\right)$ is represented by white lines, and the order parameter $\left(s^0\right)^2 = \left(P^0_{11}\right)^2 + \left(P^0_{12}\right)^2$ is represented by colour from blue to red. Reproduced from [16] with permission from Society for Industrial and Applied Mathematics

everywhere away from the vertices, where

$$\mathbf{P}^{\infty} = \frac{B}{2C}\left(\mathbf{n}^{\infty} \otimes \mathbf{n}^{\infty} - \frac{1}{2}\mathbf{I}_2\right),$$

$\mathbf{n}^{\infty} = (\cos\gamma^{\infty}, \sin\gamma^{\infty})$ and γ^{∞} is a global minimizer of the energy

$$I[\gamma] := \int_{E_K} |\nabla\gamma|^2 \, \mathrm{dA}$$

subject to Dirichlet conditions, $\gamma = \gamma_b$ on ∂E_K. The angle γ_b is determined by the fixed boundary condition, $\mathbf{P}_b$, where $\mathbf{n}_b = (\cos\gamma_b, \sin\gamma_b)$. We have $\mathbf{n}_b$ is tangent to the polygon edges, which constrains the values of γ_b, and if $\deg(\mathbf{n}_b, \partial E_K) = 0$ (interpreted as the winding number of $\mathbf{n}_b$ around ∂E_K), then γ^{∞} is a solution of the Laplace equation

$$\Delta\gamma^{\infty} = 0, \; on \; E_K \tag{22}$$

subject to $\gamma = \gamma_b$ on ∂E_K [3, 29]. The solution γ^{∞} may be explicitly computed for a given γ_b, but the tangent boundary conditions necessarily imply that γ_b is discontinuous at the polygon vertices, and hence, $I[\gamma^{\infty}]$ is infinite. This is consistent with a vanishingly small interpolation region around the vertices as described above, i.e., taking the $d \to 0$ limit above, and the structural details of the equilibria are largely unaffected by how we define the boundary conditions at the vertices, provided the size of the interpolation region is sufficiently small.

There are multiple choices of γ_b consistent with the tangent boundary conditions, which implies that there are multiple local/global minima of (17) for large λ. We present a simple estimate of the number of stable states if we restrict γ_b, so that $\mathbf{n}_b$ rotates by either $2\pi/K - \pi$ or $2\pi/K$ at a vertex (see Fig. 5a, b, referred to as "splay" and "bend" vertices, respectively). Since we require $\deg(\mathbf{n}_b, \partial E_K) = 0$, we necessarily have 2 "splay" vertices and $(K - 2)$ "bend" vertices. So we have at least $\binom{K}{2}$ minima of (17), for λ sufficiently large.

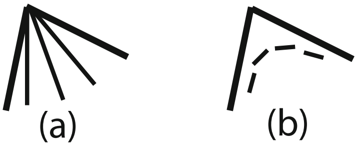


Fig. 5 Two arrangements of nematics in the corner: (**a**) splay and (**b**) bend. Reproduced from [16] with permission from Society for Industrial and Applied Mathematics

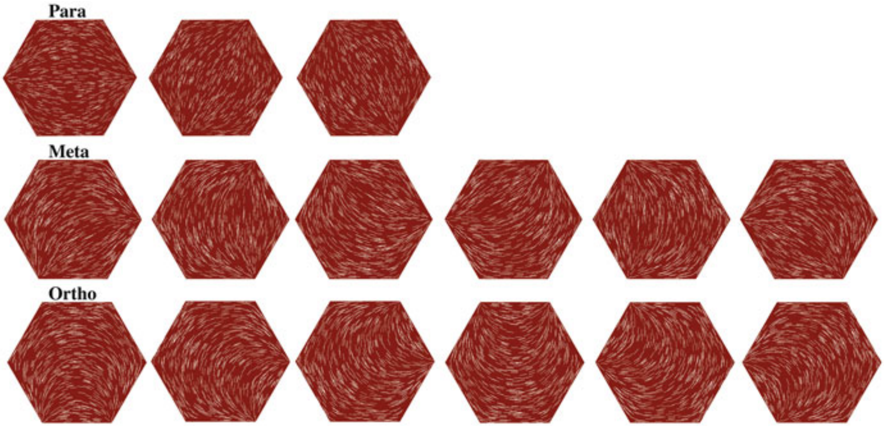


Fig. 6 $\binom{6}{2=15}$ solutions of (22) subject to boundary condition (23) in hexagon domain. The vector $(\cos\gamma^{\infty}, \sin\gamma^{\infty})$ is represented by white lines. Reproduced from [16] with permission from Society for Industrial and Applied Mathematics

As an illustrative example, we take the hexagon E_6 in Fig. 6. The Dirichlet boundary conditions are

$$\gamma_b = \gamma_k \; on \; C_k, \;\; k = 1, \ldots, K, \tag{23}$$

where

$$\gamma_1 = \frac{\pi}{K} - \frac{\pi}{2}, \;\; \gamma_{k+1} = \gamma_k + jump_k, \;\; k = 1, 2, .., K-1.$$

We need to choose the two splay vertices where γ rotates as in Fig. 5a. If the chosen corner is between the edges C_k and C_{k+1}, then $jump_k = 2\pi/K - \pi$; otherwise $jump_k = 2\pi/K$, $k = 1, \ldots, K-1$. We have 15 different choices for the two "splay" vertices, (i) 3 of which correspond to the three pairs of diagonally opposite vertices, (ii) 6 of which correspond to pairs of vertices that are separated by one vertex, and (iii) 6 of which correspond to "adjacent" vertices connected by an edge (see Fig. 6). We refer to (i) as *Para* states, (ii) as *Meta* states, and (iii) as *Ortho* states.

Next, we present two bifurcation diagrams on a hexagon and pentagon as a function of λ, as illustrative examples of a polygon with an even or odd number of edges. We discuss the bifurcation diagram on E_6 in Fig. 7a. For λ sufficiently small, there is a unique $Ring$-like minimizer. Our numerics show that the $Ring$-like solution (with the unique zero at the polygon centre) exists for all λ, but there is a critical point $\lambda = \lambda^*$, such that the $Ring$-like solution is unstable for $\lambda > \lambda^*$ and bifurcates into two kinds of branches: stable $Para$ solution branches and unstable BD branches. In the BD state, the hexagon is separated into three regions by two "defective low-order lines" (low $|\mathbf{P}|^2$) such that the corresponding director (eigenvector with largest positive eigenvalue) is approximately constant in each region. There are at least three different BD states. The unstable BD branches further bifurcate into unstable $Meta$ solutions at $\lambda = \lambda^{**}$. There is a further critical point $\lambda = \lambda^{***}$ at which the $Meta$ solutions gain stability and continue as stable solution branches as λ increases. Stable $Ortho$ solutions appear as solution branches for λ is large enough. For large λ, there are multiple stable solutions: three $Para$, six $Meta$, and six $Ortho$, in Fig. 6. The $Para$ states have the lowest energy, and the $Ortho$ states are energetically the most expensive, as can be explained on the heuristic grounds that bending between neighbouring vertices is energetically unfavourable. The case of a pentagon is different. In Fig. 7b, there is no analogue of the $Para$ states, and there are 10 different stable states for large λ—(1) five $Meta$ states featured by a pair of splay vertices that are separated by a vertex and (2) five $Ortho$ states featured by a pair of adjacent splay vertices. There are five analogues of the BD states that are featured by a single line of "low" order along an edge and an opposite splay vertex.

These examples and the numerical results are not exhaustive, but they do showcase the beautiful complexity and ordering transitions feasible in two-dimensional polygonal frameworks. Similar methodologies can also be applied to other non-regular polygons, convex or concave polygons.

5 Effects of Geometrical Anisotropy

The prototype problem of nematics inside square and other regular polygon domains has been discussed in Sect. 3 and 4. A natural question is what will happen if we break the symmetry of geometry? For example, what are the solution landscapes for NLCs on two-dimensional rectangles, as opposed to squares, and how sensitive is the landscape to the geometrical anisotropy? Is there the counterpart of $WORS$ on a rectangle?

We review results from [10]. The working domain is $\Omega = [0, a] \times [0, 1]$, with $a > 1$, and let ϵ be a dimensionless parameter that is inversely proportional to λ^2 in the reduced LdG free energy (17) in Sect. 4. We use a combination of formal calculations and elegant maximum principle arguments to analyze solution landscapes in the $\epsilon \to \infty$ and $\epsilon \to 0$ limits.

In the $\epsilon \to \infty$ limit, i.e., the $\lambda \to 0$ limit, the limiting problem is a system of Laplace equations with Dirichlet tangent boundary conditions. Analogous to the calculations in [29], the unique solution $\mathbf{P}^0$ can be calculated explicitly as $P_{12}^0 = 0$

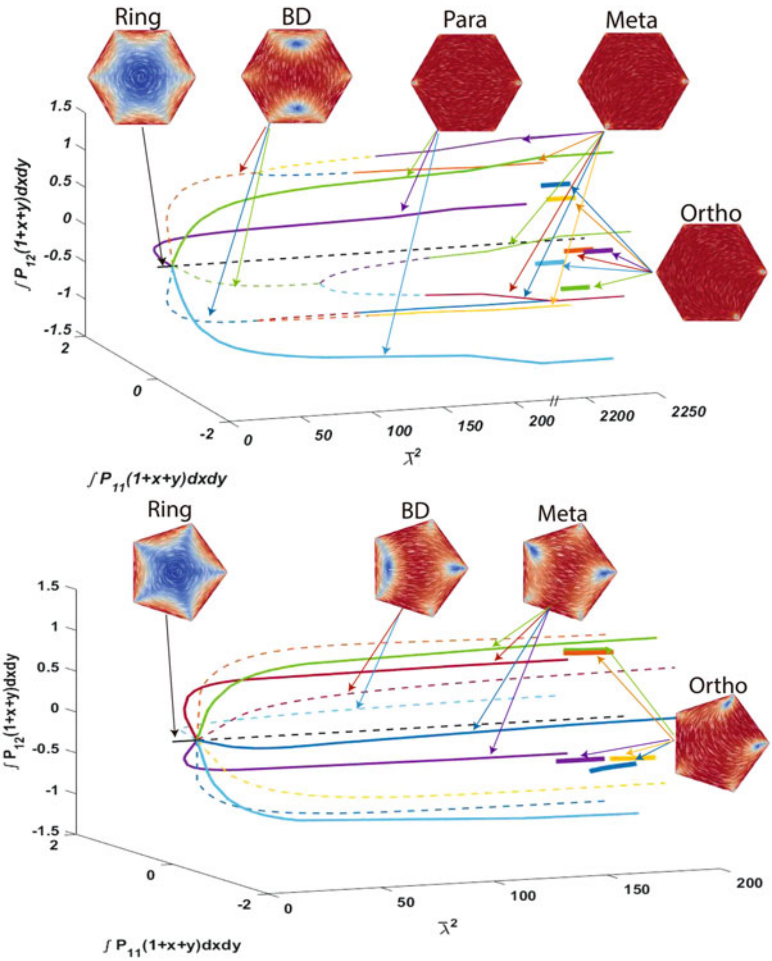


Fig. 7 Bifurcation diagrams for reduced LdG model in regular hexagon (top) and pentagon (bottom) domains, as a function of $\bar{\lambda}^2 = \frac{2C\lambda^2}{L}$. Reproduced from [16] with permission from Society for Industrial and Applied Mathematics

and

$$
\begin{aligned}
P_{11}^0(x, y) \; &= \sum_{k\ odd} \frac{4\sin(k\pi d/a)}{k^2\pi^2 d/a} \sin\left(\frac{k\pi x}{a}\right) \frac{\sinh(k\pi(1-y)/a) + \sinh(k\pi y/a)}{\sinh(k\pi/a)} \\
&\quad - \sum_{k\ odd} \frac{4\sin(k\pi d)}{k^2\pi^2 d} \sin(k\pi y) \frac{\sinh(k\pi(a-x)) + \sinh(k\pi x)}{\sinh(k\pi a)}, \qquad (24)
\end{aligned}
$$

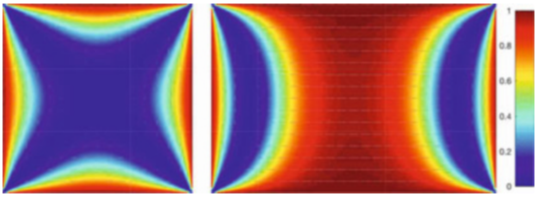


Fig. 8 Left: $WORS$ on square. Right: $BD2$ on rectangle with $a = 1.5$. The colour bar represents the value of $s^2 = |\mathbf{P}|/2$ in this and the next figure. Reproduced from [10] with permission from SAGE Publications

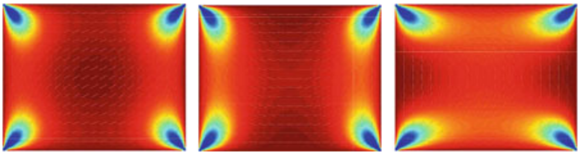


Fig. 9 From left to right: D, $R1$, $R3$ on rectangle with $a = 1.25$. Reproduced from [10] with permission from SAGE Publications

where d is the size of mismatch region near the rectangular vertices; we linearly interpolate between the boundary conditions on the two intersecting edges to define the boundary value at the vertices.

On a square with $a = 1$, the $WORS$ solution is distinguished by $\mathbf{P} = 0$ on the square diagonals. In fact, we can use the symmetry of the Laplace equations, Dirichlet boundary condition, and the geometry, to show that $\mathbf{P}^0(1/2, 1/2) = 0$. However, by constructing multiple auxiliary boundary-value problem on $[0, a] \times [0, a]$, $[0, a] \times [0, 1]$, $[0, 1] \times [0, 1]$ with suitable boundary conditions and using the maximum principle multiple times, one can prove that $P^0_{11}(a/2, 1/2) > 0$ on a rectangle with $a > 1$. The details are omitted here for conciseness, and readers are referred to the elegant arguments in Proposition 3.3 in [10]. In the $d \to 0$ limit, the result $P^0_{11}(a/2, 1/2) > 0$ still holds. Hence, we lose the $WORS$-cross structure for $a \neq 1$, i.e., as soon as we break the symmetry of the square domain. In Fig. 8, we show the differences between the $WORS$ on a square and the limiting profile on a rectangle, labelled as $BD2$. The $BD2$ is featured by disentangled line defects near opposite short edges.

In $\epsilon \to 0$ limit relevant for macroscopic domains or large λ, analogous to the approach in [29], the energy minimizers of (17) can be studied in terms of Dirichlet boundary-value problems for the director angle. As with the square, there are two diagonal D states for which $\mathbf{n}$ in (19) is aligned along a diagonal of the rectangle, the rotated $R1$ and $R2$ states for which $\mathbf{n}$ rotates by π radians between a pair of parallel horizontal edges, and the rotated $R3$, $R4$ states for which $\mathbf{n}$ rotates by π radians between a pair of parallel vertical edges (Fig. 9). For $a > 1$, the $R3$, $R4$ states have higher energies than the $R1$, $R2$ states (see [29, 47] for details), breaking the energy degeneracy of the rotated solutions on a square. Interested readers are referred to [10] for bifurcation diagrams on rectangles, for different values of a, that capture the effects of geometrical anisotropy on NLC solution landscapes.

6 Effects of Elastic Anisotropy

In this section, we study the effects of elastic anisotropy on the critical points of the reduced LdG energy (11) on square domains, with tangent boundary conditions. The elastic anisotropy is captured by a parameter $\hat{L}_2$ in (6). As in previous sections, we restrict ourselves to **Q**-tensors with three degrees of freedom q_1, q_2, and q_3 in (12). In the following paragraphs, we review the modelling details, theoretical analyses, and numerical results from [18].

Substituting the **Q**-tensor ansatz (12) into (11), and writing the energy functional as a function of $(q_1, q_2, q_3) \in W^{1,2}(\Omega; \mathbb{R}^3)$, we have

$$J[q_1, q_2, q_3] := \int_\Omega f_{el}(q_1, q_2, q_3) + \tfrac{\lambda^2}{L} f_b(q_1, q_2, q_3)\,\mathrm{dA}, \tag{25}$$

where

$$f_b(q_1, q_2, q_3) := A(q_1^2 + q_2^2 + 3q_3^2) + C(q_1^2 + q_2^2 + 3q_3^2)^2 + 2Bq_3(q_1^2 + q_2^2 - q_3^2), \tag{26}$$

and

$$f_{el}(q_1, q_2, q_3) := \left(1 + \frac{\hat{L}_2}{2}\right)|\nabla q_1|^2 + \left(1 + \frac{\hat{L}_2}{2}\right)|\nabla q_2|^2 + \left(3 + \frac{\hat{L}_2}{2}\right)|\nabla q_3|^2$$
$$+\hat{L}_2(q_{1,y}q_{3,y} - q_{1,x}q_{3,x} - q_{2,y}q_{3,x} - q_{2,x}q_{3,y}) + |\hat{L}_2|(q_{2,y}q_{1,x} - q_{1,y}q_{2,x}). \tag{27}$$

The elastic energy density can be rewritten in the following two ways: if $\hat{L}_2 \geq 0$,

$$f_{el} = |\nabla q_1|^2 + |\nabla q_2|^2 + 3|\nabla q_3|^2 + \frac{\hat{L}_2}{2}((q_{1,x} + q_{2,y} - q_{3,x})^2 + (q_{2,x} - q_{1,y} - q_{3,y})^2), \tag{28}$$

and if $\hat{L}_2 < 0$,

$$f_{el} = (1 + \hat{L}_2)(|\nabla q_1|^2 + |\nabla q_2|^2 + 3|\nabla q_3|^2) \tag{29}$$
$$-\frac{\hat{L}_2}{2}((-q_{3,x} - q_{1,x} - q_{2,y})^2 + (q_{2,x} - q_{1,y} + q_{3,y})^2 + 4|\nabla q_3|^2). \tag{30}$$

To ensure the non-negativity of the elastic energy density, we assume $\hat{L}_2 \in (-1, \infty)$. The corresponding EL equations are

$$\left(1 + \frac{\hat{L}_2}{2}\right)\Delta q_1 + \frac{\hat{L}_2}{2}(q_{3,yy} - q_{3,xx}) = \frac{\lambda^2}{L} q_1(A + 2Bq_3 + 2C(q_1^2 + q_2^2 + 3q_3^2)), \tag{31}$$

$$\left(1+\frac{\hat{L}_2}{2}\right)\Delta q_2 - \hat{L}_2 q_{3,xy} = \frac{\lambda^2}{L} q_2(A + 2Bq_3 + 2C(q_1^2 + q_2^2 + 3q_3^2)), \tag{32}$$

$$\left(1+\frac{\hat{L}_2}{6}\right)\Delta q_3 + \frac{\hat{L}_2}{6}(q_{1,yy} - q_{1,xx}) - \frac{\hat{L}_2}{3} q_{2,xy} = \frac{\lambda^2}{L} q_3(A - Bq_3 + 2C(q_1^2 + q_2^2 + 3q_3^2)) + \frac{\lambda^2 B}{3L}(q_1^2 + q_2^2). \tag{33}$$

The EL equations (31)–(33) do not have the elegant Laplace structure and hence are not readily amenable to analytic methods. Notably, we do not have an explicit maximum principle argument for the solutions of (31)–(33), as in the Dirichlet case with $\hat{L}_2 = 0$.

Analogous to Theorem 2.2 in [2], we can prove the existence of minimizers of J in (25) in the admissible class

$$\mathcal{A}_0 := \{(q_1, q_2, q_3) \in W^{1,2}(\Omega; \mathbb{R}^3) : q_1 = q_b,\ q_2 = 0,\ q_3 = -s_+/6 \text{ on } \partial\Omega\}, \tag{34}$$

where q_b is piecewise of class C^1, and q_1 is prescribed to ensure that the tangent boundary conditions are satisfied (at least away from the vertices). For λ small enough, we can prove that the LdG energy (25) has a unique critical point, but the proof is more involved than in [28], with additional embedding theorems and functional inequalities.

We can analytically construct a symmetric critical point for all admissible values of $\hat{L}_2$ and edge lengths λ, and we quote the relevant proposition from [18] below. Here, to avoid discussing the mismatch of the boundary conditions at the square vertices, we assume the domain Ω to be a truncated square, see Fig. 10. The proposition is also valid for a non-truncated square, with interpolation-type boundary condition near the vertices.

Proposition 3 *There exists a critical point (q_1^s, q_2^s, q_3^s) of the energy functional (25) in the admissible space $\mathcal{A}_0$, for all $\lambda > 0$, such that q_1^s is odd about the square diagonals and x- and y-axis (see Fig. 10), q_2^s has even reflections about the square diagonals and odd reflection about x- and y-axis, q_3^s has even reflections about the square diagonals and x- and y-axis. Subsequently, $q_1 : \Omega \to \mathbb{R}$ vanishes along the square diagonals $y = x$ and $y = -x$, and the function $q_2 : \Omega \to \mathbb{R}$ vanishes along $y = 0$ and $x = 0$.*

Subsequently, we can exploit the structure of the Eqs. (31)–(33) and the boundary conditions to prove that for $A < 0$ and $\hat{L}_2 \neq 0$, the critical point constructed in Proposition 3, has non-constant q_2^s on Ω, for all $\lambda > 0$. This symmetric critical point is globally stable for small domains size, i.e., the edge length λ is small enough (see Fig. 11). When $\hat{L}_2 = 0$, this symmetric critical point is the $WORS$ defined by (13).

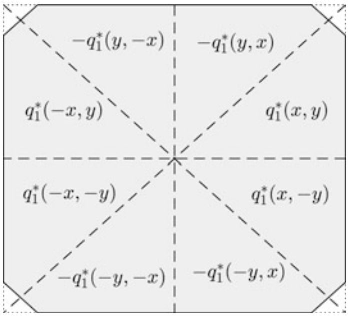


Fig. 10 (Credit [18]). The reflected solution $q_1^s(x, y)$ in Proposition 3

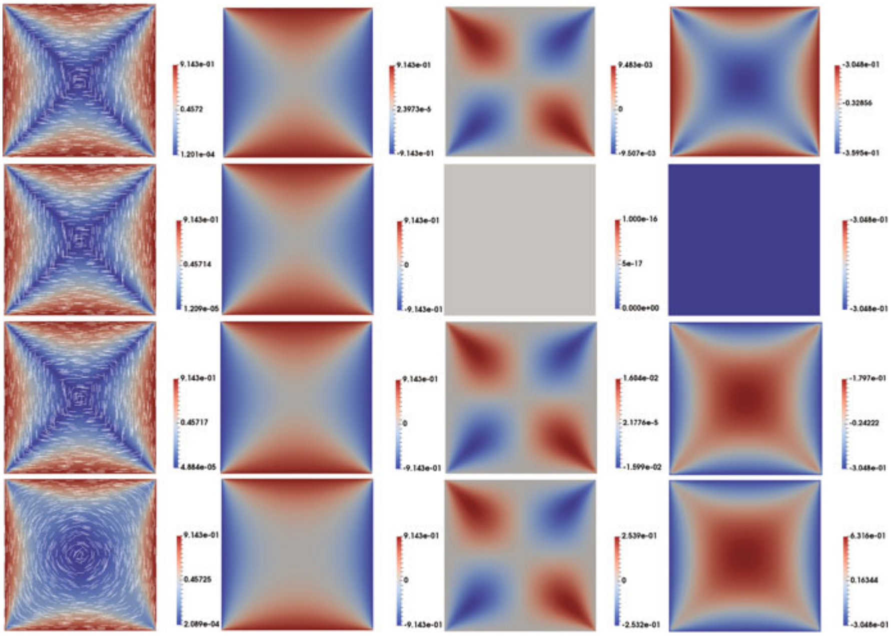


Fig. 11 (Credit [18]). The unique stable solution of the Euler–Lagrange equations (31)–(33), with $\bar{\lambda}^2 = 5$, and (from the first to fourth rows) $\hat{L}_2 = -0.5$, 0, 1, and 10, respectively. In the first column, we plot the (q_1, q_2) profile. We plot the corresponding q_1, q_2, and q_3 profiles, in the second to fourth columns, respectively

Notably, $q_2 = 0$ everywhere for the $WORS$ (refer to (12)), which is equivalent to having a set of constant eigenvectors in the plane of Ω. When $|\hat{L}_2| > 0$, q_2 and q_3 are non-constant, which means that we lose the constant eigenvectors and subsequently the cross structure in $WORS$. When $\hat{L}_2 = -0.5$, 1, and 10, we have a central +1-point defect in the profile of (q_1, q_2), and we label this as the $Ring^+$ solution (Fig. 11).

We investigate the effect of $\hat{L}_2$ on the $WORS$ profile, using asymptotic methods. The $WORS$ solution is of the form of (12), given by the triplet $(q, 0, -B/6C)$ at the

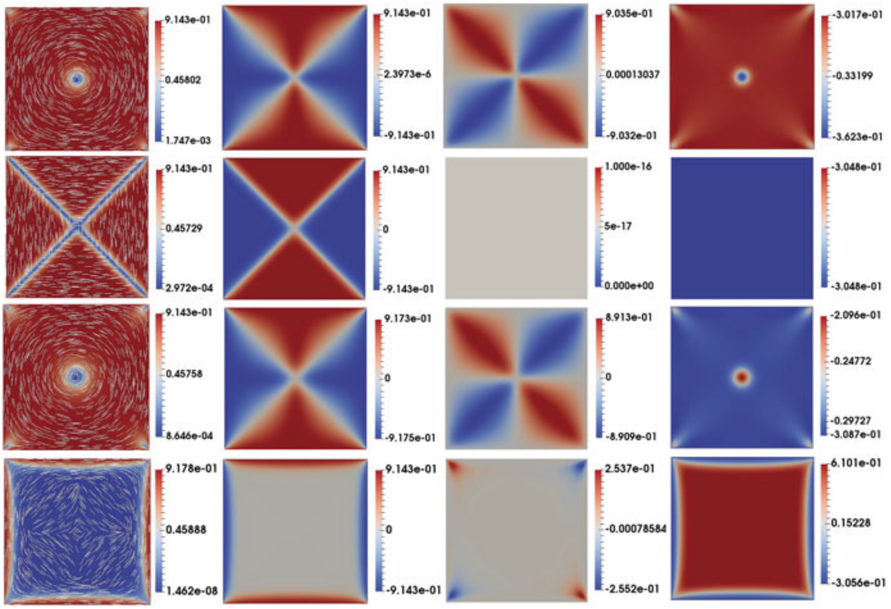


Fig. 12 (Credit [18]) A solution branch for the Euler–Lagrange system (31)–(33) with $\bar{\lambda}^2 = 500$, and $\hat{L}_2 = -0.5$, 0, 1, and 10, respectively, plotted in the first to fourth rows, respectively. This solution branch is a symmetric solution branch, as described in Proposition 3. When $\hat{L}_2 = -0.5$, 0, and 1, the plotted solution is unstable. When $\hat{L}_2 = 10$, the plotted solution is stable. The first column contains plots of (q_1, q_2). In the second to fourth columns, we plot the corresponding, q_1, q_2 and q_3, profiles

fixed temperature $A = -B^2/3C$, where q is a solution of the Allen–Cahn equation, as in [6]. With the leading order approximation given by $(q, 0, -B/6C)$, we expand q_1, q_2, q_3 in powers of $\hat{L}_2$ as follows:

$$\begin{aligned} q_1(x, y) &= q(x, y) + \hat{L}_2 f(x, y) + \dots \\ q_2(x, y) &= \hat{L}_2 g(x, y) + \dots \\ q_3(x, y) &= -\frac{B}{6C} + \hat{L}_2 h(x, y) + \dots \end{aligned} \tag{35}$$

for some functions f, g, h, which vanish on the boundary.

For λ small enough, one can show that the corrections (f, g, h) are unique, $g \equiv 0$ on Ω, and $f(x, y) = 0$ on diagonals. Hence, for λ small enough, the cross structure of the $WORS$ is lost mainly because of effects of $\hat{L}_2$ on the component, q_3.

We work at the fixed temperature $A = -B^2/3C$ for all the following numerical results. We perform a parameter sweep of $\bar{\lambda}^2$, from 5 to 500, and find one of the symmetric solution branches constructed in Proposition 3, with various fixed $\hat{L}_2$. The solutions with $\bar{\lambda}^2 = 500$ are plotted in Fig. 12. When $\hat{L}_2 = 0$, we recover

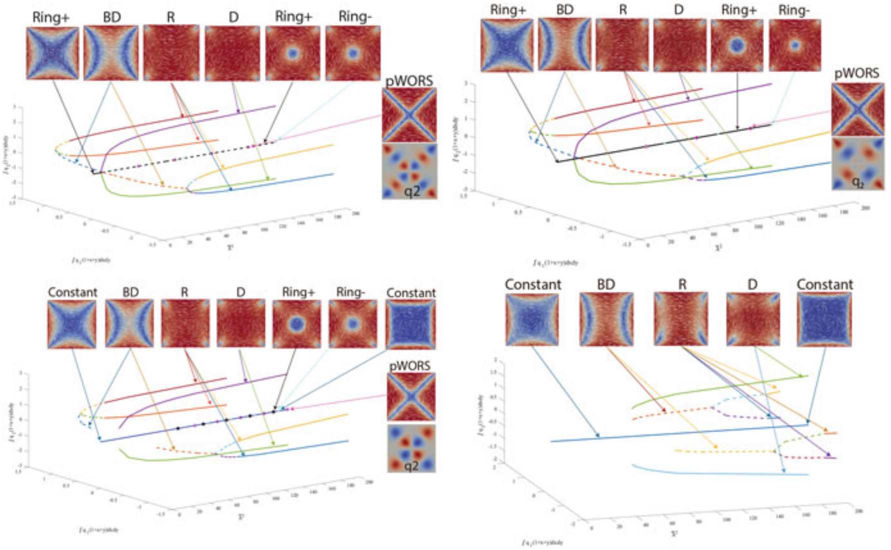


Fig. 13 (Credit [18]) Bifurcation diagrams for the LdG model in square domain with $\hat{L}_2 = 1, 2.6,$ 3, and 10 from top to bottom

the familiar $WORS$ for all $\lambda > 0$. When $-1 < \hat{L}_2 < 0$, the solution exhibits a $+1$ defect at the square centre, and we refer to it as the $Ring^+$ solution. When $\hat{L}_2$ is positive and moderate in value, we recover the $Ring^+$ solution branch and $q_3 > -s_+/6$ at the square centre. When $\hat{L}_2$ is large enough, we discover a new symmetric solution that is approximately constant, $(q_1, q_2, q_3) = (0, 0, s_+/3)$, away from the square edges, as shown in the fourth row of Fig. 12 for $\hat{L}_2 = 10$. We refer to this novel solution as the *Constant* solution.

As stated in Sect. 3, for large λ and with $\hat{L}_2 = 0$, the D and R states are the competing energy minimizers in this reduced framework. For large λ, with small or moderate $\hat{L}_2$, the D and R states still survive. When $\hat{L}_2 = 0$ and for fixed $A < 0$, $s^2 = q_1^2 + q_2^2 \approx s_+^2/4$, $q_3 = -s_+/6$ almost everywhere on Ω. As $|\hat{L}_2|$ increases, q_3 deviates significantly from the limiting value $q_3^\infty = -s_+/6$, near the square vertices; the deviation being more significant near the bend vertices compared to the splay vertices. From an optical perspective, we expect to observe larger defects near the square vertices for anisotropic materials with $\hat{L}_2 \gg 1$. For large λ and large $\hat{L}_2$, the D, R, and $Constant$ states are three energetically competing states. In [29], as $\lambda \to \infty$, the authors compute the limiting energies of D and R solutions, and the energy estimates are linear in $\hat{L}_2$. The $Constant$ solution has transition layers near the square edges and as in Section 4 of [49], by using the geodesic distance theory, we can show that there is a critical value $\hat{L}_2^*$, such that for $\hat{L}_2 > \hat{L}_2^*$, the limiting $Constant$ solution has lower energy than the competing D and R solutions.

In what follows, we compute bifurcation diagrams as a function of λ, with fixed temperature $A = -B^2/3C$, for five different values of $\hat{L}_2 = 1, 2.6, 3, 10$

in Fig. 13. We numerically discover at least 5 classes of symmetric critical points constructed in Proposition 1—the $WORS$, $Ring^{\pm}$, $Constant$, and the $pWORS$ solutions, of which the $WORS$, $Ring+$, and the $Constant$ solutions can be stable. The bifurcation diagram for $\hat{L}_2 = 0$, the elastically isotropic case, is discussed in Sect. 3. For $\hat{L}_2 = 1$, the $WORS$ ceases to exist, and the unique solution in the $\lambda \to 0$ limit is the stable $Ring^+$ solution. At the first bifurcation point $\lambda = \lambda^*$, the $Ring^+$ solution bifurcates into an unstable $Ring^+$ and two stable D solutions. At the second bifurcation point, $\lambda = \lambda^{**} > \lambda^*$, the unstable $Ring^+$ bifurcates into two unstable BD solutions, and for $\lambda = \lambda^{***} > \lambda^{**}$, the unstable $Ring^-$ and unstable $pWORS$ solution branches appear. In the (q_1, q_2) plane, the $pWORS$ has a constant set of eigenvectors away from the diagonals and has multiple $\pm 1/2$-point defects on the two diagonals, so that the $pWORS$ is similar to the $WORS$ away from the square diagonals. The $Ring^-$ and $pWORS$ are always unstable, and the $Ring^+$ solution has slightly lower energy than the $Ring^-$. The unstable $pWORS$ has higher energy than the unstable $Ring^{\pm}$ solutions when λ is large. The solution landscapes for $\hat{L}_2 = 1$ and $\hat{L}_2 = 2.6$ are qualitatively similar. For $\hat{L}_2 = 3$, the unique stable solution, for small λ, is the $Constant$ solution, which remains stable for $\bar{\lambda}^2 \leq 200$. The $Constant$ solution approaches $(q_1, q_2, q_3) \to (0, 0, s_+/3)$, as λ or $\hat{L}_2$ gets large. The BD and D solution branches are disconnected from the stable $Constant$ solution branch. For $\lambda = \lambda^*$, the stable $Ring^+$ appears, and for $\lambda = \lambda^{**} > \lambda^*$, the unstable $Ring^-$ and $pWORS$ appear. For $\hat{L}_2 = 10$, i.e., for very anisotropic materials, the $pWORS$ and $Ring^{\pm}$ states disappear, and the $Constant$ solution does not bifurcate to any known states. The $Constant$ solution has lower energy than the R and D solutions, for large λ. For much larger values of $\hat{L}_2$, we only numerically observe the $Constant$ solution branch, for the numerically accessible values of λ.

To summarize, the primary effect of the anisotropy parameter, $\hat{L}_2$, is on the unique stable solution for small λ. The elastic anisotropy destroys the cross structure of the $WORS$ and also enhances the stability of the $Ring^+$ and $Constant$ solutions. In fact, the $Constant$ solution is only observed for large $\hat{L}_2$. A further interesting feature for large $\hat{L}_2$ is the disconnectedness of the D and R solution branches from the parent $Constant$ solution branch. This indicates novel hidden solutions for large $\hat{L}_2$, which may have different structural profiles to the discussed solution branches.

7 NLC Solution Landscapes on a Hexagon

We have studied NLC equilibria on regular polygons, with or without elastic anisotropy. In this section, we investigate the solution landscape of a thin layer of NLC on a 2D hexagon, including stable and unstable critical points of the reduced LdG energy (11). The hexagon is a generic example of a 2D polygon with an even number of sides: the hexagon supports the generic $Ring$ solution for small domains, does not support the special symmetric solutions exclusive to a square (constructed in Proposition 1), and is better suited to capture generic trends with respect to geometrical parameters, as illustrated in Sect. 4.

First, we recap the essential concepts of a solution landscape. A *Solution Landscape* is a pathway map of connected solutions of a system of partial differential equations, in this case the Euler–Lagrange equations of the reduced LdG energy in (11). The solution landscape starts at a parent state (typically an unstable critical point of the LdG energy) and connects to stable energy minimizers via intermediate unstable critical points. More precisely, we can measure the degree of instability of an unstable critical point by means of its Morse index [38]. The Morse index of a critical/stationary point of the free energy is the number of negative eigenvalues of its Hessian matrix [38]. Energy minima or experimentally observable stable states are index-0 stationary points of the free energy with no unstable directions. A confined NLC system can switch between different energy minima or stable states, by means of an external field, thermal fluctuations, and mechanical perturbations. The switching requires the system to cross an energy barrier separating the two stable states, typically with an intermediate transition state. The transition state is an index-1 saddle point, the highest energy state along the transition pathway connecting the two stable states [54]. There are typically multiple transition pathways, with distinct transition states, and the optimal transition pathway has the lowest energy barrier. The reader is referred to [25] for transition pathways on a square domain with tangent boundary conditions and to [15] for transition pathways on a cylindrical domain with homeotropic/normal boundary conditions. Transition states are the simplest kind of saddle points of the free energy. Besides stable states and transition states, there are high-index saddle points with highly symmetric profiles and multiple interior defects, all of which offer fundamentally new scientific prospects.

In Sect. 4, we have reviewed the typical solutions, including $Ring$, BD, $Para(P)$, and $Meta(M)$, and the bifurcation diagram (Fig. 7a) of the critical points of (11), on a 2D hexagon. In what follows, we review results from [19] for NLC solution landscapes on regular 2D hexagons, as a function of the hexagon edge length, λ, at the fixed temperature, $A = -B^2/3C$. When λ^2 is sufficiently small, the $Ring$ solution is the unique stable solution as stated in Sect. 4. For $\bar{\lambda}^2 \approx 10$, the $Ring$ solution transitions from being a zero-index solution to an index-2 saddle point solution (with two equal negative eigenvalues), and we additionally have index-1 BD solutions and the index-0 P solutions. The solution landscape for $\bar{\lambda}^2 = 70$ is illustrated in Fig. 14, showing the relationships between $Ring$, BD, and P solutions. The $Ring$ solution is the parent state, i.e., the highest-index saddle point solution. Following each unstable eigen-direction of the $Ring$ solution shown in Fig. 14, the central $+1$ point defect splits into two defects that relax around a pair of opposite edges, i.e., the BD solutions. The two BD defects move from opposite edges to opposite vertices, following the single unstable eigenvector of the BD solution and converging to the corresponding P solution.

The solution landscape is quite complicated for $\bar{\lambda}^2 = 600$, as shown in Fig. 15a. There are three notable numerical findings in this regime: a new stable T solution with an interior $-1/2$ defect; new classes of saddle point solutions, H and TD,

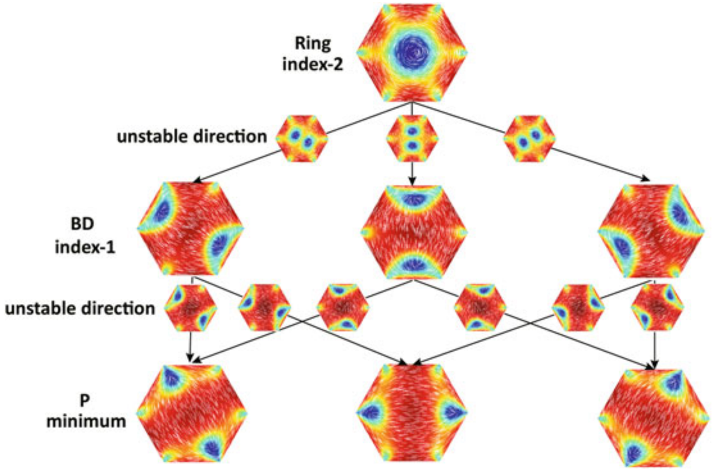


Fig. 14 Solution landscape at $\bar{\lambda}^2 = 70$. The index-2 $Ring$ is the parent state and connects to three index-1 BD solutions along its unstable directions. Each BD solution connects to two P minima along BD's single unstable direction. Reproduced from [19] with permission from IOP Publishing and the London Mathematical Society

with high symmetry and high indices; new saddle points with asymmetric defect locations.

The stable index-0 T solution is our first stable solution with an interior $-1/2$ defect at the centre of the hexagon, for $\bar{\lambda}^2 > 250$. The competing stable states, P and M, have defects pinned to the vertices, and these vertex defects are a natural consequence of the tangent boundary conditions and topological considerations (the total topological degree of the boundary condition is zero). The T solution on a hexagon (for large λ) is strongly reminiscent of the $Ring$ solution on a regular triangle (Fig. 15c), as reported in Sect. 4, suggesting that we can build new solutions by tessellating solutions on simpler building block-type polygons, such as the triangle and the square.

We numerically find a new class of saddle point solutions with high Morse indices and multiple interior defects, labelled as H class solutions, which have Morse indices ranging from 8 to 14 (Fig. 16). Notably, the parent state is the index-14 $H*$ saddle point solution connecting to the lowest index-8 saddle point solution, labelled as H. The saddle point $H*$ has no splay-like vertices, whereas H has 6 splay-like vertices. Numerically, we find that an index-m solution in the H class has $(m - 8)$ bend-like vertices, e.g., the index-8 H solution has no bend-like vertices, whereas the index-14 $H*$ solution has 6 bend-like vertices. Similar remarks apply to the saddle points in the TD class, i.e., a TD-type saddle point with m bend-like vertices is index-$(m + 3)$.

Next, we illustrate a comprehensive network of transition pathways between stable states including two T, six M, and three P solutions, for $\bar{\lambda}^2 = 600$ in Fig. 17. First, we remark that some stable and configurationally close solutions

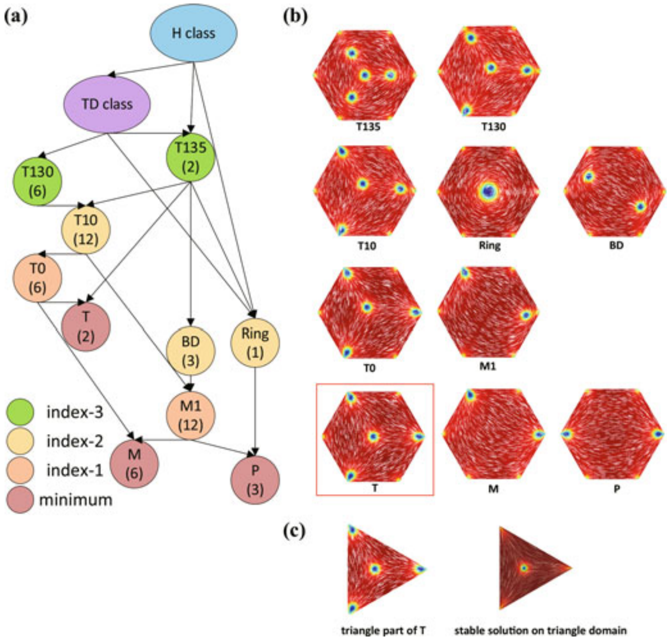


Fig. 15 (**a**) Solution landscape at $\bar{\lambda}^2 = 600$. (**b**) The configurations corresponding to (**a**). (**c**) The triangle part of T solution on a hexagonal domain Ω and stable $Ring$ solution on a triangle domain with $\bar{\lambda}^2 = 450$. Reproduced from [19] with permission from IOP Publishing and the London Mathematical Society

can be connected by a single transition state (index-1 saddle point) in Fig. 17. For example, the transition state between T_{left} and M_{26} is $T0_4$, and the transition state between M_{26} and P_{25} is $M1_{62}$. However, two different M or P solutions cannot be connected by means of a single index-1 transition state, i.e., the transition pathway typically involves an intermediate stable P or M state, risking entrapment.

The most complicated transition pathway appears to be the pathway between the two stable T solutions: T_{left} and T_{right}. In fact, one numerically computed transition pathway between T_{left} and T_{right} is T_{left}–$T0_4$–M_{26}–$M1_{62}$–P_{25}–$M1_{15}$–M_{15}–$T0_3$–T_{right}, where $T0_4$, $M1_{62}$, $M1_{15}$, and $T0_3$ are transition states (index-1 saddle points). This shows that a transition between two energetically close but configurationally far T solutions may have to overcome four energy barriers and could be easily trapped by the stable M or P solutions. This is not a reliable way of achieving switching because of the intermediate stable states.

An alternative approach is to use higher-index saddle points with multiple unstable directions, to connect configurationally far stable solutions. Figure 18 shows how the different P, M, and T solutions are connected by high-index saddle points. Two M solutions or two P solutions can be connected by the index-2 BD solution, and the system will not be trapped by the transient local minima along this pathway. The T_{left} and T_{right} solutions are configurationally far and can be connected

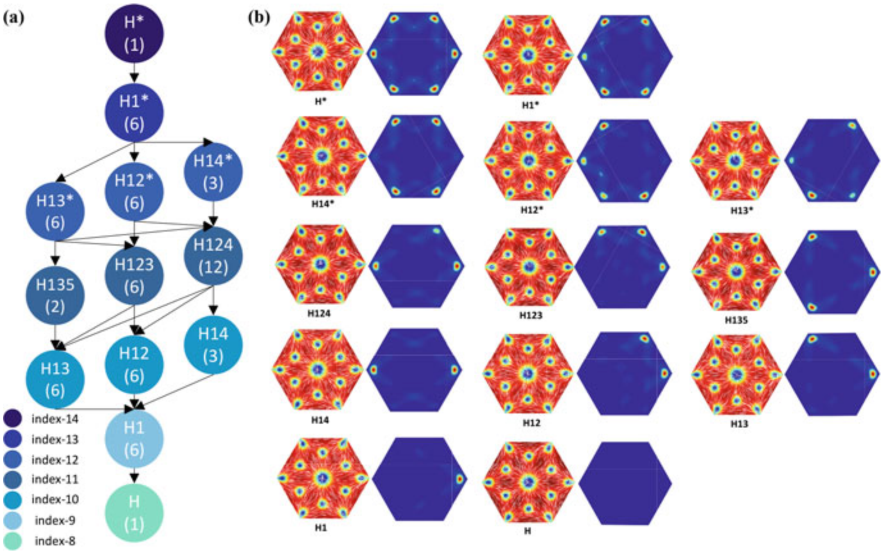


Fig. 16 (**a**) Solution landscape of the H class. (**b**) The corresponding configurations and plots of $|\mathbf{P} - \mathbf{P}^H|$, where $\mathbf{P}$ is any solution in the H class, and $\mathbf{P}^H$ is the index-8 H solution. Reproduced from [19] with permission from IOP Publishing and the London Mathematical Society

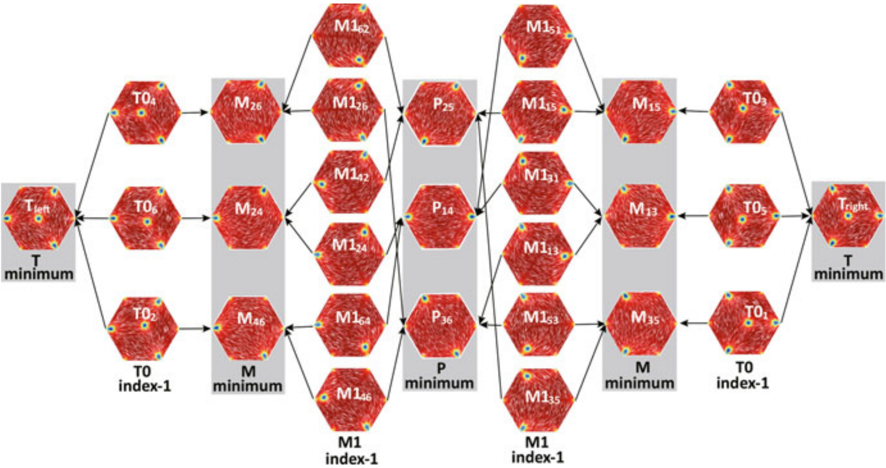


Fig. 17 The transition pathways between stable states including two T, six M, and three P solutions, for $\bar{\lambda}^2 = 600$. Reproduced from [19] with permission from IOP Publishing and the London Mathematical Society

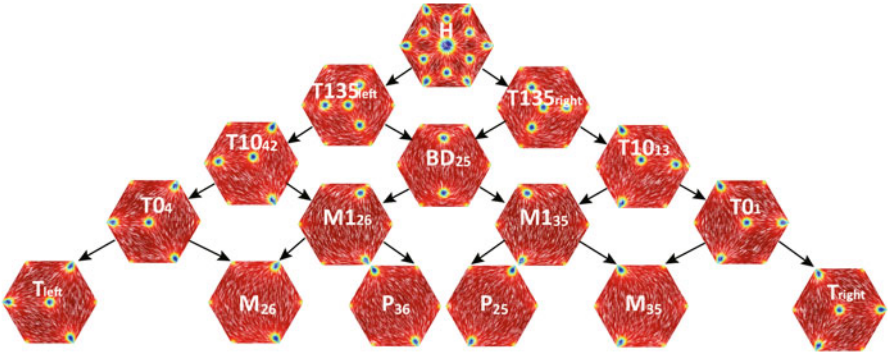


Fig. 18 Solution landscape starting from the H solution. All local minima such as $T_{\rm left}$, M_{26}, P_{36}, P_{25}, M_{35}, and $T_{\bf right}$ are connected by the index-8 H solution. Reproduced from [19] with permission from IOP Publishing and the London Mathematical Society

by an index-8 H solution: $T_{\rm left} \leftarrow T135_{\rm left} \leftarrow H \rightarrow T135_{\rm right} \rightarrow T_{\rm right}$. The index-8 H saddle point is connected to every stable solution, and we can thus construct dynamical pathways from the H solution to every individual stable solution.

Our numerical results highlight the differences between transition pathways mediated by index-1 saddle points and pathways mediated by high-index saddle points. We deduce that index-1 saddle points are efficient for connecting configurationally close stable solutions. For configurationally far stable states, they are generally connected by multiple transition states and intermediate stable states, or it may be possible to find a dynamical pathway between these configurationally far stable states via high-index saddle points. The selection of dynamical pathways is an open problem of tremendous scientific and practical interest.

Finally, let us compare the solution landscapes on a hexagon with that on a square domain. This illustrates the effects of geometry on solution landscapes. The most obvious difference is on the parent state. The Morse index of the $WORS$ increases with the domain size, λ, and the $WORS$ is always the parent state for a square domain [53]. Intuitively, this is because the diagonal defect lines become longer, so that the Morse index of the $WORS$ also increases with increasing edge length/increasing λ. The $Ring$ solution, which is the analogue of the $WORS$ on a hexagon, is index-0 for λ small enough and is an index-2 saddle point solution for larger λ, i.e., the Morse index does not increase with increasing λ. The highest-index parent saddle point on a hexagon changes from the $Ring$ solution to the index-3 $T135$ and index-14 $H*$ (see Fig. 19b), respectively, where $T135$ and $H*$ solutions emerge through saddle node bifurcations, as λ increases. We believe that the hexagon is a more generic example of a regular polygon with an even number of sides than a square, and hence, we expect that the qualitative aspects of our numerical study on a hexagon will extend to arbitrary polygons with an even number of sides.

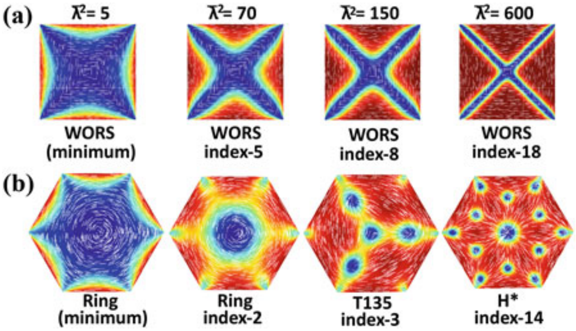


Fig. 19 Comparison of the parent states of the solution landscapes on the square (**a**) and the hexagon (**b**). On square, the parent state is always $WORS$, while, on hexagon, the parent state changes from $Ring$, $T135$, to $H*$ state. Reproduced from [19] with permission from IOP Publishing and the London Mathematical Society

8 Conclusions and Discussions

This review focuses on NLC equilibria in reduced 2D settings, within the reduced LdG framework (11). We look at regular polygons and the effects of elastic anisotropy captured by a parameter $\hat{L}_2$, with some preliminary work on the effects of geometrical anisotropy. The geometrical size is captured by a typical length (e.g., edge length of a polygon), denoted by λ. The $\lambda \to 0$ and $\lambda \to \infty$ limits are analytically tractable. In fact, we have a unique globally stable NLC equilibrium for λ sufficiently small and multistability for λ sufficiently large. The shape of the geometry plays a crucial role in the structural details. For example, the $WORS$ with a pair of mutually orthogonal defect lines along the diagonals is exclusive to a square domain, where we observe the $Ring$ solution with a central $+1$ defect for all other polygons, except the equilateral triangle. For a regular triangle, the stable NLC equilibrium has a central $-1/2$ defect in the $\lambda \to 0$ limit. For a K-regular polygon with K edges, there are at least $\left[\frac{K}{2}\right]$-classes of stable NLC equilibria for λ large enough, so that the shape of the polygon has a crucial role in multistability.

The effects of $\hat{L}_2$ have only been reviewed on square domains. Elastic anisotropy destroys the perfect $WORS$-cross structure for small λ, enhances the stability of some symmetric critical points, and very importantly introduces a novel $Constant$ solution branch for large values of $\hat{L}_2$. The square is special, and we need more comprehensive studies on generic 2D domains to capture the effects of $\hat{L}_2$ on solution landscapes.

Of particular interest are the study of saddle points and dynamical pathways between NLC equilibria on a 2D hexagon, at a fixed temperature below the nematic supercooling temperature. We review results on high-index saddle points from [19], focusing on the effects of λ, and numerically illustrate dynamical pathways, with intermediate index-1 saddle points/transition states versus dynamical pathways with intermediate high-index saddle points. The high-index saddle points are poorly understood in the literature but can play a crucial role in switching, selection of stable states and transient non-equilibrium dynamics, all of which are relevant to applications of confined NLC systems.

With regard to future research avenues, the possibilities are tremendous. A natural question concerns the sensitivity of solution landscapes to shape variations, i.e., if the geometry is not fixed but can be optimized with regard to prescribed properties. In other words, can we use shape and topology to tune the Morse indices of critical points in the LdG framework? Similarly, can we mathematically analyze new composite materials with multiple order parameters, e.g., a nematic order parameter and a magnetic order parameter; see [17] for detailed numerical studies of a prototype model for ferronematics, on 2D polygons with tangent boundary conditions. Last but not the least, these problems rely on a delicate and challenging combination of tools from variational analysis, numerical analysis, simulations, and experiments. In [34] and [35], the authors perform numerical analyses of some finite-element methods for the reduced LdG model in (11), with *a priori* and *a posteriori* estimates for the discontinuous Galerkin Method and the Nitsche's method. Of course, the possibilities are endless, and our vision is to design and implement generic algorithms for partially ordered materials, which can select the best mathematical model for the system under consideration, and then do comprehensive searches of the solution landscapes, yielding deterministic recipes for NLC-based systems that are predicted, designed, and controlled by mathematical toolboxes.

9 Supplement: Numerical Methods

We have used various methods to discretize the domain Ω, in the case studies of this review. In Sect. 4 and 6, on arbitrary regular polygons, we use standard finite-element methods to solve the linear systems including the Laplace equation and the limiting problems in the $\lambda \to \infty$ limits. All finite-element simulations and numerical integrals are performed using the open-source package FEniCS [31], along with the LU solver and the Newton's method. Newton's method strongly depends on the initial condition. We typically use the analytic solutions in the asymptotic limits—e.g., the $\lambda \to 0$ or $\lambda \to \infty$ limit and the $\hat{L}_2 \to 0$ limit, or perturbations of these solutions, as the initial conditions for the numerical solver, for a range of values of λ. In Sect. 5, on rectangle domain, we use traditional finite difference schemes for square mesh. In Sect. 7, on hexagon domain, we apply finite difference schemes over triangular elements to approximate the spatial derivatives, by analogy with the conventional discretization of a square domain [9].

To compute the bifurcation diagrams consisting of known stable and unstable solution branches, we perform an increasing λ sweep for the unique solution branch such as $WORS$, $Ring$, $Constant$ for small λ, and decreasing λ sweep for the distinct $Para$, $Meta$, $Ortho$, D or R solution branches. We distinguish between the distinct solution branches by defining two new measures, e.g.,

$\int_\Omega P_{12}\,(1+x+y)\,dxdy$ and $\int_\Omega P_{11}\,(1+x+y)\,dxdy$, and plot these measures versus λ^2 for the different solutions. Actually, the specific form of measure depends on the central point and the shape of the domain Ω and the mathematical model. We study the stability of the solutions by numerically calculating the smallest real eigenvalue of the Hessian of the free energy and the corresponding eigenfunction using the LOBPCG (locally optimal block preconditioned conjugate gradient) method in [52] (which is an iterative algorithm to find the smallest (largest) k eigenvalues of a real symmetric matrix). A negative eigenvalue is a signature of instability, and we have local stability if all eigenvalues are positive.

To investigate unstable solutions of the Euler–Lagrange equations, labelled as saddle points, in Sect. 7, we use the high-index optimization-based shrinking dimer (HiOSD) method to compute any-index saddle points [52]. The high-index saddle dynamics for finding an index-k saddle point can be viewed as a transformed gradient flow for the state variable $\mathbf{x}$ and k direction variables $\mathbf{v}_i$. The stability analysis is performed to show that a linearly stable steady state of this dynamical system is exactly an index$-k$ saddle point.The HiOSD method is an efficient tool for the computation of unstable saddle points and (local and global) minimizers, without good initial guesses. The connectivity of saddle points, including transition pathways, can be well established via the downward search and upward search algorithms. By combining the HiOSD method with downward and upward search algorithms, we can construct the solution landscape systematically. For more details, the readers are referred to the reference [53]. In general, we track bifurcations by tracking the indices of solutions; a change in the index is a signature of a bifurcation and a possible change of stability properties.

Acknowledgments AM gratefully acknowledges support from the University of Strathclyde New Professor Fund, a Leverhulme International Academic Fellowship, a Royal Society Newton Advanced Fellowship and the DST-UKIERI grant on “Theoretical and experimental studies of suspensions of magnetic nanoparticles, their applications and generalizations.” AM gratefully acknowledges support from an OCIAM Visiting Fellowship, a Visiting Professorship from the University of Bath, and a Visiting Professorship from IIT Bombay (India). AM also acknowledges support from a London Mathematical Society Research Reboot Grant, which facilitated the preparation of this review article. YH is fully supported by a Royal Society Newton International Fellowship. AM and YH are grateful to Professor Lei Zhang (Peking University), Professor Lei Zhang (Shanghai Jiao Tong University), Dr Lidong Fang, and Mr Joseph Harris (University of Strathclyde) for helpful suggestions.

References

1. B. Bahadur, *Liquid Crystal-Applications And Uses*, vol. 1 (World Scientific, Singapore, 1990)
2. P. Bauman, J. Park, D. Phillips, Analysis of nematic liquid crystals with disclination lines. Arch. Rational Mech. Anal. **205**(3), 795–826 (2012)
3. F. Bethuel, H. Brezis, F. Hélein, Asymptotics for the minimization of a Ginzburg-Landau functional. Calculus Var. Partial Differ. Equ. **1**(2), 123–148 (1993)

4. F. Bethuel, H. Brezis, F. Hélein, et al., *Ginzburg-Landau Vortices*, vol. 13 (Springer, Berlin, 1994)
5. M.A. Brilleslyper, M.J. Dorff, J.M. McDougall, J.S. Rolf, L.E. Schaubroek, R.L. Stankewitz, K. Stephenson, *Explorations in Complex Analysis*, vol. 40 (Mathematical Association of America, Washington, 2012)
6. G. Canevari, A. Majumdar, A. Spicer, Order reconstruction for nematics on squares and hexagons: a Landau-de Gennes study. SIAM J. Appl. Math. **77**(1), 267–293 (2017)
7. G. Canevari, J. Harris, A. Majumdar, Y. Wang, The well order reconstruction solution for three-dimensional wells, in the Landau–de Gennes theory. Int. J. Non-Linear Mech. **119**, 103342 (2020)
8. J.L. Ericksen, Liquid crystals with variable degree of orientation. Arch. Rational Mech. Anal. **113**(2), 97–120 (1990). https://doi.org/10.1007/BF00380413
9. J. Fabero, A. Bautista, L. Casasús, An explicit finite differences scheme over hexagonal tessellation. Appl. Math. Lett. **14**(5), 593–598 (2001)
10. L. Fang, A. Majumdar, L. Zhang, Surface, size and topological effects for some nematic equilibria on rectangular domains. Math. Mech. Solids **25**(5), 1101–1123 (2020)
11. G. Friedel, Les états mésomorphes de la matière. Ann. Phys. **9**, 273–474 (1922)
12. S. Gantenbein, K. Masania, W. Woigk, J.P. Sesseg, T.A. Tervoort, A.R. Studart, Three-dimensional printing of hierarchical liquid-crystal-polymer structures. Nature **561**(7722), 226–230 (2018)
13. P.G. de Gennes, J. Prost, *The Physics of Liquid Crystals*, vol. 83 (Oxford University Press, Oxford, 1995)
14. D. Golovaty, J.A. Montero, P. Sternberg, Dimension reduction for the Landau-de Gennes model in planar nematic thin films. J. Nonlinear Sci. **25**(6), 1431–1451 (2015)
15. Y.C. Han, Y.C. Hu, P.W. Zhang, L. Zhang, Transition pathways between defect patterns in confined nematic liquid crystals. J. Comput. Phys. **396**, 1–11 (2019)
16. Y. Han, A. Majumdar, L. Zhang, A reduced study for nematic equilibria on two-dimensional polygons. SIAM J. Appl. Math. **80**(4), 1678–1703 (2020)
17. Y. Han, J. Harris, A. Majumdar, Tailored nematic and magnetization profiles on two-dimensional polygons. Phys. Rev. E **103**, 052702 (2021)
18. Y. Han, J. Harris, L. Zhang, A. Majumdar, Elastic anisotropy of nematic liquid crystals in the two-dimensional Landau–de Gennes model (2021, preprint). arXiv:2105.10253
19. Y. Han, J. Yin, P. Zhang, A. Majumdar, L. Zhang, Solution landscape of a reduced Landau–de Gennes model on a hexagon. Nonlinearity **34**(4), 2048–2069 (2021)
20. V. Jampani, R. Volpe, K.R. de Sousa, J.F. Machado, C. Yakacki, J. Lagerwall, Liquid crystal elastomer shell actuators with negative order parameter. Sci. Adv. **5**(4), eaaw2476 (2019)
21. J.C. Jones, Defects, flexoelectricity and RF communications: the ZBD story. Liquid Crystals **44**(12–13), 2133–2160 (2017)
22. J.H. Kim, M. Yoneya, H. Yokoyama, Tristable nematic liquid-crystal device using micropatterned surface alignment. Nature **420**(6912), 159–162 (2002)
23. M. Kléman, Defects in liquid crystals. Rep. Progress Phys. **52**(5), 555 (1989)
24. S. Kralj, A. Majumdar, Order reconstruction patterns in nematic liquid crystal wells. Proc. R. Soc. A Math. Phys. Eng. Sci. **470**(2169), 20140276 (2014)
25. H. Kusumaatmaja, A. Majumdar, Free energy pathways of a multistable liquid crystal device. Soft Matter **11**(24), 4809–4817 (2015)
26. J.P. Lagerwall, An introduction to the physics of liquid crystals, in *Fluids, Colloids and Soft Materials: An Introduction to Soft Matter Physics*, ed. by A. Fernandez-Nieves, A.M. Puertas (Wiley, Hoboken, 2016), pp. 307–340
27. J.P. Lagerwall, G. Scalia, A new era for liquid crystal research: applications of liquid crystals in soft matter nano-, bio-and microtechnology. Curr. Appl. Phys. **12**(6), 1387–1412 (2012)
28. X. Lamy, Bifurcation analysis in a frustrated nematic cell. J. Nonlinear Sci. **24**(6), 197–1230 (2014). http://dx.doi.org/10.1007/s00332-014-9216-7
29. A.H. Lewis, I. Garlea, J. Alvarado, O.J. Dammone, P.D. Howell, A. Majumdar, B.M. Mulder, M. Lettinga, G.H. Koenderink, D.G. Aarts, Colloidal liquid crystals in rectangular confinement: theory and experiment. Soft Matter **10**(39), 7865–7873 (2014)

30. F.H. Lin, C. Liu, Static and dynamic theories of liquid crystals. J. Partial Differ. Equ. **14**(4), 289–330 (2001)
31. A. Logg, K.A. Mardal, G.N. Wells, *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book*. Springer Science and Business Media, vol. 84 (Springer, Berlin, 2012)
32. C. Luo, A. Majumdar, R. Erban, Multistability in planar liquid crystal wells. Phys. Rev. E **85**(6), 061702 (2012)
33. W. Maier, A. Saupe, Eine einfache molekulare theorie des nematischen kristallinflüssigen zustandes. Z. Naturforsch. A **13**(7), 564–566 (1958)
34. R.R. Maity, A. Majumdar, N. Nataraj, Discontinuous Galerkin finite element methods for the landau–de gennes minimization problem of liquid crystals. IMA J. Numer. Anal. **41**(2), 1130–1163 (2021)
35. R.R. Maity, A. Majumdar, N. Nataraj, Error analysis of nitsche's and discontinuous Galerkin methods of a reduced landau–de gennes problem. Comput. Methods Appl. Math. **21**(1), 179–209 (2021)
36. A. Majumdar, Equilibrium order parameters of nematic liquid crystals in the landau-de gennes theory. Eur. J. Appl. Math. **21**(2), 181–203 (2010)
37. A. Majumdar, A. Zarnescu, Landau-de Gennes theory of nematic liquid crystals: the Oseen-Frank limit and beyond. Arch. Rational Mech. Anal. **196**(1), 227–280 (2010)
38. J.W. Milnor, M. Spivak, R. Wells, *Morse Theory*, vol. 1 (Princeton University Press, Princeton, 1969)
39. N.J. Mottram, C. Newton, Introduction to Q-tensor theory. Tech. Rep. 10, Department of Mathematics, University of Strathclyde (2004)
40. L. Onsager, The effects of shape on the interaction of colloidal particles. Ann. New York Acad. Sci. **51**(4), 627–659 (1949)
41. C. Oseen, The theory of liquid crystals. Trans. Faraday Soc. **29**(140), 883–899 (1933)
42. P. Palffy-Muhoray, Orientationally ordered soft matter: the diverse world of liquid crystals. Electronic-Liquid Crystal Communications (e-LC) (2007)
43. F. Reinitzer, Beiträge zur kenntniss des cholesterins. Monatshefte für Chemie **9**(1), 421–441 (1888)
44. M. Robinson, C. Luo, P.E. Farrell, R. Erban, A. Majumdar, From molecular to continuum modelling of bistable liquid crystal devices. Liquid Crystals **44**(14–15), 2267–2284 (2017)
45. A. Sonnet, A. Kilian, S. Hess, Alignment tensor versus director: description of defects in nematic liquid crystals. Phys. Rev. E **52**(1), 718 (1995)
46. I.W. Stewart, *The Static and Dynamic Continuum Theory of Liquid Crystals: A Mathematical Introduction* (CRC Press, Boca Raton, 2019)
47. C. Tsakonas, A. Davidson, C. Brown, N.J. Mottram, Multistable alignment states in nematic liquid crystal filled wells. Appl. Phys. Lett. **90**(11), 111913 (2007)
48. E.G. Virga, *Variational Theories for Liquid Crystals*, vol. 8 (CRC Press, Boca Raton, 1995)
49. Y. Wang, G. Canevari, A. Majumdar, Order reconstruction for nematics on squares with isotropic inclusions: A Landau-de Gennes study. SIAM J. Appl. Math. **79**(4), 1314–1340 (2019)
50. H.H. Wensink, Polymeric nematics of associating rods: phase behavior, chiral propagation, and elasticity. Macromolecules **52**(21), 7994–8005 (2019)
51. P.J. Wojtowicz, P. Sheng, E. Priestley, *Introduction to Liquid Crystals* (Springer, Berlin, 1975)
52. J. Yin, L. Zhang, P. Zhang, High-index optimization-based shrinking dimer method for finding high-index saddle points. SIAM J. Sci. Comput. **41**(6), A3576–A3595 (2019)
53. J. Yin, Y. Wang, J.Z. Chen, P. Zhang, L. Zhang, Construction of a pathway map on a complicated energy landscape. Phys. Rev. Lett. **124**(9), 090601 (2020)
54. L. Zhang, W.Q. Ren, A. Samanta, Q. Du, Recent developments in computational modelling of nucleation in phase transformations. NPJ Comput. Mater. **2**, 16003 (2016)

On Applications of Herglotz-Nevanlinna Functions in Material Sciences, I: Classical Theory and Applications of Sum Rules

Annemarie Luger and Miao-Jung Yvonne Ou

1 Introduction

This chapter deals with theory and applications of Herglotz-Nevanlinna functions, which are functions analytic in the complex upper halfplane and with non-negative imaginary part. They appear in surprisingly many circumstances and have been studied and utilized for a long time, which also explains why they do appear under several names. Here we are going to call them Herglotz-Nevanlinna functions (or Herglotz for short).

Even if the definition at first sight does not seem to be very restrictive, it does have strong implications. For more than a century, it is known that the set of all Herglotz-Nevanlinna functions is described via an integral representation using three parameters only, two numbers and a positive Borel measure (satisfying a reasonable growth condition). This explicit parameterization has made them a very powerful tool that has been used effectively both in pure mathematics and in applications.

It turns out that with such relatively simple functions, amazingly much information can be encoded. For example, Herglotz-Nevanlinna functions are in one-to-one correspondence with passive (one-port) systems. This means that the corresponding function "knows everything about the system." Another example are Sturm–Liouville differential operators, appearing in mathematical physics. Here for a given operator, its spectrum can be completely described in terms of the singularities of

A. Luger
Department of Mathematics, Stockholm University, Stockholm, Sweden
e-mail: luger@math.su.se

M.-J. Y. Ou (✉)
Department of Mathematical Sciences, University of Delaware, Newark, DE, USA
e-mail: mou@udel.edu

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_19

the corresponding Titchmarsh–Weyl coefficient, which is a Herglotz-Nevanlinna function. And even more, this function can still be used in order to describe the spectrum when the boundary conditions are changed. But these functions are not only used when working with a single system or operator, but can also be employed to deal with a whole class of problems simultaneously, as for instance when finding common bounds for the performance of all antennas that fit into a given volume (e.g., a ball of given radius), independently of their particular shape. In the study of composite materials, a similar situation arises in deriving bounds on effective properties when only the volume fractions are given; these bounds only depend on the volume fraction.

In recent years, there has been a series of workshops where mathematicians working in pure mathematics and in applied mathematics and experts in various applications have met. All participants have one common interest, Herglotz-Nevanlinna functions, but with very different perspectives and approaches. This two-part review article is an attempt to reflect and to present in a systematic and unified way the various pieces of mathematical theorems underpinning a diverse set of applications.

The structure of the current paper is as follows. After this introduction, in Sect. 2, we review the mathematical background for Herglotz-Nevanlinna functions and provide a common basis for the applications presented in Sect. 3 and in Part II, which is concluded with possible generalizations of the theory.

Section 2 starts with the well-known integral representation (Sect. 2.2), followed by various aspects that we consider to be relevant in the chosen applications. In particular, the behavior of a Herglotz-Nevanlinna function on/toward the real line (i.e., at the boundary of the domain) is detailed in Sects. 2.3 and 2.7. In material sciences, often the functions do have more specific properties, which are discussed in Sect. 2.4; in particular, Stieltjes functions are characterized. Besides the integral representation, other (equivalent) representations are also presented in Sect. 2.5. In Sect. 2.6, it is explained how Herglotz-Nevanlinna functions appear in the mathematical description of passive systems, and in Sect. 2.8, we briefly review matrix- (and operator-)valued Herglotz-Nevanlinna functions.

Section 3 (as well as Sect. 2 in Part II) is devoted to applications, where we present a diverse set of applications in material sciences with the underlying common theme of Herglotz-Nevanlinna functions. The common feature here is that the use of Herglotz-Nevanlinna functions makes it possible to handle a large class of problems at once, instead of changing the models according to details such as the shape of inclusions. In particular, in several situations, physical bounds can be derived, which provide estimates of, e.g., performance under certain conditions. In the applications presented here, the independent variable is either the frequency (in electromagnetics, poroelastics, quasi-static cloaking, as well as time-dispersive, dissipative systems) or the material contrasts (for composite material).

In Sect. 3.1, we describe how sum rules can be employed for deriving bounds for electromagnetic structures, and in Sect. 3.2, passive realizations/approximations of non-passive systems are found via optimization in terms of the corresponding Herglotz-Nevanlinna functions.

More applications can be found in Part II. They involve bounds on effective properties of composite materials, numerical treatment of a costly memory term in the modeling of poroelastic materials, as well as bounds for quasi-static cloaking and identifying certain time-dispersive and dissipative systems as restrictions of Hamiltonian systems.

Even if all these examples demonstrate the effectiveness of Herglotz-Nevanlinna functions, there are situations in applications that cannot be treated by these methods but would require more general classes of functions. This applies for instance for non-passive systems, e.g., appearing in electromagnetics, for which the analytic function in question might have non-positive imaginary part as well. Another example are composite materials with more than two phases. Then, even if the corresponding analytic functions still have positive imaginary part, they are not covered by the treatment above, since they depend on more than only one complex variable.

Therefore, in Sect. 3 of Part II, we provide an overview of the mathematics that is available for different classes of functions that extend the classical Herglotz-Nevanlinna class, and we expect them to be relevant for applications in material sciences.

We hope that this two-part review paper can be both helpful for people working in applications (by providing mathematical references for different aspects of Herglotz-Nevanlinna functions as well as their generalizations for future work) and interesting for pure mathematicians (by pointing out some relevant applications of Herglotz-Nevanlinna functions).

2 Mathematical Background

2.1 Definition and First Examples

In this chapter, the complex upper halfplane is denoted by $\mathbb{C}^+ := \{z \in \mathbb{C} : \operatorname{Im} z > 0\}$ and the right halfplane by $\mathbb{C}_+ := \{z \in \mathbb{C} : \operatorname{Re} z > 0\}$.

Definition 1 A function $h : \mathbb{C}^+ \to \mathbb{C}$ is called a Herglotz-Nevanlinna function if it is analytic in $\mathbb{C}^+$ and satisfies $\operatorname{Im} h(z) \geq 0$ for all $z \in \mathbb{C}^+$.

These functions appear at various places with different names: Herglotz, Nevanlinna, Pick, R-function (or some combination of these). In pure mathematics, Nevanlinna seems to be most used, whereas in applications often Herglotz is preferred.

Example 1 It is easy to check that the following functions belong to this class

$$f_1(z) = -\frac{1}{z-3} \quad f_2(z) = i \quad f_3(z) = -\frac{1}{z+i} \quad f_4(z) = \operatorname{Log} z \quad f_5(z) = \sqrt{z},$$

where for the last two functions the branch is chosen such that the functions map $\mathbb{C}^+$ into the upper halfplane. Other, maybe less obvious, examples are

$$f_6(z) = \tan z \qquad f_7(z) = \frac{\log\big(\Gamma(z+1)\big)}{z \log z},$$

where $\Gamma(z)$ denotes the Gamma function; see [6, 7].

Remark 1 By definition for a Herglotz-Nevanlinna function $\operatorname{Im} f(z) \geq 0$ for all $z \in \mathbb{C}^+$. However, it follows from a version of the maximum principle that if there is a point $z^* \in \mathbb{C}^+$ such that $\operatorname{Im} f(z^*) = 0$, then f is a (real) constant function.

Hence, if f and g are non-constant Herglotz-Nevanlinna functions, then the composition $F(z) := f\big(g(z)\big)$ is a Herglotz-Nevanlinna function as well. In particular, if $f \not\equiv 0$ is Herglotz-Nevanlinna, then both $g_1(z) := f\big(-\frac{1}{z}\big)$ and $g_2(z) := -\frac{1}{f(z)}$ are Herglotz-Nevanlinna functions.

When considering limits toward real points, then usually only non-tangential limits $z \hat{\to} x_0$ are considered, this means that z tends to $x_0 \in \mathbb{R}$ in some Stolz domain $D_\theta := \{z \in \mathbb{C}^+ : \theta < \operatorname{Arg}(z - x_0) < \pi - \theta\}$, where $0 < \theta < \frac{\pi}{2}$.

Remark 2 Herglotz-Nevanlinna functions can also be characterized via the boundary behavior only, namely an analytic function $f : \mathbb{C}^+ \to \mathbb{C}$ is Herglotz-Nevanlinna if and only if it holds $\limsup\limits_{z \hat{\to} x_0} \operatorname{Im} f(z) \geq 0$ (as a finite number or $+\infty$) for all $x_0 \in \mathbb{R} \cup \{\infty\}$.

2.2 Integral Representation

The main tool in the work with Herglotz-Nevanlinna functions is the following explicit representation, which in principle has been known for more than a century; see e.g., [25] and also [12].

Theorem 1 *A function $f : \mathbb{C}^+ \to \mathbb{C}$ is a Herglotz-Nevanlinna function if and only if there are numbers $a \in \mathbb{R}$, $b \geq 0$ and a (positive) Borel measure μ with $\int_{\mathbb{R}} \frac{1}{1+\xi^2} d\mu(\xi) < \infty$ such that*

$$f(z) = a + bz + \int_{\mathbb{R}} \left(\frac{1}{\xi - z} - \frac{\xi}{1+\xi^2} \right) d\mu(\xi). \tag{1}$$

Moreover, a, b, and μ are unique with this property.

Note that the term $\frac{\xi}{1+\xi^2}$ is needed for assuring the convergence of the integral.

Remark 3 Alternatively, representation (1) can also be written as

$$f(z) = a + bz + \int_{\mathbb{R}} \frac{1 + \xi z}{\xi - z} d\sigma(\xi) \tag{2}$$

with the finite measure σ given by $d\sigma(\xi) := \frac{d\mu(\xi)}{1+\xi^2}$.

Given a Herglotz-Nevanlinna function, the constants a and b can be read off directly, namely, it holds

$$a = \operatorname{Re} f(i) \quad \text{and} \quad b = \lim_{y\to\infty} \frac{f(iy)}{iy}. \tag{3}$$

Example 2 For the functions in Example 1, we have for instance $\mu_1 = \delta_3$, the point measure with mass 1 at the point $\xi_0 = 3$, is the representing measure for f_1, for f_2 the measure is a multiple of the Lebesgue measure $\mu_2 = \frac{1}{\pi}\lambda_{\mathbb{R}}$, whereas the representing measure μ_3 of f_3 is absolutely continuous with respect to the Lebesgue measure and has density $\frac{1}{\pi(1+\xi^2)}$, i.e., $d\mu_3(\xi) = \frac{1}{\pi(1+\xi^2)} d\lambda_{\mathbb{R}}(\xi)$.

Given the function, its representing measure can be reconstructed via the following formula, known as the Stieltjes inversion formula; see, e.g., [25].

Proposition 1 *Let f be a Herglotz-Nevanlinna function with integral representation (1). Then for $x_1 < x_2$, it holds*

$$\mu\big((x_1, x_2)\big) + \frac{1}{2}\mu(\{x_1\}) + \frac{1}{2}\mu(\{x_2\}) = \lim_{y\to 0+} \frac{1}{\pi} \int_{x_1}^{x_2} \operatorname{Im} f(x + iy)\, dx, \tag{4}$$

or, in a weak formulation, if h is a compactly supported smooth function in $C_0^1(\mathbb{R})$, then

$$\int_{\mathbb{R}} h(\xi) d\mu(\xi) = \lim_{y\to 0+} \frac{1}{\pi} \int_{\mathbb{R}} h(x) \operatorname{Im} f(x + iy)\, dx.$$

Moreover, point masses are given by

$$\lim_{z \hat{\to} \alpha} (\alpha - z) f(z) = \mu\big(\{\alpha\}\big). \tag{5}$$

By definition, a Herglotz-Nevanlinna function is defined in the upper halfplane $\mathbb{C}^+$ only. However, it can be extended naturally also to the lower halfplane $\mathbb{C}^-$, since the integral in the right-hand side of (1) is well-defined for all $z \in \mathbb{C} \setminus \mathbb{R}$. This extension is symmetric with respect to the real line, i.e.,

$$f(\overline{z}) = \overline{f(z)} \qquad z \in \mathbb{C} \setminus \mathbb{R}, \tag{6}$$

and is hence called *symmetric extension*.

Example 3 For some of the functions from Example 1, the symmetric extensions are

$$f_1(z) = -\frac{1}{z-3} \qquad f_2(z) = \begin{cases} i & \operatorname{Im} z > 0 \\ -i & \operatorname{Im} z < 0 \end{cases} \qquad f_3(z) = \begin{cases} -\frac{1}{z+i} & \operatorname{Im} z > 0 \\ -\frac{1}{z-i} & \operatorname{Im} z < 0 \end{cases}.$$

2.3 Boundary Behavior

We first note that for a Herglotz-Nevanlinna function f

$$\lim_{y \to 0+} f(x+iy) \text{ exists for almost all } x \in \mathbb{R}.$$

To see this, let φ be a Möbius transform that maps the unit disk $\mathbb{D}$ onto the open upper halfplane $\mathbb{C}^+$, e.g. $\varphi(w) = i\frac{1+w}{1-w}$. If f is a Herglotz-Nevanlinna function, then the function $h(w) := \varphi^{-1}\big(f(\varphi(w))\big)$ is a bounded analytic function in $\mathbb{D}$ and hence has boundary values almost everywhere. Therefore, it is also true for the Herglotz-Nevanlinna function f.

The weak form of the Stieltjes inversion formula also shows that the limit of the imaginary part always exists in the distributional sense. However, for pointwise limits, and good properties of the function on the boundary, more assumptions on the measure have to be imposed.

Let f be given with integral representation (1). If there is an interval (x_1, x_2) such that $(x_1, x_2) \cap \operatorname{supp} \mu = \emptyset$, then for every $x \in (x_1, x_2)$, the integral in (1) exists and is real analytic. Hence, the function can be extended analytically to the lower halfplane, and this analytic extension coincides with the symmetric extension.

But also in other cases, it can be possible to extend the Herglotz-Nevanlinna function analytically over (some part of) the real line. But then, in general, the continuation will not coincide with the symmetric extension. A characterization of this situation in terms of the measure is given in the following theorem; see [18].

Proposition 2 *Let f be a Herglotz-Nevanlinna function with representation* (1). *Then f can be continued analytically onto the interval* (x_1, x_2) *if and only if the measure μ is absolutely continuous with respect to the Lebesgue measure λ on this interval, and the density $\varrho(t)$ is real analytic on* (x_1, x_2). *In this case,*

$$f(z) = \overline{f(\overline{z})} + 2\pi i \varrho(z),$$

where $\varrho(z)$ denotes the analytic continuation of the density ϱ.

Example 4 The function f_2 in Example 1 can be extended as an entire function, $f_2(z) \equiv i$, whereas f_3 can be extended analytically only to the punctured plane $\mathbb{C} \setminus \{-i\}$.

Loosely speaking, an analytic density guarantees an analytic boundary function. However, for the boundary function to be continuous, it is not sufficient to assume that μ has a continuous density. As a counter example, consider the density

$$\varrho(\xi) = \begin{cases} -\dfrac{1}{\ln \xi}, & \xi \in (0, \gamma], \\ 0, & \xi \in [-\gamma, 0], \end{cases} \qquad (7)$$

which is continuous on the $[-\gamma, \gamma]$ for any $\gamma \in (0, 1)$, but for which the corresponding Herglotz-Nevanlinna function does not admit a continuous extension to $x = 0$.

The appropriate assumption here turns out to be Hölder continuity. A function $\varrho : (x_1, x_2) \to \mathbb{R}$ is called *Hölder continuous with exponent* α, that is $\varrho \in C^{0,\alpha}(x_1, x_2)$, if there exists a constant $C > 0$ such that

$$|\varrho(\xi_1) - \varrho(\xi_2)| \leq C \cdot |\xi_1 - \xi_2|^{\alpha} \quad \text{for all } \xi_1, \xi_2 \in (x_1, x_2).$$

The following theorem relies on some well-known results; a detailed proof for the current situation is given in [22, Theorem 2.2].

Proposition 3 *Let f be a Herglotz-Nevanlinna function with representation (1), and assume that there is an interval (x_1, x_2) where the measure μ is absolutely continuous with respect to the Lebesgue measure λ with Hölder continuous density ϱ. Then for every compact interval $I \subset (x_1, x_2)$, the function f admits a continuous extension to $\mathbb{C}^+ \cup I$. This continuation is given via the Hilbert transform*

$$f(x) = a + bx + p.v. \int_{\mathbb{R}} \left(\frac{1}{\xi - x} - \frac{\xi}{1 + \xi^2} \right) d\mu(\xi) + i\pi\varrho(x), \quad x \in I,$$

where the integral is taken as a principal value at $\xi = x$.

2.4 Subclasses

In this section, we focus on how properties of the measure in the integral representation (1) are related to properties of the function.

We start with the so-called symmetric functions, which are important for instance in connection with passive systems, cf., Sect. 2.6.

Definition 2 A Herglotz-Nevanlinna function is called symmetric if

$$f(-\overline{z}) = -\overline{f(z)}. \qquad (8)$$

Such functions are purely imaginary on the imaginary axes and can be characterized in the following way.

Proposition 4 *A Herglotz-Nevanlinna function f with representation* (1) *is symmetric if and only if $a = 0$ and μ is symmetric with respect to* 0, *i.e., $\mu(B) = \mu(-B)$ for every Borel set B in $\mathbb{R}$. In this case, the representation can be written as*

$$f(z) = bz + p.v. \int_{\mathbb{R}} \frac{1}{t-z} d\mu(t) \quad \textit{for } z \in \mathbb{C}^+,$$

where $p.v.$ denotes the principle value at ∞.

The function behavior at ∞ is closely related to the properties of the representing measure μ and related simplifications of the representation. The following statements can be found in [25]. The first theorem characterizes when the term $\frac{\xi}{1+\xi^2}$ is needed in the integral.

Theorem 2 *Let f be a Herglotz-Nevanlinna function with representation* (1). *Then the following are equivalent:*

(i) $\displaystyle\int_1^\infty \frac{\operatorname{Im} f(iy)}{y} dy < \infty.$

(ii) $\displaystyle\int_{\mathbb{R}} \frac{1}{1+|\xi|} d\mu(\xi) < \infty.$

(iii) $\displaystyle f(z) = s + \int_{\mathbb{R}} \frac{1}{\xi - z} d\mu(\xi)$ *with some* $s \in \mathbb{R}$.

In this case, $s = \lim\limits_{y\to\infty} f(iy) = \lim\limits_{y\to\infty} \operatorname{Re} f(iy) = a - \int_{\mathbb{R}} \frac{\xi}{1+\xi^2} d\mu(\xi).$

The next theorem characterizes functions with bounded measure.

Theorem 3 *Let f be a Herglotz-Nevanlinna function with representation* (1). *Then the following are equivalent:*

(i) $\displaystyle\lim_{z \hat{\to} \infty} \frac{f(z)}{\operatorname{Im} z} = 0 \quad \textit{and} \qquad \limsup_{z \hat{\to} \infty} |z| \operatorname{Im} f(z) < \infty.$

(ii) $\displaystyle\int_{\mathbb{R}} d\mu(\xi) < \infty.$

Hence also in this case, $\displaystyle f(z) = s + \int_{\mathbb{R}} \frac{1}{\xi - z} d\mu(\xi)$, *with* $s \in \mathbb{R}$.

An important subclass of Herglotz-Nevanlinna functions are Stieltjes functions; see also [25].

Definition 3 A holomorphic function $f : \mathbb{C} \setminus [0, +\infty) \to \mathbb{C}$ is called a Stieltjes function if:

- $\operatorname{Im} f(z) \geq 0$ for $\operatorname{Im} z > 0$.
- $f(x) \geq 0$ for $x \in (-\infty, 0)$.

These functions can be characterized in several different ways.

Theorem 4 *Let f be holomorphic in the domain $\mathbb{C} \setminus [0, +\infty)$. Then the following are equivalent:*

(a) f is a Stieltjes function.
(b) f can be represented as

$$f(z) = s + \int_{[0,\infty)} \frac{1}{\xi - z} d\mu(\xi)$$

with $s \geq 0$ and $\int_{[0,\infty)} \frac{1}{1+\xi} d\mu(\xi) < \infty$.
(c) f is a Herglotz-Nevanlinna function (analytically continued onto $\mathbb{R}^-$), which satisfies $\int_1^\infty \frac{\operatorname{Im} f(iy)}{y} dy < \infty$, and $\lim\limits_{y\to\infty} f(iy) \geq 0$.
(d) The functions $f(z)$ and $h_1(z) := zf(z)$ are Herglotz-Nevanlinna functions.
(e) The functions $f(z)$ and $h_2(z) := zf(z^2)$ are Herglotz-Nevanlinna functions.

In this case, $s = \lim\limits_{x\to-\infty} f(x)$.

Moreover, symmetric Herglotz-Nevanlinna functions can be represented via Stieltjes functions.

Theorem 5 *A function f is a symmetric Herglotz-Nevanlinna function, i.e., $f(-\overline{z}) = -\overline{f(z)}$, if and only if there exists a Stieltjes function h such that $f(z) = zh(z^2)$.*

Note that in some places the notion Stieltjes function means that additionally all moments of the representing measure exist. Other versions of Stieltjes functions where the functions are analytic on the other halfline are used in Sect. 2.1.1 of Part II.

Another important subclass is rational Herglotz-Nevanlinna functions. Here, the term *rational* might be understood in two different ways. One way is to think about functions for which there exists a rational function in $\mathbb{C}$ such that its restriction to the upper halfplane coincides with the given function, e.g., f_1, f_2, and f_3 in Example 1, as well as in connection with electrical circuit networks, cf., Example 9. Note that these functions might have absolutely continuous measures, such as f_2 and f_3.

But *rational* can also be interpreted in a more strict way, namely that the integral representation gives a rational function in $\mathbb{C}$, or in other words, that the symmetric extension is rational in $\mathbb{C}$. Among the above named examples, only f_1 is rational also in this sense. Rational functions in this stricter meaning are exact those functions for which the measure is a finite sum of Dirac measures, as e.g., when deriving bounds in Sect. 2.1.1 of Part II.

Also, more generally, meromorphic Herglotz-Nevanlinna functions have been investigated, e.g., in connection with inverse problems. An important property are the interlacing of zeros and poles on the real line.

2.5 *Other Representations*

Besides the integral representation, there also exist other ways to represent Herglotz-Nevanlinna functions.

2.5.1 Operator Representations

Representations using resolvents have been used in different contexts. The theorem below follows straightforwardly from Example 5 or can be seen as a special case of the results in, e.g., [27]. Here, self-adjoint linear relations are used; they can be viewed as multi-valued operators. For a detailed overview of relations in inner product spaces, see [14] or [5, Chapter 1].

Theorem 6 *A function f is a Herglotz-Nevanlinna function if and only if there exist a Hilbert space $\mathcal{H}$, a self-adjoint linear relation A in $\mathcal{H}$, a point $z_0 \in \mathbb{C}^+$, and an element $v \in \mathcal{H}$ such that*

$$f(z) = \overline{f(z_0)} + (z - \overline{z_0})\left((I + (z - z_0)(A - z)^{-1})v, v\right)_{\mathcal{H}}. \tag{9}$$

Moreover, if $\mathcal{H} = \overline{span}\{(I + (z - z_0)(A - z)^{-1})v : z \in \varrho(A)\}$, where $\overline{span}$ denotes closed linear span and $\varrho(A)$ the resolvent set of A, then the representation is called minimal. In this case, the representation is unique up to unitary equivalence.

If the representation is minimal, then it can be shown that $\text{hol}(f) = \varrho(A)$, meaning that the function f (more precisely, its symmetric continuation to the lower halfplane and to those real points where possible) is analytic exactly in the resolvent set of the representing relation A. In particular, isolated eigenvalues of A are poles of f. Non-isolated eigenvalues are then called generalized poles and can be characterized analytically as well. Since unitarily equivalent relations do have the same spectral properties, these are intrinsic for the function as well.

There are different (equivalent) ways to construct such an operator representation.

Example 5 If, for instance, the integral representation (1) is given, then the above representation can be realized as follows: If in the integral representation $b = 0$, then $\mathcal{H} = L^2_\mu$ and A is actually an operator. Namely, A is multiplication by the independent variable, i.e., $g(\xi) \mapsto \xi \cdot g(\xi)$. If z_0 is fixed, then $v \in L^2_\mu$ might be chosen as $v(\xi) = \frac{1}{\xi - \overline{z_0}}$.

If $b > 0$, then the space has an additional one-dimensional component, namely, $\mathcal{H} = L^2_\mu \oplus \mathbb{C}$ and A is not an operator but a relation with non-trivial multi-valued part $A(0)$. The relation A is acting in L^2_μ as multiplication by the independent variable and has the second component as multi-valued part, i.e., $A(0) = \{0\} \times \mathbb{C}$.

In Theorems 2 and 3, some properties of the function have been related to certain properties of the measure that lead to simplifications of the integral representation. In the following theorem, these results are extended to the operator representation.

Theorem 7 *Let f be a Herglotz-Nevanlinna function given by representation* (9). *Then:*

1. $\lim\limits_{y\to\infty} \dfrac{f(iy)}{y} = 0$ *if and only if the relation A is an operator, i.e., its multi-valued part is trivial.*
2. $\displaystyle\int_1^\infty \frac{\operatorname{Im} f(iy)}{y} dy < \infty$ *if and only if* $v \in \operatorname{dom}((|A| + I)^{1/2})$.
3. $\lim\limits_{z\hat{\to}\infty} \dfrac{f(z)}{\operatorname{Im} z} = 0$ *and* $\limsup\limits_{z\hat{\to}\infty} |z| \operatorname{Im} f(z) < \infty$ *if and only if A is an operator and* $v \in \operatorname{dom}(A)$. *In this case,*
$$f(z) = s + \Big((A - z)^{-1} u, u\Big)_{\mathcal{H}}$$
with $s \in \mathbb{R}$ and $u := (A - \overline{z_0})v$.

Operator representations appear naturally in connection with spectral problems for self-adjoint operators. For instance, the spectrum of a Sturm–Liouville operator can be characterized in terms of the singularities of the corresponding Titchmarsh–Weyl function, which in many cases is a Herglotz-Nevanlinna function. Then A is the differential operator, and μ can be interpreted as the spectral measure, see, e.g., [16] and references therein or Chapter 6 in [5].

Abstractly speaking, scalar Herglotz-Nevanlinna functions do appear in connection with rank one perturbations of self-adjoint operators, see, e.g., [2], or in connection with self-adjoint extensions of a symmetric operator with deficiency indices $(1, 1)$ [1]. Given such a symmetric operator and one fixed self-adjoint extension, then there exists a Herglotz-Nevanlinna function, the so-called Q-function (in the sense of Krein) or abstract Weyl function, such that all self-adjoint extensions can be parameterized via Krein's resolvent formula. Moreover, also the spectrum of any (minimal) extension is given in terms of (the singularities of fractional linear transformations of) this Herglotz-Nevanlinna function.

2.5.2 Exponential Representation

If f is a Herglotz-Nevanlinna function, then the function $F(z) := \operatorname{Log}(f(z))$ is also Herglotz-Nevanlinna. Since $\operatorname{Im} F$ is bounded, it follows that F has an integral representation with an absolute continuous measure and no linear term, i.e., $b = 0$. This observation leads to the following representation.

Proposition 5 *A function f is a Herglotz-Nevanlinna function if and only if there exists a real constant γ and a density ϑ such that*

$$f(z) = \exp\left(\gamma + \int_{\mathbb{R}} \left(\frac{1}{t-z} - \frac{t}{1+t^2}\right) \vartheta(t) d\lambda_{\mathbb{R}}(t)\right).$$

For details, in particular, concerning the relation between μ from (1) and ϑ, see [3] and [4].

2.6 Passive Systems

Symmetric Herglotz-Nevanlinna functions are also characterized in terms of Laplace transforms of certain distributions, see, e.g., the classical text [32].

Consider an operator R that acts on distributions $\mathcal{D}'(\mathbb{R}, \mathbb{C})$ as a convolution operator, i.e., there exists $Y \in \mathcal{D}'$ such that $R(\varphi) = Y \star \varphi$ for all $\varphi \in \mathcal{D}'$ such that this action is well-defined.

Definition 4 A convolution operator $R = Y\star$ is called (admittance-)passive if for every test function $\varphi \in \mathcal{D}$ the output $R(\varphi) =: \psi$ is locally integrable and

$$\operatorname{Re}\left[\int_{-\infty}^{t} \overline{\varphi(\tau)}\psi(\tau)d\tau\right] \geq 0, \quad \forall t \in \mathbb{R}.$$

It can be shown that every passive operator R is *causal* (i.e., supp$Y \subseteq [0, \infty)$), and it is of slow growth (i.e., $Y \in \mathcal{S}'$, where $\mathcal{S}'$ denotes the set of Schwartz distributions).

For a convolution operator that is causal and of slow growth, the Laplace transform $W := \mathcal{L}(Y)$ of its defining distribution is well-defined and holomorphic in the right halfplane, see, e.g., [32] for details.

Furthermore, a *real distribution* is a distribution that maps real test functions to real numbers, and a convolution operator is called *real* if it maps real distributions into real distributions. A holomorphic function is called *positive real* (or for short PR) if it maps the right halfplane into itself and takes real values on the real line.

Passive operators are in a one-to-one correspondence with the positive real functions in the sense of the following theorem, which, however, is formulated in terms of Herglotz-Nevanlinna functions.

Theorem 8 *Given a real passive operator $R = Y\star$, the function $f(z) := iW(\frac{z}{i})$ is a symmetric Herglotz-Nevanlinna function (where $W = \mathcal{L}(Y)$).*

Conversely, given a symmetric Herglotz-Nevanlinna function f, the convolution operator $R := \mathcal{L}^{-1}(W)\star$ for $W(s) := \frac{1}{i} f(is)$ is passive and real.

Remark 4 Here the Laplace transform W is itself a positive real function. In applications sometimes this transfer function is considered directly; see e.g., Example 10, or alternatively the Laplace transform is combined with a multiplication of i in the independent variable and is then called the Fourier–Laplace transform, as in Eq. (38) of Part II.

2.7 Asymptotic Behavior

Generally speaking, the growth of the function at a boundary point in $\mathbb{R} \cup \{\infty\}$ is closely related to the behavior of the measure at this point, e.g., (5). In this section, we demonstrate how the function's asymptotic behavior and the moments of the measure are related; see [29] for an overview and [8] for the proofs.

We start with noting that for every Herglotz-Nevanlinna function f, one has

$$f(z) = b_1 z + o(z) \qquad \text{as } z \mathrel{\hat{\to}} \infty,$$

and

$$f(z) = \frac{a_{-1}}{z} + o(\frac{1}{z}) \qquad \text{as } z \mathrel{\hat{\to}} 0,$$

where $b_1 = b$ in the integral representation (1) and $a_{-1} = -\mu(\{0\})$. Some functions do even admit expansions of higher order. We first consider expansions at ∞.

Definition 5 A Herglotz-Nevanlinna function f has an *asymptotic expansion of order* K at $z = \infty$ if for $K \geq -1$ there exist real numbers $b_1, b_0, b_{-1}, \dots, b_{-K}$ such that f can be written as

$$f(z) = b_1 z + b_0 + \frac{b_{-1}}{z} + \ldots + \frac{b_{-K}}{z^K} + o\Big(\frac{1}{z^K}\Big) \qquad \text{as } z \mathrel{\hat{\to}} \infty. \tag{10}$$

Remark 5 This means that

$$\lim_{z \mathrel{\hat{\to}} \infty} z^K \Big(f(z) - b_1 z - b_0 - \frac{b_{-1}}{z} - \ldots - \frac{b_{-K}}{z^K}\Big) = 0. \tag{11}$$

Moreover, the coefficients b_{-j} are given by

$$b_{-j} = \lim_{z \mathrel{\hat{\to}} \infty} z^j \Big(f(z) - b_1 z - b_0 - \frac{b_{-1}}{z} - \ldots - \frac{b_{-(j-1)}}{z^{j-1}}\Big). \tag{12}$$

The following theorem relates the asymptotic expansion to the moments of the measure.

Theorem 9 *Let f be a Herglotz-Nevanlinna function with representing measure μ in* (1) *and $N_\infty \geq 0$. Then f has an asymptotic expansion of order $2N_\infty + 1$ at $z = \infty$ if and only if the measure μ has finite moments up to order $2N_\infty$, i.e., $\int_{\mathbb{R}} \xi^{2N_\infty} d\mu(\xi) < \infty$. Moreover, in this case*

$$\int_{\mathbb{R}} \xi^k d\mu(\xi) = -b_{-k-1} \quad \textit{for } 0 < k \leq N_\infty. \tag{13}$$

Since these moments can be calculated by a modified version of the Stieltjes inversion formula, this result can be reformulated in the following way, known as *sum rules*. See [8] for a rigorous derivation.

Theorem 10 *Let f be a Herglotz-Nevanlinna function. Then, for some integer $N_\infty \geq 0$, the limit*

$$\lim_{\varepsilon\to 0^+}\lim_{y\to 0^+}\int_{\varepsilon<|x|<\frac{1}{\varepsilon}} x^{2N_\infty}\mathrm{Im}\, f(x+iy)dx \tag{14}$$

exists as a finite number if and only if the function f admits at $z=\infty$ an asymptotic expansion of order $2N_\infty+1$. In this case, the following sum rules hold

$$\lim_{\varepsilon\to 0^+}\lim_{y\to 0^+}\frac{1}{\pi}\int_{\varepsilon<|x|<\frac{1}{\varepsilon}} x^{n}\mathrm{Im}\, f(x+iy)dx = \begin{cases} a_{-1}-b_{-1}, & n=0 \\ -b_{-n-1}, & 0<n\leq 2N_\infty \end{cases} \tag{15}$$

Example 6 Note that the assumption that the coefficients in expansions (10) are real is essential. Consider, e.g., the function $f(z)=i$ for $z\in\mathbb{C}^+$, which admits expansions of arbitrary order if non-real coefficients are allowed. However, the limits (14) do not exist. This example also shows that not every Herglotz-Nevanlinna function does admit a sum rule.

Expansions at $z=0$ are defined analogously. This can either be done explicitly, as below, or via the expansion at ∞ for the Herglotz-Nevanlinna function $\widetilde{f}(z):=f(-1/z)$. The above remark applies then accordingly.

Definition 6 A Herglotz-Nevanlinna function f has an *asymptotic expansion of order K* at $z=0$ if for $K\geq -1$ there exist real numbers $a_{-1},a_0,a_1,\ldots,a_K$ such that f can be written as

$$f(z)=\frac{a_{-1}}{z}+a_0+a_1z+\ldots+a_Kz^K+o\big(z^K\big) \qquad \text{as } z\hat{\rightarrow}0. \tag{16}$$

Theorem 11 *Let f be a Herglotz-Nevanlinna function. Then, for some integer $N_0\geq 1$, the limit*

$$\lim_{\varepsilon\to 0^+}\lim_{y\to 0^+}\int_{\varepsilon<|x|<\frac{1}{\varepsilon}} \frac{\mathrm{Im}\, f(x+iy)}{x^{2N_0}}dx \tag{17}$$

exists as a finite number if and only if f admits at $z=0$ an asymptotic expansion of order $2N_0-1$. In this case, the following sum rules hold

$$\lim_{\varepsilon\to 0^+}\lim_{y\to 0^+}\frac{1}{\pi}\int_{\varepsilon<|x|<\frac{1}{\varepsilon}} \frac{\mathrm{Im}\, f(x+iy)}{x^{p}}dx = \begin{cases} a_1-b_1, & p=2 \\ a_{p-1}, & 2<p\leq 2N_0 \end{cases}. \tag{18}$$

Example 7 The Herglotz-Nevanlinna function $f(z) = \tan(z)$ has the asymptotic expansion

$$\tan(z) = z + \frac{z^3}{3} + \frac{2z^5}{15} + \ldots \text{ as } z \hat{\to} 0 \tag{19}$$

and $\tan(z) = i + o(1)$ as $z \hat{\to} \infty$ (which, however, is not an asymptotic expansion in the sense of (10)). We thus find that $a_1 = 1$, $a_3 = 1/3$, $a_5 = 2/15$, and $b_1 = 0$ (whereas b_0 does not exist), and hence, the following sum rules apply.

$$\lim_{\epsilon \to 0^+} \lim_{y \to 0^+} \frac{1}{\pi} \int_{\epsilon \le |x| \le 1/\epsilon} \frac{\operatorname{Im} \tan(x+iy)}{x^p} dx = \begin{cases} 1 & p = 2 \\ 1/3 & p = 4 \\ 2/15 & p = 6 \end{cases} . \tag{20}$$

Remark 6 Note that the case of $p = 1$ is not included in Theorem 11. In order to guarantee this limit to be finite, it is required that f admits asymptotic expansions of order 1 at both $z = \infty$ and $z = 0$. In this case, the limit equals $a_0 - b_0$.

Remark 7 Note that the exponents in (14) and (17) are even. A corresponding statement for odd exponents, meaning that the existence of the limit is equivalent to the existence of the expansion, does not hold. A counterexample is given in [8, p. 9].

Remark 8 The counterpart of Theorem 9 for the operator representation (9) is $v \in \operatorname{dom}(A^{N_\infty})$ if and only if an asymptotic expansion of order $2N_\infty + 1$ at $z = \infty$ exists.

For symmetric Herglotz-Nevanlinna functions (8), the non-zero coefficients of odd and even order in an asymptotic expansion are necessarily real-valued and purely imaginary, respectively, and hence expansions (16) and (10) stop at the appearance of the first imaginary term, or the first non-existing term. If the assumptions in both Theorems 10 and 11 are satisfied, i.e., that both asymptotic expansions exist up to order $2N_0 - 1$ and $2N_\infty + 1$, respectively, these together with Remark 6 can be summarized as

$$\frac{2}{\pi} \int_{0^+}^{\infty} \frac{\operatorname{Im} f(x)]}{x^{2n}} dx := \lim_{\varepsilon \to 0^+} \lim_{y \to 0^+} \frac{2}{\pi} \int_{\varepsilon}^{1/\varepsilon} \frac{\operatorname{Im} h(x+iy)]}{x^{2n}} dx = a_{2n-1} - b_{2n-1} \tag{21}$$

for $n = -N_\infty, \ldots, N_0$.

2.8 Matrix- and Operator-Valued Herglotz-Nevanlinna Functions

So far in this text the values of the functions considered have been complex numbers, but much of the theory can be extended to matrix- or even operator-valued functions; see [15] for a detailed overview.

Let $\mathcal{H}_0$ be a complex Hilbert space and denote by $\mathcal{L}(\mathcal{H}_0)$ and $\mathcal{B}(\mathcal{H}_0)$ the spaces of linear and bounded linear operators in $\mathcal{H}_0$, respectively. In case of finite-dimensional $\mathcal{H}_0$, say $\dim\mathcal{H}_0 = n$, these two spaces coincide and are identified with the space of matrices $\mathbb{C}^{n\times n}$. For $T \in \mathcal{L}(\mathcal{H}_0)$, we denote by T^* the adjoint operator; for $T \in \mathbb{C}^{n\times n}$, this is the conjugate transpose of the matrix T.

Definition 7 A function $F : \mathbb{C}^+ \to \mathcal{B}(\mathcal{H}_0)$ is called Herglotz-Nevanlinna if it is analytic and $\operatorname{Im} F(z) \geq 0$ for $z \in \mathbb{C}^+$, where $\operatorname{Im} F(z) := \frac{1}{2i}(F(z) - F(z)^*)$.

Also these functions can be represented via an integral representation as in Theorem 1.

Theorem 12 *A function $F : \mathbb{C}^+ \to \mathcal{B}(\mathcal{H}_0)$ is a Herglotz-Nevanlinna function if and only if there are operators $C = C^*$ and $D \geq 0 \in \mathcal{B}(\mathcal{H}_0)$ and a (positive) $\mathcal{B}(\mathcal{H}_0)$-valued Borel measure Ω with $\int_{\mathbb{R}} \frac{1}{1+\xi^2} d\,(\Omega(\xi)\mathbf{x}, \mathbf{x})_{\mathcal{B}(\mathcal{H}_0)} < \infty$ for all $\mathbf{x} \in \mathcal{H}_0$ such that*

$$F(z) = C + Dz + \int_{\mathbb{R}} \left(\frac{1}{\xi - z} - \frac{\xi}{1+\xi^2} \right) d\Omega(\xi). \tag{22}$$

Moreover, C, D, and Ω are unique with this property.

Here an operator-valued measure is defined via a non-decreasing operator-valued (distribution) function; see [15].

Remark 9 As in Theorems 2 and 3, the representation simplifies under certain growth conditions. More precisely, these theorems hold true even in the operator-valued case if the growth conditions are considered weakly, e.g., (i) in Theorem 2 becomes

$$\int_0^\infty \frac{(\operatorname{Im} F(iy)\mathbf{x}, \mathbf{x})_{\mathcal{H}_0}}{y} dy \leq \infty$$

for all $\mathbf{x} \in \mathcal{H}_0$. Also the results in Sect. 2.3 hold in this weak sense.

Also the operator representations can be extended to this case.

Theorem 13 *A function $F : \mathbb{C}^+ \to \mathcal{B}(\mathcal{H}_0)$ is a Herglotz-Nevanlinna function if and only if there exist a Hilbert space $\mathcal{H}$, a self-adjoint linear relation A, a point $z_0 \in \mathbb{C}^+$, and a map $\Gamma \in \mathcal{B}(\mathcal{H}_0, \mathcal{H})$ such that*

$$F(z) = F(z_0)^* + (z - \overline{z_0})\Gamma^*(I + (z - z_0)(A - z)^{-1})\Gamma. \tag{23}$$

Moreover, if $\mathcal{H} = \overline{\operatorname{span}}\{(I + (z - z_0)(A - z)^{-1})\Gamma\mathbf{x} : z \in \varrho(A)$ *and* $\mathbf{x} \in \mathcal{H}_0\}$*, then the representation is called minimal. In this case, the representation is unique up to unitary equivalence.*

For scalar functions, i.e., $\mathcal{H}_0 = \mathbb{C}$, the linear mapping $\Gamma : \mathbb{C} \to \mathcal{H}$ acts as $1 \mapsto v$, where v is the element in the scalar representation Theorem 6.

Similarly as in Theorems 2 and 3, certain assumptions on the growth of the function F guarantee simplified representations. As an example, we give one result, which will be used in Sect. 2.4 of Part II.

Theorem 14 *Let* $F : \mathbb{C}^+ \to \mathcal{B}(\mathcal{H}_0)$ *be a Herglotz-Nevanlinna function with representation* (23)*. Then*

$$\lim_{z \hat{\to} \infty} \frac{\|F(z)\|}{\operatorname{Im} z} = 0 \textit{ and } \limsup_{z \hat{\to} \infty} |z| \cdot \|\operatorname{Im} F(z)\| < \infty$$

if and only if A *is an operator and* $ran\Gamma \subset \operatorname{dom}(A)$*. In this case,*

$$F(z) = S + \Gamma_0^*(A - z)^{-1}\Gamma_0 \tag{24}$$

with $\Gamma_0 := (A - \overline{z_0})\Gamma$ *and* $S = S^* \in \mathcal{B}(\mathcal{H}_0)$*.*

In particular, this theorem implies the following corollary.

Corollary 1 *For a Herglotz-Nevanlinna function* $F : \mathbb{C}^+ \to \mathcal{B}(\mathcal{H}_0)$ *the growth condition* $\limsup\limits_{y \to \infty} y\|F(iy)\| < \infty$ *implies that*

$$F(z) = \Gamma_0^*(A - z)^{-1}\Gamma_0, \tag{25}$$

where A *is a self-adjoint operator in a Hilbert space* $\mathcal{H}$ *and* $\Gamma_0 \in \mathcal{B}(\mathcal{H}_0, \mathcal{H})$*. Moreover, there exists a minimal representation, that is, a representation for which it holds* $\mathcal{H} = \overline{\operatorname{span}}\{(A - z)^{-1})\Gamma_0\mathbf{x} : z \in \varrho(A)$ *and* $\mathbf{x} \in \mathcal{H}_0\}$*, that is unique up to unitary equivalence.*

Example 8 Both the functions

$$F(z) = \begin{pmatrix} z & 1 \\ 1 & -\frac{1}{z} \end{pmatrix} \quad \text{and} \quad \tilde{F}(z) := -F(z)^{-1} = \frac{1}{2} \cdot \begin{pmatrix} -\frac{1}{z} & -1 \\ -1 & z \end{pmatrix}$$

are Herglotz-Nevanlinna functions.

The above example illustrates a general phenomenon for matrix (and operator) functions, namely, the point $z = 0$ is both a pole and a zero of F; it is also a pole of the inverse F^{-1}. In particular, $\det F(z) \equiv -2$, and hence, the poles of F cannot be read off from the scalar function det $F(z)$, but the matrix structure has to be taken into account.

Whereas scalar Herglotz-Nevanlinna functions do appear in connection with extensions of symmetric operators with deficiency index 1, higher defect leads to matrix-valued functions (for finite deficiency index) or operator-valued functions (for infinite deficiency index). As an example, consider differential operators. If such an operator acts on functions defined on the halfline $\mathbb{R}^+$ (which has only one boundary point, $x = 0$), then the minimal operator will in general have deficiency index 1, and hence, the corresponding Titchmarsh–Weyl function is a scalar Herglotz-Nevanlinna function. If however, one considers either a compact interval (with 2 boundary points) or differential operators on finite graphs (with finitely many boundary points), the corresponding Weyl function is a matrix-valued Herglotz-Nevanlinna function, where the number of boundary points determines its size. Partial differential operators defined on some domain in $\mathbb{R}^n$ (with boundary that consists of infinitely many points) give rise to operator-valued Herglotz-Nevanlinna functions. See, e.g., the recent books [5, 28] and references therein.

Other examples for matrix-valued Herglotz-Nevanlinna functions do appear, e.g., in connection with array antennas [24].

3 Applications

In this section, as well as in Part II, we give examples of applications, where Herglotz-Nevanlinna functions are utilized. They stem from quite different areas, but in terms of the underlying mathematics, they have a lot in common. Here, we focus on applications in electromagnetics and techniques that are related to the sum rules. As is mentioned in the introduction, there are also applications where the functions depend on the contrast of materials rather than frequency; see Sect. 2.1 of Part II. Here we want to point out these similarities in an informal way, and more precise definitions are then given in the respective application below or in Part II.

First of all, the description of most of the problems in some way involves a *convolution* operator. This might be related to time invariance (also called time homogeneity), or it can appear as a memory term or a time-dispersive integral term.

Another common feature is *causality*, which means that the current state depends only on the time evolution in the past but not on the future. Mathematically, causality amounts to the fact that the convolution kernel is supported on one halfline only, which implies that its Fourier (or Laplace) transform is an analytic function, in the upper (or right) halfplane. In the applications with contrast, the analyticity arises from the coercivity of a certain sesquilinear form.

In general, the analytic functions given in this way will not be Herglotz-Nevanlinna, but an additional assumption is needed. This might be, e.g., *passivity* or power *dissipation*, which imposes a sign restriction on the imaginary (or real) part, and this is how Herglotz-Nevanlinna functions appear. In many situations, there is a one-to-one correspondence between the systems and the Herglotz-Nevanlinna functions describing them.

In the following sections as well as in Part II, we summarize results from different areas and try to make their connections to the mathematical background in Sect. 2 more explicit. We try to use the notations as close as possible to the original papers in order to make them more accessible to the reader. Unfortunately, this leads to unavoidable clashes in some notations, which we will point out explicitly if the context there is not enough to resolve the ambiguity of notation.

3.1 Sum Rules and Physical Bounds in Electromagnetics

In Sect. 2.6, the mathematical definition of passive systems was given, and it was explained that such systems are in one-to-one correspondence with symmetric Herglotz-Nevanlinna functions. Here we are going to give a physical motivation including an example from electromagnetics and demonstrate how the sum rules are used to derive physical bounds. We are following closely the exposition in [29], where also additional references can be found.

Physical objects that cannot produce energy are usually considered as passive. However, whether a system is passive or not (in the mathematical sense) depends very much on the definition of the input and the output.

More precisely, consider one-port systems. These are systems consisting of one input and one output parameter, which can be measured at the so-called *ports* of these systems. As an example, one might think of an electric circuit with two nodes to which one can input a signal, e.g., a current, and measure a voltage.

The one-port systems we consider here are assumed to be linear, continuous, and time-translationally invariant. Hence, the system is in convolution form [32], i.e., if $u(t)$ denotes the input, then the output $v(t)$ is given by

$$v(t) = (w \star u)(t) := \int_{\mathbb{R}} w(\tau)u(t-\tau)d\tau, \tag{26}$$

with impulse response $w(t)$. As before, we restrict ourselves to real-valued systems, i.e., the systems where the impulse response w is real-valued. One way to define passivity for such systems is the so-called admittance passivity defined in Definition 4 [31, 32], where

$$\mathcal{W}_{\text{adm}}(T) := \operatorname{Re} \int_{-\infty}^{T} v(t)\overline{u(t)}dt \geq 0 \tag{27}$$

for all $T \in \mathbb{R}$ and all $u \in C_0^\infty$ (i.e., smooth functions with compact support).

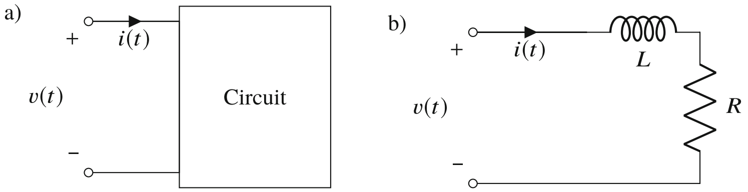


Fig. 1 (**a**) A general electric circuit; (**b**) A simple circuit example

Here, $\mathcal{W}_{\text{adm}}(T)$ represents all energy the system has absorbed until time T, and hence, this definition means that the system absorbs more energy than it emits, or in other words, the system does not produce energy.

It can be shown [32] that the impulse response w of a passive system has the representation

$$w(t) = b\delta'(t) + H(t)\int_{\mathbb{R}} \cos(\xi t)d\mu(\xi), \tag{28}$$

where $b \geq 0$, δ' denotes the derivative of the Dirac distribution, H the Heaviside step function, and μ a Borel measure satisfying the growth condition from Theorem 1. This implies that the Laplace transform of the impulse response (28), $W(s)$, gives rise to a symmetric Herglotz-Nevanlinna function, cf., Theorem 8, which has exactly the parameters b and μ.

Let us have a closer look at a few examples of passive systems in electromagnetics from [29].

Example 9 (Input Impedance of Electrical Circuit Networks) Consider a simple electric one-port circuit containing passive components, i.e., each resistance R, inductance L, and capacitance C are positive. The input signal to this system is the real-valued electric current $i(t)$, and its output signal is the voltage $v(t)$, see Fig. 1a. As an explicit example, consider the simple circuit in Fig. 1b. In order to check that this system is passive, we calculate $\mathcal{W}_{\text{adm}}(T)$ from (27).

For a given input current $i(t)$, the output voltage is given by $v(t) = L\frac{d\,i(t)}{dt} + Ri(t)$ and can be written as $v = w\star i$, where $w = L\delta' + R\delta$ is the impulse response. Hence, the integral (27) becomes

$$\mathcal{W}_{\text{adm}}(T) = \int_{-\infty}^{T}\left(L\frac{d\,i(t)}{dt}i(t) + Ri(t)^2\right)dt = \frac{L}{2}i(T)^2 + R\int_{-\infty}^{T} i(t)^2 dt \geq 0, \tag{29}$$

and the system is admittance-passive. The transfer function (i.e., here the input impedance), which by definition is the Laplace transform of the impulse response, becomes, in this case, the positive real (PR)-function

$$Z_{\text{in}}(s) = sL + R, \tag{30}$$

and hence, $f(z) := iZ_{\text{in}}(-is)$ is a Herglotz-Nevanlinna function. This simple example generalizes to circuit networks composed of arbitrary number and combinations of passive resistors, capacitances, and inductances resulting in rational PR functions [19]. Moreover, it is straightforward to include transformers and transmission lines as well as multiple input and output systems resulting in matrix-valued PR functions [11].

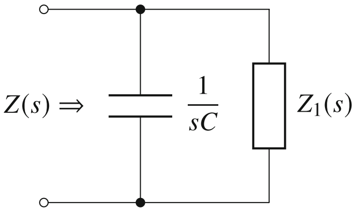
Given a Herglotz-Nevanlinna function, the integral identities in Theorems 10 and 11 have been applied in order to derive physical bounds on passive systems, see e.g., [8]. In the engineering and physics literature, these integral identities appear in various forms and special cases and are also often referred to as *sum rules* [8, 26].

For Herglotz-Nevanlinna functions, the integral identities are given on the real axis where $z = x$ is often interpreted as angular frequency ω (in rad/s), wave number $k = \omega/c_0$ (in m^{-1}), or as wavelength $\lambda = 2\pi/k$ (in m).

In many practical electromagnetic applications, it is reasonable to assume some partial knowledge regarding the low- and/or high-frequency asymptotic expansions of the corresponding Herglotz-Nevanlinna function, such as the static and the optical responses of a material, or a structure. In these cases, the sum rules can be used to obtain inequalities by constraining the integration interval to a finite bandwidth in the frequency (or wavelength) domain and thereby yielding useful physical limitations in a variety of applications.

As illustration, we treat the following classical example by applying the theory presented in Sect. 2.7, even though residue calculus could also be used to solve this problem.

Example 10 (The Resistance-Integral Theorem) Consider a passive circuit consisting of a parallel connection of a capacitance C and an impedance $Z_1(s)$ that does not contain a shunt capacitance (i.e., $Z_1(0)$ is finite and $Z_1(\infty) \neq 0$), see the figure besides. Then the input impedance of this circuit is given by $Z(s) = 1/(sC + 1/Z_1(s))$, which is a PR function in the Laplace variable $s \in \mathbb{C}_+$, and hence, the system is admittance-passive.



The asymptotic expansions are $Z(s) = Z_1(0) + o(s)$ as $s \hat{\to} 0$ and $Z(s) = 1/(sC) + o(s^{-1})$ as $s \hat{\to} \infty$. Here, the corresponding Herglotz-Nevanlinna function is $h(\omega) := iZ(-i\omega)$ for $\omega \in \mathbb{C}^+$. Its low- and high-frequency asymptotics are

$$h(\omega) = o(\omega^{-1}) \text{ as } \omega \hat{\to} 0 \text{ and } h(\omega) = -\frac{1}{\omega C} + o(\omega^{-1}) \text{ as } \omega \hat{\to} \infty. \qquad (31)$$

In terms of (16) and (10), we have $a_{-1} = 0$ and $b_{-1} = -1/C$, and thus the sum rule (21) with $n = 0$ gives

$$\frac{2}{\pi}\int_{0+}^{\infty} \mathrm{Re}\,[Z(-i\omega)]d\omega = \frac{2}{\pi}\int_{0+}^{\infty} \mathrm{Im}\,[h(\omega)]d\omega = a_{-1} - b_{-1} = \frac{1}{C}. \tag{32}$$

By integrating only over a finite frequency interval $\Omega := [\omega_1, \omega_2]$, and estimating this integral from below, we obtain the bound

$$\Delta\omega \inf_{\omega\in\Omega} \mathrm{Re}\,[Z(-i\omega)] \leq \int_{0+}^{\infty} \mathrm{Re}\,[Z(-i\omega)]d\omega = \frac{\pi}{2C}, \tag{33}$$

where $\Delta\omega := \omega_2 - \omega_1$. Consequently, inequality (33) limits the product between the bandwidth and the minimum resistance over the given frequency interval; see also [9].

Compositions of Herglotz-Nevanlinna functions can be used to construct new Herglotz-Nevanlinna functions and, hence, also new sum rules, cf., also Sect. 2.3 in Part II. Here, we illustrate this for a case where the minimal temporal dispersion for metamaterials is determined, by first transforming the problem into the question of determining the minimum amplitude of a Herglotz-Nevanlinna function over a bandwidth [8, 20].

When a dielectric medium is specified to have inductive properties (i.e., has negative permittivity) over a given bandwidth, it is regarded as a metamaterial. A given negative permittivity value at a single frequency is always possible to achieve. For instance, the plasmonic resonances in small metal particles can be explained by, e.g., using Drude or Lorentz models. However, when a constant negative permittivity value is prescribed over a given bandwidth, the passivity of the material will imply severe bandwidth limitations, see e.g., [20].

To derive these limitations based on Herglotz-Nevanlinna functions, we start by considering the following general situation: Let h_0 be a fixed Herglotz-Nevanlinna function that can be extended continuously to a neighborhood of the compact interval $\Omega \subset \mathbb{R}$ and has the large argument asymptotics $h_0(z) = b_1^0 z + o(z)$ as $z\hat{\to}\infty$. Denote by $F(x) := -h_0(x)$ the negative of h_0. We are now looking for a Herglotz-Nevanlinna function h that has the same continuity property on the real line as h_0 and with an asymptotic expansion $h(z) = b_1 z + o(z)$ as $z\hat{\to}\infty$ and lies as close as possible to the given anti-Herglotz function F. In particular, we aim to derive a lower bound for the error norm

$$\|h - F\|_{L^\infty(\Omega)} := \sup_{x\in\Omega} |h(x) - F(x)|. \tag{34}$$

To this end, the following auxiliary Herglotz-Nevanlinna function $h_\Delta(z)$, for $\Delta > 0$, is used

$$h_{\Delta}(z) := \frac{1}{\pi}\int_{-\Delta}^{\Delta}\frac{1}{\xi - z}d\xi = \frac{1}{\pi}\mathrm{Log}\frac{z-\Delta}{z+\Delta} = \begin{cases} i + o(1) & \text{as } z\hat{\to}0 \\ \dfrac{-2\Delta}{\pi z} + o(z^{-1}) & \text{as } z\hat{\to}\infty. \end{cases} \tag{35}$$

Note that $\mathrm{Im}\, h_{\Delta}(z) \geq \frac{1}{2}$ for $|z| \leq \Delta$ and $\mathrm{Im}\, z \geq 0$. Next, consider the composite Herglotz-Nevanlinna function $h_1(z) := h_{\Delta}\big(h(z) + h_0(z)\big)$. Since $h(z) + h_0(z) = (b_1 + b_1^0)z + o(z)$ as $z\hat{\to}\infty$ the new function h_{Δ} has the asymptotic expansions

$$h_1(z) = o(z^{-1}) \text{ as } z\hat{\to}0 \text{ and } h_1(z) = \frac{-2\Delta}{\pi(b_1 + b_1^0)}z^{-1} + o(z^{-1}) \text{ as } z\hat{\to}\infty. \tag{36}$$

Then the sum rule (21) with $n = 0$ becomes

$$\frac{2}{\pi}\int_{0+}^{\infty}\mathrm{Im}\, h_1(x)dx = a_{-1} - b_{-1} = \frac{2\Delta}{\pi(b_1 + b_1^0)}. \tag{37}$$

Choosing $\Delta := \sup_{x\in\Omega}|h(x) + h_0(x)|$, the following integral inequalities follow

$$\frac{1}{\pi}|\Omega| \leq \frac{2}{\pi}\int_{\Omega}\underbrace{\mathrm{Im}\, h_1(x)}_{\geq\frac{1}{2}}dx \leq \frac{2}{\pi}\int_{0+}^{\infty}\mathrm{Im}\, h_1(x)dx = \frac{2\sup_{x\in\Omega}|h(x) + h_0(x)|}{\pi(b_1 + b_1^0)} \tag{38}$$

or

$$\|h + h_0\|_{L^{\infty}(\Omega)} \geq (b_1 + b_1^0)\frac{1}{2}|\Omega|, \text{ where } |\Omega| = \int_{\Omega}dx. \tag{39}$$

Example 11 (Metamaterials and Temporal Dispersion) Consider now a dielectric metamaterial with a constant, real-valued, and negative target permittivity $\epsilon_{\mathrm{t}} < 0$ to be approximated over an interval Ω. In this case, the function of interest is $F(z) = z\epsilon_{\mathrm{t}}$, and hence, we have $h_0(z) = -F(z)$ with $b_1^0 = -\epsilon_{\mathrm{t}}$. Let $\epsilon(z)$ be the permittivity function of the approximating passive dielectric material, and $h(z) = z\epsilon(z)$ the corresponding Herglotz-Nevanlinna function with $b_1 = \epsilon_{\infty}$, the assumed high-frequency permittivity of the material, and the approximation interval $\Omega = \omega_0[1 - B/2, 1 + B/2]$, where ω_0 is the center frequency and B the relative bandwidth with $0 < B < 2$. The resulting physical bound obtained from (39) is given by

$$\|\epsilon(\cdot) - \epsilon_{\mathrm{t}}\|_{L^{\infty}(\Omega)} \geq \frac{(\epsilon_{\infty} - \epsilon_{\mathrm{t}})B}{2 + B}. \tag{40}$$

Note that the variable x corresponds here to angular frequency, also commonly denoted as ω (in rad/s).

Other applications are related to scattering passive systems, see, e.g., [8, 32] for a precise definition. Scattering passive systems have transfer functions that map $\mathbb{C}^+$ to the unit disk. To use (21), one then first constructs a Herglotz-Nevanlinna function by mapping the unit disk to $\mathbb{C}^+$. This map can be made in many different ways, and the particular choice depends on the asymptotic expansion and the physical interpretation of the system. The Cayley transform, logarithm, and addition are most common in applications. For example, see, e.g., [8].

3.2 Physical Bounds via Convex Optimization

In this section, it is exemplified how Herglotz-Nevanlinna functions can be used to identify or approximate passive systems with given properties. This approach is based on convex optimization related to the function integral representation.

To facilitate the computation of a numerical solution using a software such as, e.g., CVX [17], it is necessary to first impose some a priori constraints on the class of approximating Herglotz-Nevanlinna functions. In view of Sect. 2.3, we restrict ourselves here to approximating Herglotz-Nevanlinna functions that are locally Hölder continuous on some given intervals on the real line.

A passive approximation problem is considered where the target function F is an arbitrary complex-valued continuous function defined on an approximation domain $\Omega \subset \mathbb{R}$ consisting of a finite union of closed and bounded intervals of the real axis. The norms used, denoted by $\|\cdot\|_{L^p(w,\Omega)}$, are weighted $L^p(\Omega)$-norms with a positive continuous weight function w on Ω, and where $1 \leq p \leq \infty$.

Here for any approximating function h, we assume that it is the Hölder continuous extension (to Ω) of some Herglotz-Nevanlinna function generated by an absolutely continuous measure μ having a density μ' that is Hölder continuous on the closure $\overline{U}$ of an arbitrary neighborhood $U \supset \Omega$ of the approximation domain. Then, cf., Proposition 3, both the real and imaginary parts of h are continuous functions on Ω. Moreover, it holds that $\operatorname{Im} h(x) = \pi \mu'(x)$ on $\overline{U}$, and the real part is given by the associated Hilbert transform. As we consider real systems only, the approximating Herglotz-Nevanlinna function h can be assumed to be symmetric, and its real part hence admits the representation

$$\operatorname{Re} h(x) = bx + p.v. \int_{\mathbb{R}} \frac{\mu'(\tau)}{\tau - x} d\tau \qquad \text{for } x \in \Omega, \tag{41}$$

where $p.v.$ denotes the principal values both at ∞ and x.

The continuity of h on Ω implies that the norm $\|h\|_{L^p(w,\Omega)}$ is well-defined for $1 \leq p \leq \infty$.

If approximating the function F by Herglotz-Nevanlinna functions h on Ω, one is interested in the greatest lower bound on the approximation error by

$$d := \inf_{h} \|h - F\|_{L^p(w,\Omega)}, \tag{42}$$

where the infinum is taken over all Herglotz-Nevanlinna functions h generated by a measure having a Hölder continuous density on $\overline{U}$.

In general, a best approximation achieving the bound d in (42) does not exist. In practice, however, the problem is approached by using numerical algorithms such as CVX, solving finite-dimensional approximation problems using, e.g., B-splines, with the number of basis functions N fixed during the optimization, cf. [22, 30]. Here, a B-spline of order $m \geq 2$ is an $m-2$ times continuously differentiable and compactly supported positive basis spline function consisting of piecewise polynomial functions of order $m-1$, i.e., linear, quadratic, cubic, etc., and which is defined by $m+1$ breakpoints [13]. For the density $\operatorname{Im} h(x)$ of the approximating symmetric function h, here it is made the ansatz of a finite B-spline expansion

$$\pi\mu'(x) = \sum_{n=1}^{N} \zeta_n \left(p_n(x) + p_n(-x)\right) \tag{43}$$

for $x \in \mathbb{R}$, where ζ_n are optimization variables for $n = 1, \ldots, N$, and $p_n(x)$ are B-spline basis functions of fixed order m that are defined on the given partition. The real part $\operatorname{Re} h(x)$ for $x \in \Omega$ is then given by (41) and can be expressed as

$$\operatorname{Re} h(x) = bx - \frac{\zeta_0}{x} + \sum_{n=1}^{N} \zeta_n \left(\hat{p}_n(x) - \hat{p}_n(-x)\right), \quad x \in \Omega, \tag{44}$$

where $\hat{p}_n(x)$ is the (negative) Hilbert transform of the B-spline function $p_n(x)$ and where a point mass at $x = 0$ with amplitude c_0 has been included. Any other a priori assumed point masses can be included in a similar way.

Consider now the following convex optimization problem:

$$\begin{aligned} \text{minimize} \quad & \|h - F\|_{L^p(w,\Omega)} \\ \text{subject to} \quad & \zeta_n \geq 0, \ \text{for}\ n = 0, \ldots N, \\ & b \geq 0, \end{aligned} \tag{45}$$

where the optimization is over the variables $(\zeta_0, \zeta_1, \ldots, \zeta_N, b)$. Note that the objective function in (45) above is the norm of an affine form in the optimization variables. Hence, the objective function is a convex function in the variables $(\zeta_0, \zeta_1, \ldots, \zeta_N, b)$.

The uniform continuity of all functions involved implies that the solution to (45) can be approximated within an arbitrary accuracy by discretizing the approximation domain Ω (and the computation of the norm) using only a finite number of sample points. The corresponding numerical problem (45) can now be solved efficiently by using the CVX Matlab software for disciplined convex programming. The convex optimization formulation (45) offers a great advantage in the flexibility in which additional or alternative convex constraints and formulations can be implemented; see also [22, 30].

Example 12 A canonical example for convex optimization is passive approximation of metamaterials; see also [20, 22, 30]. As in Example 11, the variable x corresponds here to angular frequency, also commonly denoted as ω (in rad/s). A typical application is with the study of optimal plasmonic resonances in small structures (or particles) for which the absorption cross section can be approximated by

$$\sigma_{\mathrm{abs}} \approx k \mathrm{Im}\, \gamma, \tag{46}$$

where $k = 2\pi/\lambda$ is the wave number, λ is the wavelength and where γ is the electric polarizability of the particle; see [10]. As, e.g., the polarizability of a dielectric sphere with radius a is given by $\gamma(x) = 4\pi a^3(\epsilon(x) - 1)/(\epsilon(x) + 2)$, where $\epsilon(x)$ is the permittivity function of the dielectric material inside the sphere.

A surface plasmon resonance is obtained when $\epsilon(x) \approx -2$, and, hence, we specify that the target permittivity of our metamaterial is $\epsilon_{\mathrm{t}} = -2$. However, a metamaterial with a negative real part cannot, in general, be implemented as a passive material over a given bandwidth, cf. [21]. Based on the theory of Herglotz-Nevanlinna functions and associated sum rules, the physical bound in (40) can be derived, where ϵ_∞ is the high-frequency permittivity of the material, $\epsilon_{\mathrm{t}} < \epsilon_\infty$, $\Omega = \omega_0[1 - B/2, 1 + B/2]$, ω_0 the center frequency, and B the relative bandwidth with $0 < B < 2$, cf. [21]. The convex optimization formulation (45) can be used to study passive realizations (43) and (44) that satisfy the bound (40) as close as possible. Here, the approximating Herglotz-Nevanlinna function is $h(x) = x\epsilon(x)$, the target function $F(x) = x\epsilon_{\mathrm{t}}$, ζ_0 the amplitude of a point mass at $x = 0$, $b = \epsilon_\infty$, and a weighted norm is used defined by $\|f\|_{L^\infty(w,\Omega)} = \max_{x\in\Omega} |f(x)/x|$ assuming that $0 \notin \Omega$. For numerical examples of these kinds of approximations as well as with non-passive systems employing quasi-Herglotz functions (Sect. 3.1 in Part II), see [22, 23, 29].

References

1. N.I. Akhiezer, I.M. Glazman, *Theory of Linear Operators in Hilbert Space*, vol. 1 (Dover Publications, 1993)
2. S. Albeverio, P. Kurasov, *Singular Perturbations of Differential Operators*, volume 271 of London Mathematical Society Lecture Note Series (Cambridge University Press, Cambridge, 2000). Solvable Schrödinger type operators
3. N. Aronszajn, W.F. Donoghue, On exponential representations of analytic functions. J. Analyse Math. **5**, 321–388 (1956)
4. N. Aronszajn, W.F. Donoghue, A supplement to the paper on exponential representations of analytic functions in the upper half-plane with positive imaginary part. J. Analyse Math. **12**, 113–127 (1964)
5. J. Behrndt, S. Hassi, H. de Snoo, *Boundary Value Problems, Weyl Functions, and Differential Operators*, volume 108 of Monographs in Mathematics (Birkhäuser/Springer, Cham, 2020)

6. C. Berg, H.L. Pedersen, Pick functions related to the gamma function. volume 32, pages 507–525. 2002, in *Conference on Special Functions* (Tempe, AZ, 2000)
7. C. Berg, H.L. Pedersen, A one-parameter family of Pick functions defined by the gamma function and related to the volume of the unit ball in n-space. Proc. Amer. Math. Soc. **139**(6), 2121–2132 (2011)
8. A. Bernland, A. Luger, M. Gustafsson, Sum rules and constraints on passive systems. J. Phys. A Math. Theor. **44**(14), 145205 (2011)
9. H.W. Bode, *Network Analysis and Feedback Amplifier Design* (Van Nostrand, 1945)
10. C.F. Bohren, D.R. Huffman, *Absorption and Scattering of Light by Small Particles* (John Wiley & Sons, 1983)
11. H. Carlin, D. Youla, L. Castriota, Bounded real scattering matrices and the foundations of linear passive network theory. IRE Trans. Circuit Theory **6**(1), 102–124 (1959)
12. W. Cauer, The Poisson integral for functions with positive real part. Bull. Amer. Math. Soc. **38**(10), 713–717 (1932)
13. C. de Boor, On calculating with B-splines. J. Approx. Theory **6**, 50–62 (1972)
14. A. Dijksma, H.S.V. de Snoo, Symmetric and selfadjoint relations in Kreĭn spaces. I, in *Operators in Indefinite Metric Spaces, Scattering Theory and Other Topics (Bucharest, 1985)*, volume 24 of Oper. Theory Adv. Appl. (Birkhäuser, Basel, 1987), pp. 145–166
15. F. Gesztesy, E. Tsekanovskii, On matrix-valued Herglotz functions. Math. Nachr. **218**(1), 61–138 (2000)
16. F. Gesztesy, M. Zinchenko, On spectral theory for Schrödinger operators with strongly singular potentials. Math. Nachr. **279**(9–10), 1041–1082 (2006)
17. M. Grant, S. Boyd, *CVX: A System for Disciplined Convex Programming, Release 2.0* (CVX Research, Inc., Austin, 2012)
18. D.S. Greenstein, On the analytic continuation of functions which map the upper half plane into itself. J. Math. Anal. Appl. **1**, 355–362 (1960)
19. E.A. Guillemin, *Synthesis of Passive Networks* (John Wiley & Sons, 1957)
20. M. Gustafsson, D. Sjöberg, Sum rules and physical bounds on passive metamaterials. New J. Phys. **12**(4), 043046 (2010)
21. M. Gustafsson, D. Sjöberg, Physical bounds and sum rules for high-impedance surfaces. IEEE Trans. Antennas Propag. **59**(6), 2196–2204 (2011)
22. Y. Ivanenko, M. Gustafsson, B.L.G. Jonsson, A. Luger, B. Nilsson, S. Nordebo, J. Toft, Passive approximation and optimization using B-splines. SIAM J. Appl. Math. **79**(1), 436–458 (2019)
23. Y. Ivanenko, M. Nedic, M. Gustafsson, B.L.G. Jonsson, A. Luger, S. Nordebo, Quasi-Herglotz functions and convex optimization. R. Soc. Open Sci. **7**, 191541 (2020)
24. B.L.G. Jonsson, C.I. Kolitsidas, N. Hussain, Array antenna limitations. Antennas Wirel. Propag. Lett, IEEE **12**, 1539–1542 (2013)
25. I.S. Kac, M.G. Krein, R-functions-analytic functions mapping the upper halfplane into itself. AMS Transl. **103**, 1–18 (1974)
26. F.W. King, *Hilbert Transforms. Vol. 2*, volume 125 of Encyclopedia of Mathematics and its Applications (Cambridge University Press, Cambridge, 2009)
27. M.G. Kreĭn, H. Langer, Über einige Fortsetzungsprobleme, die eng mit der Theorie hermitescher Operatoren im Raume Π_κ zusammenhängen. I. Einige Funktionenklassen und ihre Darstellungen. Math. Nachr. **77**, 187–236 (1977)
28. P. Kurasov, *Spectral Geometry of Graphs*, to appear
29. M. Nedic, C. Ehrenborg, Y. Ivanenko, A. Ludvig-Osipov, S. Nordebo, A. Luger, B.L.G. Jonsson, D. Sjöberg, M. Gustafsson, Herglotz functions and applications in electromagnetics, in *Advances in Mathematical Methods for Electromagnetics* (IET, 2019)
30. B. Nilsson, D. Sjöberg, S. Nordebo, M. Gustafsson, Optimal realizations of passive structures. IEEE Trans. Antennas Propag. **62**(9), 4686–4694 (2014)
31. M. Wohlers, E. Beltrami, Distribution theory as the basis of generalized passive-network analysis. IEEE Trans. Circuit Theory **12**(2), 164–170 (1965)
32. A.H. Zemanian, *Distribution Theory and Transform Analysis. An Introduction to Generalized Functions, with Applications* (McGraw-Hill Book Co., New York, 1965)

On Applications of Herglotz–Nevanlinna Functions in Material Sciences, II: Extended Applications and Generalized Theory

Miao-Jung Yvonne Ou and Annemarie Luger

1 Introduction

In this part of the review paper, we present a wide class of applications of Herglotz–Nevanlinna functions in material sciences. We start with the application in the static theory of two-phase composite materials, where the scalar-valued Herglotz–Nevanlinna functions correspond to the effective properties of the composite materials. Following this is an example showing how the matrix-valued Herglotz–Nevanlinna function theory can be applied to study the permeability tensor of a porous material. In both applications, the independent variable of the corresponding Herglotz–Nevanlinna functions is the contrast of material properties. The other group of applications presented in this chapter demonstrates the power of Herglotz–Nevanlinna functions in the study of systems of equations where the energy dissipation and dispersion satisfied causality and passivity, whose mathematical definition can be clearly specified in terms of the Herglotz–Nevanlinna functions. After presenting their various applications in material sciences, we conclude this chapter by introducing several classes of functions that can be considered as various generalizations of the Herglotz–Nevanlinna functions motivated by some emerging research field in physics and engineering.

This chapter is organized as follows. Section 2.1 deals with composite materials and bounds on effective properties. In Sect. 2.2, it is demonstrated how the usage of Herglotz–Nevanlinna functions can avoid a numerically costly memory term

M.-J. Y. Ou (✉)
Department of Mathematical Sciences, University of Delaware, Newark, DE, USA
e-mail: mou@udel.edu

A. Luger
Department of Mathematics, Stockholm University, Stockholm, Sweden
e-mail: luger@math.su.se

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_20

in the modeling of materials. Section 2.3 shows how the bounds for quasi-static cloaking can be derived. In Sect. 2.4, a general representation theorem of Herglotz–Nevanlinna functions is used in order to identify certain time dispersive and dissipative systems as restrictions of Hamiltonian systems.

Even if all these examples demonstrate the effectiveness of Herglotz–Nevanlinna functions, there are situations in applications that cannot be treated by these methods but would require more general classes of functions. This applies for instance for non-passive systems, appearing in electromagnetics, for which the analytic function in question might have non-positive imaginary part as well. Another example are composite materials with more than two phases. Then, even if the corresponding analytic functions still have positive imaginary part, they are not covered by the treatment above since they depend on more than one complex variable.

In Sect. 3, we therefore provide an overview of the mathematics that is available for different classes of functions that extend the classical Herglotz–Nevanlinna class, and that we expect to be relevant for applications in material science.

Note that items that are already defined in Part I will not be defined again in this part.

We hope that this review can be both helpful for people working in applications (by providing mathematical references for different aspects of Herglotz–Nevanlinna functions as well as their generalizations for future work) and interesting for pure mathematicians (by pointing out some relevant applications of Herglotz–Nevanlinna functions).

2 Applications

This section starts with the applications arising in the study of effective properties of composite materials, followed by the application in broadband passive quasi-static cloaking, and is concluded with a delicate application of the operator-valued Herglotz–Nevanlinna function theory for understanding the Hamiltonian structure of the time-dispersive and dissipative systems.

2.1 Effective Properties of Two-Phase Composite Materials

2.1.1 Effective Properties of Composite Materials and Bounds by Using Theory of the Stieltjes Function

Composite materials made of pure homogenous phases are abundant around us, e.g., reinforced concrete, plywood, fluid saturated sand, cancellous bones, and sea ice. Suppose the scale of the microstructure of a bulk composite sample is much smaller than the size of the sample; it makes sense to use the effective moduli to describe the properties of composite materials. For example, the effective permittivity of a complex fluid or the effective Young's modulus of a cancellous bone sample. Intuitively, these effective properties should depend on the properties of the pure phases as well as how these constituents are arranged, i.e., the microstructure of the composite. For multi-laminated microstructure, there are exact algebraic formulas for computing the effective properties as certain averages of the properties of the constituents; see [71]. However, for most microstructures, there is no exact "mixing theory formula" that can be used to compute the effective properties even though the effective properties are well defined by the homogenization theory [9, 49, 83, 87]. For the history of the development and the limitations of various formulas for computing the effective dielectric constants for simple microstructures, see [10] for details. Instead of looking for the exact formulas, many researchers have looked into the possibility of finding bounds of effective properties from the given constituent properties and information of the microstructure; see [45–47, 79, 84] [10–13, 20, 29, 42, 43, 66, 69, 70, 72], just to name a few. From this vast and rich literature emerges the beautiful bounding method based on the analytic properties of the effective moduli as a Stieltjes function of the dielectric constants of the pure phases; it was first described in [10] by David Bergman and further developed and extended from real-valued bounds to the general complex bounds by Graeme Milton in [68, 70]. This method provides a way for deriving the bounds without the use of variational principles. The first rigorous derivation of the Stieltjes function representation for the effective dielectric parameters of a two-phase composite is given in 1983 by Kenneth Golden and George Papanicolaou [43] in a random media setting. To fix ideas, we start with a brief description of the proof in [43].

Let $(\Omega, \mathcal{F}, P)$ be a probability space and the permittivity tensor $\boldsymbol{\epsilon}(\mathbf{x}, \omega)$ a stationary random field, where ω is a realization in Ω and $\mathbf{x}$ the spatial coordinates in $\mathbb{R}^d$ with $d \in \mathbb{N}$ and $d \geq 2$. Specifically, there exists a bijective group transformation $\boldsymbol{\tau}_{\mathbf{x}}$ from Ω to Ω, $\boldsymbol{\tau}_{\mathbf{x}}\boldsymbol{\tau}_{\mathbf{y}} = \boldsymbol{\tau}_{\mathbf{x}+\mathbf{y}}$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ such that $P(\boldsymbol{\tau}_{\mathbf{x}}A) = P(A)$ for all $\mathbf{x} \in \mathbb{R}^d$ and $A \in \mathcal{F}$. Suppose the permittivity tensor $\boldsymbol{\epsilon}(\mathbf{x}, \omega)$ can be represented by a measurable function $\tilde{\boldsymbol{\epsilon}}(\omega)$ on Ω as follows:

$$\boldsymbol{\epsilon}(\mathbf{x}, \omega) = \tilde{\boldsymbol{\epsilon}}(\boldsymbol{\tau}_{-\mathbf{x}}\omega). \tag{1}$$

It is further assumed to be bounded and satisfies the ellipticity condition, i.e., there exist two positive numbers α and β so that $\alpha\boldsymbol{\xi} \cdot \boldsymbol{\xi} \leq \boldsymbol{\epsilon}(\mathbf{x}, \omega)\boldsymbol{\xi} \cdot \boldsymbol{\xi} \leq \beta\boldsymbol{\xi} \cdot \boldsymbol{\xi}$ for all $\mathbf{x}, \boldsymbol{\xi} \in \mathbb{R}^d$. Since all the random fields considered here are stationary, the solutions of the form in (1) are sought for, i.e.,

$$\mathbf{E}(\mathbf{x}, \omega) = \tilde{\mathbf{E}}(\boldsymbol{\tau}_{-\mathbf{x}}\omega),\ \mathbf{D}(\mathbf{x}, \omega) = \tilde{\mathbf{D}}(\boldsymbol{\tau}_{-\mathbf{x}}\omega). \tag{2}$$

Consider the electrostatic Maxwell's equations for the random stationery electric field $\mathbf{E}(\mathbf{x}, \omega)$ and the electric induction field $\mathbf{D}(\mathbf{x}, \omega)$

$$\mathbf{D}(\mathbf{x}, \omega) = \boldsymbol{\epsilon}(\mathbf{x}, \omega)\mathbf{E}(\mathbf{x}, \omega),\ \nabla \times \mathbf{E}(\mathbf{x}, \omega) = \mathbf{0},\ \nabla \cdot \mathbf{D}(\mathbf{x}, \omega) = 0$$
$$\text{and} \int_\Omega \mathbf{E}(\mathbf{x}, \omega) P(d\omega) = \overline{\mathbf{E}} \tag{3}$$

with a prescribed constant electric field $\overline{\mathbf{E}}$ as its assemble average. Let the constant vector $\overline{E}$ be $\mathbf{e}_j$, the unit vector in the j-th direction, $j = 1, \cdots, d$, and denote the corresponding solution as $\mathbf{E}^j$ and $\mathbf{D}^j$. The effective permittivity tensor $\boldsymbol{\epsilon}^*$ is then defined as

$$\boldsymbol{\epsilon}^*\mathbf{e}_l := \int_\Omega \mathbf{D}^l(\mathbf{x}, \omega)P(d\omega) = \int_\Omega \boldsymbol{\epsilon}(\mathbf{x}, \omega)\mathbf{E}^l(\mathbf{x}, \omega)P(d\omega),\ l = 1, \cdots d. \tag{4}$$

It can be shown that these assemble averages of the solution do not depend on $\mathbf{x}$. The variational formulation plays an important role in the proof; it is described here. First consider the Hilbert space $H := L^2(\Omega, \mathcal{F}, P)$ endowed with the inner product $(\tilde{f}, \tilde{g})_H := \int_\Omega \tilde{f}(\omega)\tilde{g}(\omega)P(d\omega)$. Define the operator $T_\mathbf{x}$ acting on $\tilde{f} \in H$ as $T_\mathbf{x}\tilde{f}(\omega) := \tilde{f}(\boldsymbol{\tau}_{-\mathbf{x}}\omega)$. Because $\boldsymbol{\tau}_\mathbf{x}$ is measure preserving, $T_\mathbf{x}$ forms a unitary group and has closed densely defined infinitesimal generator $L_j := \frac{\partial}{\partial x_j} T_\mathbf{x}\Big|_{\mathbf{x}=0}$ for each $j = 1, \cdots, d$ with domain $\mathcal{D}_j \subset H$. Then, $\mathcal{D} := \bigcap_{j=1}^d \mathcal{D}_j \subset H$ is a Hilbert space with the inner product $(\tilde{f}, \tilde{g})_\mathcal{D} := \int_\Omega \tilde{f}(\omega)\tilde{g}(\omega)P(d\omega) + \sum_{i=1}^d \int_\Omega L_i\tilde{f}(\omega)L_i\tilde{g}(\omega)P(d\omega)$.

Since the problem (3) is equivalent to finding $\mathbf{E}$ in a curl-free space, and $\mathbf{E} = \overline{\mathbf{E}} + \mathbf{G}$ with a zero-average field $\mathbf{G}$, the following Hilbert space of vector-valued functions with inner product $(\cdot, \cdot)_\mathcal{H} := (\cdot, \cdot)_H$ is considered

$$\mathcal{H} := \left\{ \tilde{f}_j(\omega) \in H | L_i\tilde{f}_j = L_j\tilde{f}_i \text{ weakly },\ \int_\Omega \tilde{f}_j(\omega)P(d\omega) = 0,\ i, j = 1, \cdots, d \right\}.$$

The variational formulation of (3), after taking into account (2), is to find $\widetilde{\mathbf{G}}^l(\omega) \in \mathcal{H}$ such that

$$\int_\Omega \widetilde{\boldsymbol{\epsilon}}(\omega)(\widetilde{\mathbf{G}}^l(\omega) + \mathbf{e}_l)\widetilde{\mathbf{f}}(\omega)P(d\omega) = \mathbf{0} \text{ for all } \widetilde{\mathbf{f}} \in \mathcal{H}. \tag{5}$$

Recall that $\mathbf{e}_l$ is the unit vector in the l-th direction, $l = 1, \cdots, d$. This problem is well-posed because the bilinear form is coercive w.r.t. the H-norm and the Lax–Milgram lemma can be applied. Letting $\widetilde{\mathbf{f}} = \widetilde{\mathbf{G}}^k$ in (5) and using the definition in (4) and the fact that $\mathbf{e}_k = \widetilde{\mathbf{E}}^k - \widetilde{\mathbf{G}}^k$ by definition, one obtains the following symmetric form:

$$\boldsymbol{\epsilon}^*\mathbf{e}_l \cdot \mathbf{e}_k = \epsilon^*_{kl} = \int_\Omega \widetilde{\boldsymbol{\epsilon}}(\omega)\widetilde{\mathbf{E}}^l(\omega)\widetilde{\mathbf{E}}^k(\omega)P(d\omega), k, l = 1, \cdots, d. \tag{6}$$

To specialize to the two-phase case with isotropic constituents, consider $\boldsymbol{\epsilon}(\mathbf{x}, \omega) = \chi_1(\mathbf{x}, \omega)\epsilon_1\mathbf{I} + \chi_2(\mathbf{x}, \omega)\epsilon_2\mathbf{I}$ with $0 < \epsilon_1 \leq \epsilon_2 < \infty$ and indicator functions χ_p, p=1,2 such that $\widetilde{\chi}_1(\omega) + \widetilde{\chi}_2(\omega) = 1$. For example, $\widetilde{\chi}_1(\omega)$ equals one for all realizations $\omega \in \Omega$ for which the origin is occupied by the material with permittivity ϵ_1. Define the contrast $h := \frac{\epsilon_2}{\epsilon_1}$. The variational formulation now reads

$$\int_\Omega [\widetilde{\chi}_1(\omega) + h\widetilde{\chi}_2(\omega)](\widetilde{\mathbf{G}}^l(\omega) + \mathbf{e}_l) \cdot \widetilde{\mathbf{f}}(\omega)P(d\omega) = \mathbf{0} \text{ for all } \widetilde{\mathbf{f}} \in \mathcal{H}, \tag{7}$$

and the effective permittivity $\boldsymbol{\epsilon}$ defined in (4), as a function of h, can be expressed as follows:

$$\epsilon^*_{kl}(h) = \epsilon_1 \left[\delta_{kl} + (h-1)\int_\Omega (\widetilde{\chi}_2(\omega)\widetilde{E}^l_k(h,\omega)P(d\omega)\right], \tag{8}$$

or equivalently, one can focus on the function

$$m_{kl}(h) := (\epsilon_1)^{-1}\epsilon^*_{kl}(h) = \left[\delta_{kl} + (h-1)\int_\Omega (\widetilde{\chi}_2(\omega)\widetilde{E}^l_k(h,\omega)P(d\omega)\right]. \tag{9}$$

Note that $m_{kl}(1) = \delta_{kl}$ by definition. If one replaces the inner product $(\cdot,\cdot)_H$ with the one for complex-valued functions (by complex conjugating one of the functions), then the sesquilinear form in (7) is coercive in $h \in \mathbb{C} \setminus (-\infty, 0]$. Hence by Lax–Milgram lemma, there is a unique solution for all $\epsilon_1, \epsilon_2 \in \mathbb{C}$ such that $\frac{\epsilon_2}{\epsilon_1} \in \mathbb{C} \setminus (-\infty, 0]$. This further implies that $\boldsymbol{\epsilon}(h)$ is analytic in $h \in \mathbb{C} \setminus (-\infty, 0]$ and so is the effective permittivity $\boldsymbol{\epsilon}^*(h)$.

To obtain the spectral representation, first note that (7) can be written formally in terms of the Kronecker δ as (by thinking of $\widetilde{G}^l_k$ as the gradient of a function because it is curl-free, i.e., $\widetilde{G}^l_k = L_k\psi^l$ for some scalar function ψ^l, $k = 1, \ldots, d$)

$$\sum_{k=1}^d L_k[\widetilde{\chi}_1(\omega) + h\widetilde{\chi}_2(\omega)(\widetilde{G}^l_k + \delta_{kl})] = 0, l = 1, \cdots, d.$$

Rewriting the above expression in the following form:

$$\sum_{k=1}^d L_k\widetilde{G}^l_k + (h-1)\sum_{k=1}^d L_k\widetilde{\chi}_2(\omega)(\widetilde{G}^l_k + \delta_{kl}) = 0, l = 1, \cdots, d. \tag{10}$$

Define $\widetilde{\triangle} := \sum_{q=1}^d L_q^2$. Then we see that formally $L_j(\widetilde{\triangle})^{-1}\sum_{k=1}^d L_k\widetilde{G}^l_k = L_j(\widetilde{\triangle})^{-1}\widetilde{\triangle}\psi^l = L_j\psi^k = \widetilde{G}^l_j$. By applying $L_j(-\widetilde{\triangle}^{-1})$ to (10), followed by adding δ_{jl} on both sides, the desired expression is achieved

$$(\widetilde{G}^l_j + \delta_{jl}) + (1-h)\sum_{k=1}^d L_j(-\widetilde{\triangle}^{-1})L_k\widetilde{\chi}_2(\widetilde{G}^l_k + \delta_{kl}) = \delta_{jl},\ j, l = 1, \cdots, d. \tag{11}$$

Define a new variable $s := \frac{1}{1-h}$ and the operator $\widetilde{B}_{jk} := L_j(-\widetilde{\triangle}^{-1})L_k\widetilde{\chi}_2$. It can be shown that $\widetilde{\mathbf{B}}$ is a self-adjoint bounded linear operator with respect to the inner product

$$< \widetilde{f}, \widetilde{g} >:= \int_{\Omega} \widetilde{\chi_2}(\omega) \widetilde{f} \cdot \overline{\widetilde{g}},\ \widetilde{f}, \widetilde{g} \in (L^2(\Omega, \mathcal{F}, P))^d$$

$\|\widetilde{\mathbf{B}}\| \leq 1$. Then the integral equation above becomes

$$\widetilde{\mathbf{E}}^l(h) = \widetilde{\mathbf{G}^l}(h) + \mathbf{e}_l = \left(\mathbf{I} + \frac{\widetilde{\mathbf{B}}}{s} \right)^{-1} \mathbf{e}_l,\ s = \frac{1}{1-h}. \tag{12}$$

Applying the spectral theory of self-adjoint operators and taking into account the fact that $s \in [-1, 0]$ must be in the resolvent set, the solution is represented in terms of the projection-valued measure $\mathbf{Q}(dz)$ associated with $\widetilde{\mathbf{B}}$

$$\widetilde{G}^l_j + \delta_{jl} = s \int_0^1 \frac{(\mathbf{Q}(dz)\mathbf{e}_l)_j}{s-z}, l, j = 1, \ldots, d \text{ for all } s \in \mathbb{C} \setminus [0, 1]. \tag{13}$$

Therefore, the effective property in (8) is given by the following integral representation formula (IRF):

$$\epsilon^*_{kl}(h) = \epsilon_1 \left[\delta_{kl} - \int_0^1 \frac{\mu_{kl}(d\xi)}{s-\xi} \right], \quad \text{where } h = 1 - \frac{1}{s} \tag{14}$$

$$\text{and } \mu_{kl}(d\xi) = \int_{\Omega} \widetilde{\chi_2}(\omega)(\mathbf{Q}(d\xi)\mathbf{e}_l)_k P(d\omega).$$

In [43], the function m defined in (9) is used in the main theorem, which shows that the diagonal terms in the effective permittivity tensor can be represented in terms of a Stieltjes function with finite positive Borel measures. The main theorem is stated below.

Theorem 1 ([43]) *Let $s = \frac{1}{1-h}$ and $F_{kl}(s) = \delta_{kl} - m_{kl}(h)$. There exist (not necessarily positive) finite Borel measures $\mu_{kl}(d\xi)$ defined on $0 \leq \xi \leq 1$ such that the **diagonal elements** $\mu_{kk}(d\xi)$ are positive measures satisfying $F_{kl}(s) = \int_0^1 \frac{\mu_{kl}(d\xi)}{s-\xi}$ for all $s \in \mathbb{C} \setminus [0, 1]$.*

This theorem has been generalized for a special case of polycrystalline materials in [73].

Note that $s = \infty$ or equivalently $h = 1$ corresponds to the case of $\epsilon_1 = \epsilon_2$, i.e., homogeneous media. Once the IRF is obtained, the relation between the moments of the finite measure and the microstructure χ_2 can be established by comparing the coefficients of the Laurent series expansion of the IRF at $s = \infty$ and the Taylor series expansion of m_{ik} at $h = 1$, which involves differentiation w.r.t. h of the right-hand side of (9).

$$\mu^{(n-1)}_{kl} := \int_0^1 z^{n-1} \mu_{kl}(dz) = \frac{(-1)^{n-1}}{n!} m^{(n)}_{kl}(1), n = 1, 2, \cdots.$$

To evaluate $m_{kl}^{(n)}$, the derivatives of $\widetilde{\mathbf{E}}^l$ are needed. They can be calculated by expanding (12) near $s = \infty$ ($h = 1$) because the functions are analytic there. By comparing the Taylor coefficients at $h = 1$ on the left-hand side with the Laurent coefficents on the right-hand side, it is clear that

$$\frac{1}{n!}\frac{d^n\widetilde{\mathbf{E}}^l}{dh^n}\Big|_{h=1} = \widetilde{\mathbf{B}}^n\mathbf{e}_l, n = 0, 1, 2, \ldots, l = 1, \ldots d. \tag{15}$$

This directly implies that the n-th moments are related to the $(n+1)$-point correlation functions of the microstructure. Some explicit relations can be derived. For example, differentiating both sides of (9) leads to

$$\mu_{kl}^{(0)} = m'_{kl}(1) = \int_\Omega (\widetilde{\chi_2}(\omega)\widetilde{E}_k^l(1,\omega)P(d\omega) + \left[(h-1)\int_\Omega \widetilde{\chi_2}(\omega)\frac{d\widetilde{E}_k^l}{dh}P(d\omega)\right]\Big|_{h=1} = p_2\delta_{kl},$$

where p_2 is the volume fraction of the material with permittivity ϵ_2. If the microstructure $\chi_2(\mathbf{x},\omega)$ is spatially isotropic, then one can also obtain the exact expression of $\mu_{kl}^{(1)}$ by differentiating (9) twice, applying (15) and using the kernel function of $(-\triangle)^{-1}$ to obtain

$$\mu_{kl}^{(1)} = -\frac{1}{2}m''_{kl}(1) = -\int_\Omega \widetilde{\chi_2}(\omega)\frac{d\widetilde{E}_k^l}{dh}(1,\omega)P(d\omega) = \frac{p_1p_2}{d}\delta_{kl},$$

where $p_1 := 1 - p_2$.

Instead of in the setting of an unbounded, stationary random media, the Stieltjes IRF for two-phase composites with isotropic constituents can also be derived in a bounded and deterministic setting with various types of boundary conditions. For example, see [86] for the permittivity tensor and [21, 52, 77] for elasticity tensors.

A very nice feature of a Stieltjes IRF like (14) is the separation of influence—the contrast s is in the integrand, while all the microstructural information is encoded in the measure. It has been used to formulate the problem of finding bounds on the diagonal terms ϵ_{ii} as a linear optimization problem over the set of all measures supported in [0, 1] with constraints on the first n moments. Specifically, let $\mathcal{PM}$ be the set of all positive finite Borel measures on [0, 1], and consider $m(h) := m_{11}(h)$, its IRF, and the set of measures with constraints on the first n moments

$$1 - m(h;\mu) = F(s) = \int_0^1 \frac{\mu(d\xi)}{s-\xi},\ s = \frac{1}{1-h},\ s \in \mathbb{C}\setminus[0,1],$$

$$\mathcal{M}(a_0,\ldots,a_{n-1}) := \left\{\mu\middle|\mu \in \mathcal{PM},\ \mu^{(0)} = a_0,\ \mu^{(1)} = a_1,\ \ldots,\ \mu^{(n-1)} = a_{n-1}\right\},$$

where $a_j > 0$, $j = 0, \ldots, n-1$, form a positive definite sequence [3] so they can be the first n moments of a measure. To study the possible values of the effective properties by mixing two given materials with contrast h, consider the following set of possible values

$$\Lambda(h, a_0, \ldots, a_{n-1}) := \left\{1 - F(s, \mu) \in \mathbb{C} \middle| \mu \in \mathcal{M}(a_0, \ldots, a_{n-1}), s \in \mathbb{C} \setminus [0, 1]\right\}.$$

Note that not all measures correspond to a microstructure, and hence, Λ contains values that are not achievable by any microstructure. Nevertheless, it contains all possible values of $\epsilon^*_{11}(h)$. Clearly, with a fixed value of $s \in \mathbb{C} \setminus [0, 1]$, $1 - F(s, \mu)$ is a bounded linear map on $\mathcal{M}(a_0, \ldots, a_{n-1})$, which is a compact and convex subset of $\mathcal{M}$ in the topology of weak convergence. Therefore, $\Lambda(h, a_0, \ldots, a_{n-1})$ is a compact and convex set in $\mathbb{C}$, and the extreme points of $\mathcal{M}(a_0, \ldots, a_{n-1})$ are weak limits of measures of the form [53]

$$d\sigma(\xi) = \sum_{k=1}^{n} \alpha_k \delta(\xi - \xi_k), \alpha_k \geq 0,\ 1 > \xi_1 > \xi_2 > \cdots > \xi_n \geq 0$$

$$\text{and} \sum_{k=1}^{n} \alpha_k \xi_k^j = a_j,\ j = 0, 1, \ldots, n-1. \tag{16}$$

A crucial step in deriving the bounds is to note the structure of *interlacing poles and zeros* of the functions represented by the type of measures in (16). Consider

$$m(h; d\sigma) := 1 - \int_0^1 \sum_{k=1}^{n} \frac{\alpha_k \delta(\xi - \xi_k)}{s - \xi} = 1 - \sum_{k=1}^{n} \frac{\alpha_k}{s - \xi_k},$$

which is a rational function of s. Let $s = \rho_k$, $k = 1, \ldots, n$, $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_n$ be the zeros of $m(h)$; they must be of real-valued because of the IRF of m. Then the following expression is valid:

$$\prod_{k=1}^{n} \frac{s - \rho_k}{s - \xi_k} = 1 - \sum_{k=1}^{n} \frac{\alpha_k}{s - \xi_k} \Rightarrow \alpha_j = -\frac{\prod_{k=1}^{n}(\xi_j - \rho_k)}{\prod_{k \neq j}(\xi_j - \xi_k)}.$$

The fact that $\alpha_j \geq 0$ for all $j = 1, \ldots, n$ and the additional physical constraint $m(0, d\sigma) > 0$ then lead to the interlacing property

$$0 \leq \xi_n \leq \rho_n \leq \cdots \leq \xi_1 \leq \rho_1 \leq 1.$$

The bounds on ϵ^*_{11} can now be derived as follows. Suppose the volume fraction p_2 is given. Then the corresponding bounding function $m(h; d\sigma)$ (or extreme points of $\Lambda(h; p_2)$) has the following form because of its zero $s = \xi_1 + p_2 \leq 1$

$$m(h; d\sigma) = 1 - \frac{p_2}{s - \xi_1},\ 0 \leq \xi_1 \leq 1 - p_2. \tag{17}$$

If a real-valued h is considered and $h \geq 1$ ($s < 0$), then the bounds of $m(h; \mu)$ are the well-known geometric bound and the algebraic bound

$$1-\frac{p_2}{s-(1-p_1)} \le m(h;\mu) \le 1-\frac{p_2}{s} \Rightarrow \frac{1}{1-p_2+\frac{p_2}{h}} \le m(h;\mu) \le 1-p_2+hp_2.$$

If a complex-valued h is considered, then (17) provides the bounding curve for the convex hull of $\Lambda(h;\, p_2)$ consisting of a cord and an arc on the complex plane.

If the microstructure is assumed to be isotropic, then $\Lambda(h;\, p_2,\, \frac{p_1p_2}{d})$ is considered. When h is real and greater than one, the same procedure recovers the well-known Hashin–Shtrikman bounds

$$1+\frac{p_2}{\frac{1}{h-1}+\frac{p_1}{d}} \le m(h) \le h+\frac{p_1}{\frac{1}{1-h}+\frac{p_2}{dh}}.$$

When h is of complex values, the convex hull of $\lambda(h;\, p_2,\, \frac{p_1p_2}{d})$ is bounded by two arcs. More details about the bounding curves can be found in [27] and [71].

The bounding curves have been utilized for finding bounds on volume fractions of two-component composites from given complex-valued permittivity data [27, 67]. The separation of influence also makes this type of IRF very useful in retrieving microstructural information from given data on the effective parameters [19, 26, 28, 44, 74, 86]. The process of reconstructing the measure from data of $\epsilon_{ii}(h)$ is termed *dehomogenization*, whose theoretical foundation is established by E. Cherkaev in [26]. For applications of IRF in the study of transport in fluid, see [4, 5, 7, 73].

2.1.2 IRF for Permeability Tensors with Positive Matrix-Valued Measures

In Theorem 1, the Stieltjes function IRF is concluded only for diagonal terms of the effective permittivity tensor $\boldsymbol{\epsilon}$. A matrix-valued IRF seems to be a more suitable choice for the study of $\boldsymbol{\epsilon}$. Moreover, the IRF in Theorem 1 has such a simple form due to the fact that $|sF_{kk}(\sqrt{-1}s)| < M$ for all $s > 0$, i.e., F_{kk} decays fast enough along the imaginary axis; this simplifies the IRF for Herglotz–Nevanlinna function significantly. In this section, an application that corresponds to a matrix-valued function that is analytic in $\mathbb{C}^+ \cup (0,\infty)$ and does not satisfy the fast decay condition along the imaginary axis is presented. The Herglotz functions that are analytic on one halfline of the real axis have been studied by Kac and Krein in [51], where they are referred to as the Stieltjes function of class $\mathbf{S}$ and $\mathbf{S}^{-1}$. In different references of the literature, the definitions of these two classes sometimes appeared to be interchanged based on whether the singularities are on the left real axis or the right real axis, but the representation theorems for each class have also been modified accordingly. Listed below is the definition that best suits the application to be presented in this section and which is a matrix version of a modification of Definition 3 in Part I:

Definition 1 ([37, 54])

1. A matrix-valued function $\mathbf{F}$ holomorphic in $\mathbb{C}\setminus(-\infty,0]$ is of class $\mathbf{S}$ if the following two criteria are satisfied.

$\frac{\mathbf{F}(z)-\mathbf{F}^*(z)}{z-\bar{z}} \geq 0$ if $Im(z) \neq 0$ and $\mathbf{F}(x) \geq 0$ for $x > 0$.

2. A matrix-valued function $\mathbf{F}$ holomorphic in $\mathbb{C} \setminus (-\infty, 0]$ is of class $\mathbf{S}^{-1}$ if the following two criteria are satisfied.
$\frac{\mathbf{F}(z)-\mathbf{F}^*(z)}{z-\bar{z}} \leq 0$ if $Im(z) \neq 0$ and $\mathbf{F}(x) \geq 0$ for $x > 0$.

Theorem 2

1. $\mathbf{F}(z)$ *belongs to class* $\mathbf{S}$ *if and only if there exists a monotonically increasing matrix-valued function* $\boldsymbol{\sigma}(t)$ *such that the following IRF holds for* $z \in \mathbb{C} \setminus (-\infty, 0]$

$$\mathbf{F}(z) = \boldsymbol{A} + \boldsymbol{C}z + \int_{+0}^{\infty} \frac{z}{z+t} d\boldsymbol{\sigma}(t),$$

where $\boldsymbol{A} \geq 0$, $\boldsymbol{C} \geq 0$, $\int_{+0}^{\infty} \frac{1}{1+t} d\boldsymbol{\sigma}(t) < \infty$, *and* $\boldsymbol{A} + \boldsymbol{C} + \int_{+0}^{\infty} \frac{1}{1+t} d\boldsymbol{\sigma}(t) > 0$.

2. $\mathbf{F}(z)$ *belongs to class* $\mathbf{S}^{-1}$ *if and only if there exists a monotonically increasing matrix-valued function* $\boldsymbol{\sigma}(t)$ *such that the following IRF holds for* $z \in \mathbb{C} \setminus (-\infty, 0]$

$$\mathbf{F}(z) = \boldsymbol{A} + \frac{\boldsymbol{C}}{z} + \int_{+0}^{\infty} \frac{1}{z+t} d\boldsymbol{\sigma}(t),$$

where $\boldsymbol{A} \geq 0$, $\boldsymbol{C} \geq 0$, $\int_{+0}^{\infty} \frac{1}{1+t} d\boldsymbol{\sigma}(t) < \infty$, *and* $\boldsymbol{A} + \boldsymbol{C} + \int_{+0}^{\infty} \frac{1}{1+t} d\boldsymbol{\sigma}(t) > 0$.

The application considered here is about the transport property of porous materials in the framework of homogenization of periodic media. The Darcy permeability tensor $\mathbf{K}^{(D)}$ plays the role of quantifying the transport of fluid in porous media. To study how the microstructure of a porous media influences its Darcy permeability tensor, $\mathbf{K}^{(D)}$ is treated in [16] as the limiting case of the two-fluid problem where each isotropic fluid is characterized by its viscosity μ_j, $j = 1, 2$.The two-fluid problem is originally formulated in [58] for studying the Stokes equations of flows mixed with tiny stationery bubbles (inclusions). The mixture is assumed to occupy a region Ω, and the tiny inclusions are *periodically* distributed. The tininess of the inclusions leads to the assumption that the side of the periodic cell is $0 < \epsilon \ll 1$, while the diameter of Ω is $O(1)$. Let Q denote the unit periodic cell $(0, 1)^n$, $n = 2, 3$ that contains disjoint parts Q_1 and the inclusion Q_2 with interface $\Gamma = \partial Q_1 \cap Q_2 = \partial Q_2$ such that $Q = Q_1 \cup Q_2 \cup \Gamma$ and $\Gamma \cap \partial Q = \emptyset$. We assume inclusions Q_2 can distribute in any possible way in a scaled period cell ϵQ as long as they do not touch one another or the periodic cell boundary $\partial(\epsilon Q)$. For any given $0 < \epsilon \ll 1$, the domain Ω is covered by a periodic extension of ϵQ, which is denoted by $\widetilde{\epsilon Q}$. Note that $\widetilde{\epsilon Q} = \widetilde{\epsilon Q_1} \cup \widetilde{\epsilon Q_2} \cup \widetilde{\epsilon \Gamma}$.

For any fixed ϵ, the hosting fluid has constant viscosity μ_1 and occupies region $\Omega_1^\epsilon := \Omega \cap \widetilde{\epsilon Q_1}$, while the inclusion has constant viscosity μ_2 and occupies region $\Omega_2^\epsilon := \Omega \cap \widetilde{\epsilon Q_2}$. The interface between the hosting fluid and the inclusion fluid is denoted by Γ^ϵ, i.e., $\Gamma^\epsilon = \Omega_1^\epsilon \cap \Omega_2^\epsilon$ and $\Omega = \Omega_1^\epsilon \cup \Omega_2^\epsilon \cup \Gamma^\epsilon$. To make this problem amenable to the theory of Herglotz–Nevanlinna functions, we assume $\mu_1 > 0$ and $\mu_2 = z\mu_1$ with $z \in \mathbb{C}$. In the tensor notation, it is

$$\tilde{\mu}_{ijkl}(\mathbf{x}; z) = (\chi_2(\mathbf{x})z\mu_1 + \chi_1(\mathbf{x})\mu_1)\frac{(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})}{2} =: \mu(\mathbf{x}; z)I_{ijkl},\ z \in \mathbb{C}. \tag{18}$$

The two-fluid problem for the unknown fluid velocity $\mathbf{u}^\epsilon$ and fluid pressure p^ϵ is formulated as follows:

$$\begin{cases} \operatorname{div}\left(2\tilde{\boldsymbol{\mu}}^\epsilon(\mathbf{x}; z)e(\mathbf{u}^\epsilon)\right) - \nabla p^\epsilon = -\mathbf{f} \quad \text{in } \Omega_1^\epsilon \cup \Omega_2^\epsilon \\ \operatorname{div}\mathbf{u}^\epsilon = 0 \quad \text{in } \Omega_1^\epsilon \cup \Omega_2^\epsilon \\ \mathbf{u}^\epsilon = \mathbf{0} \quad \text{on } \partial\Omega \\ [\![\mathbf{u}^\epsilon]\!] = 0,\ \mathbf{u}^\epsilon \cdot \mathbf{n} = 0 \quad \text{on } \Gamma^\epsilon \\ [\![\boldsymbol{\pi}^\epsilon]\!] \cdot \mathbf{n} = \left([\![\boldsymbol{\pi}^\epsilon \cdot \mathbf{n}]\!] \cdot \mathbf{n}\right)\mathbf{n} \quad \text{on } \Gamma^\epsilon, \end{cases} \tag{19}$$

where $e(\mathbf{u}^\epsilon) = (\nabla\mathbf{u}^\epsilon + \nabla^T\mathbf{u}^\epsilon)/2$, $\mathbf{f}$ is a square integrable momentum source, $[\![\cdot]\!]$ denotes the jump across Γ^ϵ, $\mathbf{n}$ the exterior normal vector of Ω_2^ϵ, and the stress tensors $\pi(\mathbf{u}^\epsilon(\mathbf{x}; z))$ are defined as (will be denoted $\pi(\mathbf{u}^\epsilon)$ for brevity)

$$\pi(\mathbf{u}^\epsilon)_{ij} = 2\tilde{\mu}^\epsilon_{ijkl}e(\mathbf{u}^\epsilon(\mathbf{x}; z))_{kl} - p^\epsilon(\mathbf{x}; z)\delta_{ij}. \tag{20}$$

The first jump condition in system (19) describes the continuity of $\mathbf{u}^\epsilon$ across the interface, while the second jump condition states that only the normal traction can have a jump across the interface, i.e., the tangential (or shear) traction has to be continuous across the interface.

Similar to [58], it can be shown by using the Lax–Milgram lemma that for every fixed $0 < \epsilon \ll 1$, (19) has a unique solution $\mathbf{u}^\epsilon$, and p^ϵ is unique up to a constant for all $z \in \mathbb{C} \setminus (-\infty, 0]$. To obtain convergence results, p^ϵ is properly normalized to the new pressure $\widetilde{p^\epsilon}$ by a procedure described in [58]. As $\epsilon \to 0$, the solution $\mathbf{u}^\epsilon$ and $\widetilde{p^\epsilon}$ converge as follows:

$$\frac{\mathbf{u}^\epsilon}{\epsilon^2} \to \mathbf{u}^0 \quad \text{weakly in } L^2(\Omega)^3,\ \ \widetilde{p^\epsilon} \to P \quad \text{strongly in } L^2(\Omega)/\mathbb{C},$$

where $\mathbf{u}^0$ and P satisfy the homogenized system:

$$\begin{cases} \mathbf{u}^0 = -\mathbf{K}(\nabla P - \mathbf{f}) \quad \text{in } \Omega \\ \operatorname{div}\mathbf{u}^0 = 0 \quad \text{in } \Omega, \end{cases} \tag{21}$$

where the **self-permeability K** is defined as

$$K_{ij}(z) := \int_Q u^i_j(\mathbf{x};z)d\mathbf{y} = \int_Q \mathbf{u}^i(\mathbf{x};z)\cdot \mathbf{e}_j d\mathbf{y} \tag{22}$$

by the Q-periodic, divergence-free solution $\mathbf{u}^k$ to the following cell problem:

$$\begin{cases} \operatorname{div}_{\mathbf{y}}\left(2\tilde{\mu}(\mathbf{y};z)e(\mathbf{u}^k) - p^k\mathbf{I}\right) + \mathbf{e}_k = \mathbf{0} \quad \text{in } Q_1 \cup Q_2 \\ \qquad\qquad\qquad [\![\boldsymbol{\pi}]\!]\cdot\mathbf{n} = \left([\![\boldsymbol{\pi}\cdot\mathbf{n}]\!]\cdot\mathbf{n}\right)\mathbf{n} \text{ on } \Gamma, \end{cases} \tag{23}$$

where $\mathbf{y}$ denotes the coordinates for the unit periodic cell Q, and the viscosity tensor in $Q_1 \cup Q_2$ is

$$\tilde{\mu}_{ijkl}(\mathbf{y};z) = (\chi_2(\mathbf{y})z\mu_1 + \chi_1(\mathbf{y})\mu_1)I_{ijkl} = \mu(\mathbf{y};z)I_{ijkl}. \tag{24}$$

The function space for the cell problem is the Hilbert space

$$H(Q) := \Big\{\mathbf{v} : \mathbf{v} \in H^1(Q_1 \cup Q_2)^3 \Big| \operatorname{div}_{\mathbf{y}}\mathbf{v} = 0,\ \mathbf{v}\cdot\mathbf{n} = 0 \text{ in } H^{-\frac{1}{2}}(\Gamma),$$
$$[\![\mathbf{v}]\!]_\Gamma = \mathbf{0},\ \mathbf{v} \text{ is } Q\text{- periodic}\Big\} \tag{25}$$

endowed with inner product

$$(\mathbf{u},\mathbf{v})_Q = \int_Q 2\mu_1 e(\mathbf{u}) : \overline{e(\mathbf{v})}d\mathbf{y}, \tag{26}$$

where the induced norm is denoted by $\|\mathbf{u}\|_Q^2 := (\mathbf{u},\mathbf{u})_Q$, and the contraction product of two $n \times n$ matrices $\mathbf{A} = \{a_{ij}\}$, $\mathbf{B} = \{b_{ij}\}$ is $\mathbf{A} : \mathbf{B} = \sum_{i,j=1}^n a_{ij}b_{ij}$. Let $\mathcal{R}(Q)$ denote the space of rigid body displacement in Q, i.e., $\mathbf{u} = \mathbf{A}\mathbf{y} + \mathbf{b}$ with constant skew-symmetric matrix $\mathbf{A}$ and constant vector $\mathbf{b}$. Then we have $H(Q)\cap\mathcal{R}(Q) = \{0\}$ because $\mathbf{A} = 0$ due to the Q-periodicity, and $\mathbf{u}\cdot\mathbf{n} = 0$ implies $\mathbf{b} = \mathbf{0}$. Hence, Korn's inequality can be applied to show that the norm $\|\cdot\|_Q$ is equivalent to the $H^1(Q_1 \cup Q_2)$-norm. In this setting, it is proved in [16] by using the Lax–Milgram lemma that the cell problem has a unique solution $\mathbf{u}(\mathbf{y},z) \in H(Q)$ and $p(\mathbf{y},z) \in L^2(Q)/\mathbb{C}$ for all $z \in \mathbb{C}\setminus(-\infty,0]$. Also, $\mathbf{u}(z)$ is analytic in $\mathbb{C}\setminus(-\infty,0]$ and so is $\mathbf{K}(z)$. Moreover, $\mathbf{K}$ in (22) can be expressed as the following quadratic form (note the $(\overline{z})$ used in function $\tilde{\mu}$):

$$K_{ij}(z) = \int_Q 2\tilde{\mu}(\mathbf{y};\overline{z})\overline{e(\mathbf{u}^i(z))} : e(\mathbf{u}^j(z))d\mathbf{y}, \tag{27}$$

and its conjugate transpose $\mathbf{K}^* := \overline{\mathbf{K}^T}$ is

$$(K^*)_{ij}(z) = \int_Q 2\tilde{\mu}(\mathbf{y}; z) e(\mathbf{u}^j(z)) : \overline{e(\mathbf{u}^i(z))} d\mathbf{y}. \tag{28}$$

We observe the two properties:

1. Because $K_{ij}(z) - K^*_{ij}(z) = 2\mu_1(\overline{z} - z) \int_{Q_2} \overline{e(\mathbf{u}^i(z))} : e(\mathbf{u}^j(z)) d\mathbf{y}$, we have

$$\frac{K_{ij}(z) - K^*_{ij}(z)}{z - \overline{z}} = -2\mu_1 \int_{Q_2} \overline{e(\mathbf{u}^i(z))} : e(\mathbf{u}^j(z)) d\mathbf{y} = -(\mathbf{u}^j, \mathbf{u}^i)_{Q_2} =: -A_{ij}.$$

The matrix $\boldsymbol{A}$ is obviously Hermitian. Furthermore, for any $\boldsymbol{\xi} \in \mathbb{C}^3$, we have $\sum_{i,j=1}^n \overline{\xi_i} A_{ij} \xi_j = (\sum_{j=1}^n \xi_j \mathbf{u}^j, \sum_{i=1}^n \xi_i \mathbf{u}^i)_{Q_2} \geq 0$. Therefore,

$$\frac{\mathbf{K}(z) - \mathbf{K}^*(z)}{z - \overline{z}} \leq 0 \text{ if } Im(z) \neq 0.$$

2. For $x > 0$, recall that $K_{ij}(x) = \big((\mathbf{u}^j, \mathbf{u}^i)_{Q_1} + x(\mathbf{u}^j, \mathbf{u}^i)_{Q_2}\big)$. With a similar argument as before, we have

$$\mathbf{K}(x) \geq 0 \text{ for } x > 0.$$

With these two properties and the fact that $\mathbf{K}$ is holomorphic in $\mathbb{C} \setminus (-\infty, 0]$, we see that $\mathbf{K}(z)$ is a Stieltjes function of class $\mathbf{S}^{-1}$. Therefore, by Theorem 2, there exists a monotonically increasing matrix-valued function $\boldsymbol{\sigma}(t)$ such that the following integral representation formula holds for $z \in \mathbb{C} \setminus (-\infty, 0]$

$$\mathbf{K}(z) = \boldsymbol{A} + \frac{\boldsymbol{C}}{z} + \int_{+0}^{\infty} \frac{1}{z+t} d\boldsymbol{\sigma}(t),$$

where $\boldsymbol{A} \geq 0$, $\boldsymbol{C} \geq 0$, $\int_{+0}^{\infty} \frac{1}{1+t} d\boldsymbol{\sigma}(t) < \infty$, and $\boldsymbol{A} + \boldsymbol{C} + \int_{+0}^{\infty} \frac{1}{1+t} d\boldsymbol{\sigma}(t) > 0$. It is proved in [16] that there exist two positive numbers $E_1, E_2 > 1$ such that $\mathbf{K}(z)$ is analytic in $(-\infty, -2E_1^2) \cup (-\frac{1}{2E_2^2}, 0)$; E_1 and E_2 are the extension constants related to Q_1 and Q_2. It is also shown in [16] that $\mathbf{K}(\infty) = \mathbf{K}^{(D)}$, the Darcy permeability of porous media defined in the appendix of [82] by L. Tartar. Also, $\mathbf{K}(0) = \mathbf{K}^{(B)}$, the permeability when the inclusion is bubbles. Therefore, the IRF above can be simplified to

$$\mathbf{K}(z) = \mathbf{K}^{(D)} + \int_{\frac{1}{2E_2^2}}^{2E_1^2} \frac{1}{z+t} d\boldsymbol{\sigma}(t).$$

This shows an interesting fact that with $\mu_1 \in \mathbb{R}$ fixed, the larger the inclusion viscosity $\mu_1 z$ is, the smaller the permeability. To see how the microstructure influences $\mathbf{K}$, a new variable $s := \frac{1}{z-1}$ is defined. As a function of s, $\mathbf{K}$ can be shown to be a function of class $\mathbf{S}$ and hence can be expressed with a monotonically increasing matrix-valued function $\boldsymbol{\rho}(t)$ as follows:

$$\mathbf{K}(s) = \mathbf{K}^{(D)} + \int_{\frac{1}{1+2E_1^2}}^{\frac{2E_2^2}{1+2E_2^2}} \frac{s}{s+t} d\boldsymbol{\rho}(t), \tag{29}$$

which is valid for all $s \in \mathbb{C} \setminus [-\frac{2E_2^2}{1+2E_2^2}, -\frac{1}{1+2E_1^2}]$. Finally, the link between the moments of measure $\boldsymbol{\rho}$ and the microstructure can be established by expansion at $s = \infty$. See [16] for details.

2.2 Numerical Treatment of Memory Terms in the Modeling of Materials

In this section, the Stieltjes function structure of the memory kernel is utilized to design an efficient numerical scheme for solving the poroelastic wave equations.

In the modeling of wave propagation in poroelastic media such as bones and fluid saturated rock or viscoelastic materials such as polymeric fluid, the current state depends on the history of the time evolution of the state from the starting time. As a result, the governing equations contain a time convolution term whose integrand consists of the unknown state function and a pre-described time-dependent kernel function K; this convolution integral is referred to as the memory term. For a time-domain solver. The presence of memory terms poses the challenges of a proper time-stepping scheme. In the literature, it has been handled by storing the history of the solution such as in [65] or a proper design of quadrature rules for approximating the memory term in the time domain, e.g., [59] and the reference therein. For poroelastic wave equations, the memory term appears in the equation of the generalized Darcy's law, which relates the pore pressure p, the solid velocity $\mathbf{v}$, and $\mathbf{q}$, the fluid velocity relative to the solid, as follows

$$-\nabla p = \rho_f \frac{\partial \mathbf{v}}{\partial t} + \left(\frac{\rho_f}{\phi}\right) \check{\boldsymbol{\alpha}} \star \frac{\partial \mathbf{q}}{\partial t}, \quad t > 0, \tag{30}$$

where the matrix $\check{\boldsymbol{\alpha}}$ is the inverse Fourier–Laplace transform of the $\boldsymbol{\alpha}$ defined in (31). The physical origin of the memory term is due to the fact that at the micro-scale (scale of the pore size) in the frequency domain, the boundary layer of the viscous pore fluid is frequency-dependent, e.g., the viscous skin depth is inversely proportional to the square root of the frequency. In the seminal papers [17, 18], M. A. Biot calculated a critical frequency f_c that separates the regime of laminar pore fluid flow from that of turbulent pore fluid flow, and each regime corresponds to a different expression of $\boldsymbol{\alpha}(\omega)$.

However, the discrepancy between the model prediction and experiment observation of wave dissipation has prompted the study of high-frequency corrections that are more general than the one proposed in [18]. In order to describe these corrections, we need to introduce the physical quantity that encapsulates this complicated viscodynamics, i.e., the dynamic tortuosity tensor $\boldsymbol{\alpha}(\omega)$ and the dynamic permeability tensor $\mathbf{K}(\omega)$ with ω being the frequency. For $\omega \neq 0$, $\boldsymbol{\alpha}(\omega)$ and $\mathbf{K}(\omega)$ are related as follows:

$$\boldsymbol{\alpha}(\omega) = \frac{i\eta\phi}{\omega\rho_f}\mathbf{K}(\omega)^{-1} \tag{31}$$

with $i := \sqrt{-1}$. and ϕ, η, ρ_f being the volume fraction, the dynamic viscosity, and the density of the pore fluid, respectively.

To keep the discussion simple, we consider the isotropic case $\boldsymbol{\alpha}(\omega) = \alpha(\omega)\mathbf{I}$.

One of the most widely used corrections is derived in [50] by Johnson, Koplik and Dashen (JKD)

$$\alpha(\omega) = \alpha_\infty + \frac{i\eta\phi}{\omega K_0\rho_f}\left(1 - \frac{4i\alpha_\infty^2 K_0^2\rho_f\omega}{\eta\Lambda^2\phi^2}\right)^{\frac{1}{2}} =: \alpha^D(\omega), \tag{32}$$

where K_0 is the static permeability, α_∞ is the limit of α at infinite frequency, and Λ a structure parameter related to the surface-to-volume ratio of the pore space; all of these parameters can be measured; see [50]. An important ingredient in their derivation is the causality of $K(\omega)$. This is carried out by first considering the gradient force and the fluid velocity field that are time-harmonic with frequency ω, i.e., the one-sided Fourier transform, followed by extending $K(\omega)$ for complex-valued ω. A function defined on the complex ω-plane is **causal** if and only if it is analytic in the upper halfplane. Another requirement in the JKD derivation is that a real-valued stimulus $\nabla pe^{-i\omega t} + \nabla\overline{p}e^{i\overline{\omega}t}$ should result in a real-valued response. This leads to the symmetry constraint $K(-\overline{\omega}) = \overline{K(\omega)}$. According to [50], the function in (32) was chosen because it is the simplest form of functions that are causal and satisfies the aforementioned symmetry constraint. Of course, there is no reason why it has to be in this form. Indeed, in [25], [88], and [80], it is shown that when the cross-section of the pore space varies rapidly enough, the JKD formula in (32) severely underestimates the imaginary part of the measured dynamic tortuosity for low frequency $\omega \leq \omega_0 := \frac{\eta\phi}{K_0\rho_f\alpha_\infty}$.

In [6], the spectrum $\{\epsilon_j\}_{j=1}^\infty$ of the incompressible Stokes equation with kinetic viscosity ν, in the pore space, is used to derive the following general integral representation formula (IRF) for the dynamic permeability

$$K(\omega) = \frac{\nu}{F}\int_0^{\Theta_1}\frac{\Theta dG(\Theta)}{1 - i\omega\Theta}, \text{ with } G(\Theta) = \frac{\sum_{\Theta_n\leq\Theta} b_n^2}{\sum_{n=1}^\infty b_n^2}, F := \left(\phi\sum_{n=1}^\infty b_n^2\right)^{-1} \tag{33}$$

where $\Theta_1 := (\nu\epsilon_1)^{-1} < \infty$, dG a positive measure with mass 1, and $b_n > 0$, $n = 1, 2, \ldots$, ordered in the same order as the non-decreasing eigenvalues, are defined by the orthogonal spectral system of the Stokes equations. This implies the dependence of K on the pore space geometry is encoded in the measure dG. Note that $0 < \epsilon_1 \leq \epsilon_2 \leq \cdots$, $\epsilon_n \to \infty$ as $n \to \infty$.

The integral representation for K in (33) shows that K itself is not a Herglotz–Nevanlinna function itself. However, as was noted in [76], the permeability in (33) can be related to a Stieltjes function with the new variables $s := -i\omega$, $\xi := -\frac{1}{s}$ and

$$R(\xi) := -s\left(\frac{F}{\nu}\right) K(is) = \int_0^{\Theta_1} \frac{\Theta dG(\Theta)}{\xi - \Theta} =: \int_0^{\Theta_1} \frac{d\lambda(\Theta)}{\xi - \Theta}. \tag{34}$$

As a result, the tortuosity α can be represented as follows:

$$\alpha(\omega) = \frac{\eta\phi}{\rho_f K_0}\left(\frac{i}{\omega}\right) + \int_0^{\Theta_1} \frac{d\sigma(\Theta)}{1 - i\omega\Theta} \text{ for } \omega \text{ such that } -\frac{i}{\omega} \in \mathbb{C} \setminus [0, \Theta_1], \tag{35}$$

where $d\sigma$ is a positive Borel measure that has a Dirac mass at $\Theta = 0$ with strength α_∞. It is also shown in [76] that the JKD tortuosity in (32) is indeed a special case of (35) by finding the corresponding $d\sigma$ for (32). In the context of JKD permeability $K^D(\omega)$, this IRF result implies that the geometry parameter Λ is related to the microstructure as follows:

$$\Lambda = \sqrt{\frac{2K_0\alpha_\infty}{\phi[\frac{\mu_1(d\lambda^D)}{\mu_0^2(d\lambda^D)} - 1]}}, \tag{36}$$

where μ_0 and μ_1 are the zero-th moment and the first moment, respectively, of the corresponding measure in (34) for the JKD permeability. We note that the commonly used formula in the engineering literature is $\Lambda \approx \sqrt{\frac{2\alpha_\infty K_0}{\phi/4}}$.

According to the theorem proved in [41], the multi-point rational approximation P_{n-1}/Q_n of every Stieltjes function $f(z) = \int_a^b \frac{d\lambda(t)}{z-t}$ is itself a Stieltjes function $\int_a^b \frac{d\beta(t)}{z-t}$ with a bounded, non-decreasing $\beta(t)$. Hence, the poles of the rational approximation of a Stieltjes function are all simple with positive residue and located in $[a, b]$. Moreover, it is shown there that the convergence is geometrical with order $2n$ in any compact set on the complex plane. We state the theorem that is relevant to the approximation of α here.

Theorem 3 ([41]) *Let f be a Stieltjes function of the form $\int_a^b \frac{d\lambda(t)}{z-t}$, and let γ_k be a set of interpolation points, consisting of k_1 real points $x_1, \cdots, x_{k_1} \in \mathbb{R} \setminus [a, b]$, and k_2 non-real points $z_1, \cdots, z_{k_2} \in \mathbb{C} \setminus \mathbb{R}$. Let $P_{n-1}(z)$ and $Q_n(z)$ be polynomials of degree at most $n - 1$ and n, respectively, with $k_1 + k_2 + k_3 = 2n$ such that the following relations are satisfied:*

$$\begin{cases} f(z)Q_n(z) - P_{n-1}(z) = A(z)\prod_{j=1}^{k_1}(z-x_j)\prod_{j=1}^{k_2}(z-z_j)(z-\overline{z}_j) \\ f(z)Q_n(z) - P_{n-1}(z) = B(z)z^{n-k_3-1}, \end{cases}$$

where $A(z)$, $B(z)$ are analytic in $\mathbb{C}\setminus[a,b]$ and $B(z)$ bounded at ∞. Then for the multi-point rational approximation, it holds:

1. *$[n-1/n]_f(z) := \frac{P_{n-1}(z)}{Q_n(z)} = \int_a^b \frac{d\beta(t)}{z-t}$ for some bounded, non-decreasing function $\beta(t)$.*
2. *Denote by γ_k, $k = 1,\dots 2n$ the interpolation points. Fix one interpolation point v, and denote $G_k(z) := \frac{\psi_v(z)-\psi_v(\gamma_k)}{1-\psi_v(z)\overline{\psi_v(\gamma_k)}}$, where $\psi_v(z) = \frac{\sqrt{z-b}-\sqrt{\frac{v-b}{v-a}}\sqrt{z-a}}{\sqrt{z-b}+\sqrt{\frac{v-b}{v-a}}\sqrt{z-a}}$ a conformal mapping that maps $\mathbb{C}\setminus[a,b]$ onto the interior of the unit circle and v onto 0.*

   If a, b are finite numbers, then there exists a constant K_v, dependent on f and v but not on n, such that for $z \in \mathbb{C}\setminus[a,b]$ it holds

$$\left| f(z) - \frac{P_{n-1}(z)}{Q_n(z)} \right| \le K_v \frac{1}{\max(|z-a|,|z-b|)} \frac{1}{1-|\psi_v(z)|} \cdot \Pi_{k=1}^{2n} |G_k(z)|.$$

Moreover, if for all n the interpolation points γ_k are at least at a fixed non-zero distance away from $[a,b]$, then there exists $\triangle_F < 1$ such that $|G_k(z)| \le \triangle_F < 1$, and hence,

$$\left| f(z) - \frac{P_{n-1}(z)}{Q_n(z)} \right| \le K_v \frac{1}{\max(|z-a|,|z-b|)} \frac{1}{1-|\psi_v(z)|} (\triangle_F)^{2n}. \tag{37}$$

Therefore, the approximation converges geometrically in any compact set that does not intersect with $[a,b]$.

Now, we rewrite (35) in terms of ξ and rearrange terms to obtain

$$\alpha(\xi) + \frac{\eta\phi}{\rho_f K_0}\xi - \alpha_\infty = \xi \int_{+0}^{\Theta_1} \frac{d\sigma(\Theta)}{\xi - \Theta}.$$

Theorem 3 implies that

$$\xi \int_{+0}^{\Theta_1} \frac{d\sigma(\Theta)}{\xi - \Theta} \approx \xi \sum_{k=1}^{n} \frac{\rho_k}{\xi - \pi_k}, \text{ with } \rho_k > 0 \text{ and } 0 < \pi_k < \Theta_1$$

with error bound

$$\left| \xi \int_{+0}^{\Theta_1} \frac{d\sigma(\Theta)}{\xi - \Theta} - \xi \sum_{k=1}^{n} \frac{\rho_k}{\xi - \pi_k} \right| \le \frac{|\xi|}{\max(|\xi|, |\xi - \Theta_1|)} \frac{K_v}{|1 - \psi_v(\xi)|} (\triangle_F)^{2n}.$$

Changing the variable back to s, it is clear that there exists $r_k > 0$ and $p_k > 0$ such that

$$\alpha(s)\approx\frac{\eta\phi}{\rho_f K_0}\left(\frac{1}{s}\right)+\alpha_\infty+\sum_{k=1}^{n}\frac{r_k}{s-p_k}\ \text{for}\ s\in\mathbb{C}\setminus(-\infty,-\frac{1}{\Theta_1}].$$

The r_k and p_k can be accurately computed from given nodes $(s_j,\alpha(s_j))_{j=1}^{n}$ by using two-sided residue approximation with arbitrary precision arithmetic; see [78, 85] for details. Let $\mathcal{L}$ be the Fourier–Laplace transform (note that this differs from the Laplace transform in, e.g., Sect. 2.6 in Part I by a factor-i)

$$\mathcal{L}f(\omega):=\int_{\mathbb{R}^+}f(t)e^{i\omega t}dt=:\hat{f}(\omega). \tag{38}$$

We approximate the transform of the memory term as follows:

$$\begin{aligned}\mathcal{L}[\check{\alpha}\star\frac{\partial\mathbf{q}}{\partial t}](s)&=\alpha(s)(s\hat{\mathbf{q}}-\mathbf{q}(0))\approx\left(\alpha_\infty+\sum_{k=1}^{n}\frac{r_k}{s-p_k}+\frac{a}{s}\right)(s\hat{\mathbf{q}}-\mathbf{q}(0))\\&=\alpha_\infty(s\hat{\mathbf{q}}-\mathbf{q}(0))+\left(a+\sum_{k=1}^{n}r_k\right)\hat{\mathbf{q}}+\left(\sum_{k=1}^{n}\frac{r_kp_k}{s-p_k}\right)\hat{\mathbf{q}}\\&\quad-\left(\sum_{k=1}^{n}\frac{r_k}{s-p_k}+\frac{a}{s}\right)\mathbf{q}(0),\ \text{where}\ a:=\frac{\eta\phi}{\rho_f K_0}.\end{aligned}$$

Furthermore, for each of the terms in the sum, since all the singularities p_k are restricted to the left of $s=-\frac{1}{\Theta_1}$, the inverse Laplace transform can be performed exactly by integrating along the imaginary axis (Theorem 9.1.1 in [33])

$$\mathcal{L}^{-1}\left[\frac{1}{s-p_k}\right](t)=\frac{1}{2\pi i}\lim_{R\to\infty}\int_{-iR}^{iR}\frac{1}{\zeta-p_k}e^{\zeta t}d\zeta=r_ke^{p_kt},\ t>0.$$

This integral is calculated by integrating along $[-Ri,Ri]\cup\{s=Re^{i\theta}|\pi/2<\theta<3\pi/2\}$ and applying the residue theorem and letting $R\to\infty$. As a result, we have for $t>0$

$$\begin{aligned}\left(\check{\alpha}\star\frac{\partial\mathbf{q}}{\partial t}\right)(\mathbf{x},t)&:=\int_0^t\check{\alpha}(\tau)\frac{\partial\mathbf{q}}{\partial t}(\mathbf{x},t-\tau)d\tau\\&\approx\alpha_\infty\frac{\partial\mathbf{q}}{\partial t}+\left(a+\sum_{k=1}^{n}r_k\right)\mathbf{q}-\sum_{k=1}^{n}r_k(-p_k)e^{p_kt}\star\mathbf{q}\\&\quad-\left(\sum_{k=1}^{n}r_ke^{p_kt}+aH(t)\right)\mathbf{q}(0),\end{aligned}$$

where H denotes the Heaviside function. Applying a strategy similar to those in the literature [22], we define the auxiliary variables Θ_k, $k = 1, \dots, n$ such that

$$\theta_k(\mathbf{x}, t) := (-p_k)e^{p_k t} \star \mathbf{q}. \tag{39}$$

It can be easily checked that θ_k, $k = 1, \dots, M$, satisfies the following equation:

$$\partial_t \theta_k(\mathbf{x}, t) = p_k \theta_k(\mathbf{x}, t) - p_k \mathbf{q}(\mathbf{x}, t). \tag{40}$$

Finally, we can approximate the generalized Darcy's law (30) with the following system that has no explicit memory terms

$$\begin{cases} \partial_t \theta_k(\mathbf{x}, t) = p_k \theta_k(\mathbf{x}, t) - p_k \mathbf{q}(\mathbf{x}, t), \; k = 1, \cdots, n & (41) \\ \quad -\nabla p = \rho_f \dfrac{\partial \mathbf{v}}{\partial t} + \left(\dfrac{\rho_f \alpha_{\infty j}}{\phi}\right) \dfrac{\partial \mathbf{q}}{\partial t} + \left(\dfrac{\eta}{K_{0j}} + \dfrac{\rho_f}{\phi} \displaystyle\sum_{k=1}^{n} r_k\right) \mathbf{q} & \\ \qquad - \left(\dfrac{\rho_f}{\phi}\right) \displaystyle\sum_{k=1}^{n} r_k \theta_k - \dfrac{\rho_f}{\phi} \left(\displaystyle\sum_{k=1}^{n} r_k e^{p_k t} + a\right) \mathbf{q}(\mathbf{x}, 0), \; t > 0. & (42) \end{cases}$$

The generalization to anisotropic tortuosity function α is straightforward and has been implemented numerically in [85].

2.3 Broadband Passive Quasi-Static Cloaking

The sum rules for Herglotz–Nevanlinna functions can be applied to explain and quantify the limitations of broadband quasi-static cloaking. In this section, we summarize the results from the paper by Cassier and Milton [23].

Here the geometry is as follows: $\Omega \subset \mathbb{R}^3$ is an open bounded set, which in this context is thought of as the whole device. Let then $O \subset \Omega$ be a bounded simply connected dielectric inclusion with Lipschitz boundary such that the cloak $\Omega \setminus \overline{O}$ is open and connected.

Consider the Maxwell equations for $\mathbf{D}$ (electric induction), $\mathbf{B}$ (magnetic induction), $\mathbf{E}$ (electric field), and $\mathbf{H}$ (magnetic field)

$$\partial_t \mathbf{D} - \nabla \times \mathbf{H} = -\mathbf{J}, \;\; \partial_t \mathbf{B} + \nabla \times \mathbf{E} = -\mathbf{J}_B, \;\; \nabla \cdot \mathbf{D} = 0, \;\; \nabla \cdot \mathbf{B} = 0. \tag{43}$$

Suppose the external electric current $\mathbf{J}$ and magnetic current $\mathbf{J}_B$ are absent; one has $\mathbf{J} = \mathbf{J}_B = \mathbf{0}$. Let ϵ_0 and μ_0 denote the permittivity constant and the permeability constant of vacuum, respectively. The Maxwell equations are supplemented with the constitutive laws

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \qquad \text{and} \qquad \mathbf{B} = \mu_0 \mathbf{H} + \mathbf{M}, \tag{44}$$

where the electric polarization $\mathbf{P}$ and the magnetic polarization $\mathbf{M}$ are defined by the time convolution with the real-valued electric susceptibility function χ_E and the magnetic susceptibility function χ_M as follows:

$$\mathbf{P} = \epsilon_0 \chi_E \star \mathbf{E} \qquad \text{and} \qquad \mathbf{M} = \epsilon_0 \chi_M \star \mathbf{H}. \tag{45}$$

The functions considered are as follows: for each $\mathbf{x} \in \Omega$, we have $\chi_E(\mathbf{x}, \cdot)$, $\chi_M(\mathbf{x}, \cdot) \in L^1(\mathbb{R})$ and $\mathbf{E}$, $\mathbf{E}$ are in $H^1(\mathbb{R}; L^2(\Omega))$.

The causality assumption of the material, i.e., $\mathbf{E}(\cdot, t)$ and $\mathbf{H}(\cdot, t)$ cannot influence $\mathbf{D}(\cdot, t')$ and $\mathbf{B}(\cdot, t')$ for $t' < t$, implies that $\chi_E(\cdot, t)$ and $\chi_M(\cdot, t)$ are supported in $t \geq 0$. Applying the Laplace–Fourier transform (38) to (44) and (45), the following relations in the frequency domain are obtained:

$$\hat{\mathbf{D}}(\omega) = \epsilon_0(1 + \hat{\chi}_E(\omega))\hat{\mathbf{E}}(\omega) =: \epsilon(\omega)\hat{\mathbf{E}}(\omega), \tag{46}$$

$$\hat{\mathbf{B}}(\omega) = \mu_0(1 + \hat{\chi}_M(\omega))\hat{\mathbf{E}}(\omega) =: \mu(\omega)\hat{\mathbf{E}}(\omega). \tag{47}$$

For real-valued ω, $\epsilon(\omega)$ and $\mu(\omega)$ are the usual dielectric permittivity and the magnetic permeability, respectively. The assumption of $\chi_E(\mathbf{x}, \cdot)$, $\chi_M(\mathbf{x}, \cdot) \in L^1(\mathbb{R})$ leads to the fact that all the functions involved in (46) and (47) are analytic in $\mathbb{C}^+$ and continuous in the topological closure $cl(\mathbb{C}^+)= \mathbb{C}^+ \cup \mathbb{R}$. Moreover, by applying the Riemann–Lebesgue theorem to χ_E and χ_M, one has $\hat{\chi}_E(\omega) \to 0$ and $\hat{\chi}_M(\omega) \to 0$ as $|\omega| \to \infty$ in $cl(\mathbb{C}^+)$. Therefore, $\epsilon(\omega) \to \epsilon_0$ and $\mu(\omega) \to \mu_0$ as $cl(\mathbb{C}^+) \ni \omega \to \infty$. The passivity assumption that demands non-negative electric/magnetic energy loss is formulated as

$$\mathcal{E}_a(t) = \int_{-\infty}^{t} \int_{\Omega} \partial_t \mathbf{D}(\mathbf{x}, s) \cdot \mathbf{E}(\mathbf{x}, s) + \partial_t \mathbf{B}(\mathbf{x}, s) \cdot \mathbf{H}(\mathbf{x}, s) d\mathbf{x} ds \geq 0, t \in \mathbb{R}. \tag{48}$$

Then the Plancherel theorem implies that

$$\mathcal{E}_a(\infty) = \frac{1}{2\pi} Re \int_{\mathbb{R}} \int_{\Omega} -i\omega \left(\hat{\epsilon}(\mathbf{x}, \omega)|\hat{\mathbf{E}}(\mathbf{x}, \omega)|^2 + \hat{\mu}(\mathbf{x}, \omega)|\hat{\mathbf{H}}(\mathbf{x}, \omega)|^2 \right) d\mathbf{x} d\omega \geq 0.$$

Since this has to hold for all $\mathbf{E}$ and $\mathbf{H}$, it must be true that $\omega \mathrm{Im}\,(\epsilon(\omega)) \geq 0$, and $\omega \mathrm{Im}\,(\mu(\omega)) \geq 0$ for all real-valued ω. These properties of $\hat{\chi}_E$, $\hat{\chi}_M$, $\epsilon(\omega)$, and $\mu(\omega)$ prompt the study of functions $f : cl(\mathbb{C}^+) \to \mathbb{C}$ that satisfy the following hypotheses:

- H1: f is analytic in $\mathbb{C}^+$ and continuous in $cl(\mathbb{C}^+)$. (causality)
- H2: $f(z) \to f_\infty > 0$ as $|z| \to \infty$ in $cl(\mathbb{C}^+)$.
- H3: $f(-\overline{z}) = \overline{f(z)}$, $z \in cl(\mathbb{C}^+)$.
- H4: $\mathrm{Im}\, f(z) \geq 0$ for all $z \in \mathbb{R}^+$. (passivity)

Note that f satisfying hypotheses H1–H4 is not a Herglotz–Nevanlinna function; however, Remark 2 in Part I implies that the function $v(\omega) := \omega f(\sqrt{\omega})$ is, and this fact will be utilized below.

The problem of passive, quasi-static cloaking for incident plane waves is formulated in [23] as follows. Suppose the material in O has constant permittivity $\epsilon\mathbf{I}$ with $\epsilon > \epsilon_0$ and is non-dispersive (frequency independent) in the frequency range $[\omega_-, \omega_+]$. The cloak is assumed to have permittivity $\epsilon(\mathbf{x}, \omega)$ and occupies the space $\Omega \setminus O$ surrounding the inclusion O.

Also, it is assumed that the permittivity in $\mathbb{R}^3 \setminus \Omega$ is $\epsilon_0\mathbf{I}$. In the quasi-static case, the time derivatives in (43) are negligible, and hence, $\mathbf{E} = -\nabla V$ for some scalar potential V. Let the incident plane wave be $\mathbf{E}_0$, a uniform field in $\mathbb{R}^3$, which will interact with the device Ω, and the scattered field with potential V_s will be generated. The scattered potential V_s is related to the total potential by $V(\mathbf{x}, t) = -\mathbf{E}_0 \cdot \mathbf{x} + V_s(\mathbf{x}, t)$. V_s satisfies the equation

$$\begin{cases} \nabla \cdot (\boldsymbol{\epsilon}(\mathbf{x}, \omega)\nabla V_s) = \nabla \cdot (\boldsymbol{\epsilon}(\mathbf{x}, \omega) - \epsilon_o\mathbf{I})\mathbf{E}_0 \text{ in } \mathbb{R}^3, \\ \qquad\qquad V_s(\mathbf{x}, \omega) = O(1/|\mathbf{x}|) \text{ as } |\mathbf{x}| \to \infty. \end{cases} \tag{49}$$

Because the cloak occupying $\Omega \setminus O$ is assumed to be passive, the permittivity $\boldsymbol{\epsilon}(\mathbf{x}, \omega)$ satisfies the following conditions for almost all $\mathbf{x} \in \Omega \setminus O$:

- $\widetilde{\mathrm{H}}1$: $\boldsymbol{\epsilon}(\mathbf{x}, \cdot)$ is analytic on $\mathbb{C}^+$ and continuous on $cl(\mathbb{C}^+)$.
- $\widetilde{\mathrm{H}}2$: $\boldsymbol{\epsilon}(\mathbf{x}, \omega) \to \epsilon_0\mathbf{I}$ as $|\omega| \to \infty$ in $cl(\mathbb{C}^+)$.
- $\widetilde{\mathrm{H}}3$: $\boldsymbol{\epsilon}(\mathbf{x}, -\overline{\omega}) = \overline{\boldsymbol{\epsilon}(\mathbf{x}, \omega)}\, \forall \omega \in cl(\mathbb{C}^+)$.
- $\widetilde{\mathrm{H}}4$: $\mathrm{Im}\, \boldsymbol{\epsilon}(\mathbf{x}, \omega) \geq 0\, \forall \omega \in \mathbb{R}^+$.
- $\widetilde{\mathrm{H}}5$: $\boldsymbol{\epsilon}(\mathbf{x}, \omega)^T = \boldsymbol{\epsilon}(\mathbf{x}, \omega)$, $\forall \omega \in cl(\mathbb{C}^+)$ (reciprocity principle).

For the well-posedness of (49), two additional conditions are imposed.

- $\widetilde{\mathrm{H}}6$: $\boldsymbol{\epsilon}(\cdot, \omega) \in L^\infty(\Omega \setminus O)\, \forall \omega \in cl(\mathbb{C}^+)$ such that $\sup_{\omega \in cl(\mathbb{C}^+)} \|\boldsymbol{\epsilon}(\cdot, \omega)\|_{L^\infty(\Omega\setminus O)} \leq c_1$ with positive constant c_1.
- $\widetilde{\mathrm{H}}7$a: There exist $c_2(\omega) > 0$ and $\gamma(\omega) \in [0, 2\pi)$ such that for all $\omega \in \mathbb{C}^+$, one has $|\mathrm{Im}\, (e^{i\gamma(\omega)}\boldsymbol{\epsilon}(\mathbf{x}, \omega)\mathbf{E} \cdot \overline{\mathbf{E}})| \geq c_2(\omega)\|\mathbf{E}\|^2$, $\forall \mathbf{E} \in \mathbb{C}^3$, and for almost all $\mathbf{x} \in \Omega \setminus O$.
- $\widetilde{\mathrm{H}}7$b: For all $\omega_0 \in \mathbb{R}$, there exist $c_3(\omega_0) > 0$, $\delta > 0$, and $\gamma(\omega_0) \in [0, 2\pi)$ such that for all $\omega \in B(\omega_0, \delta) \cap cl(\mathbb{C}^+)$, one has $|\mathrm{Im}\, (e^{i\gamma(\omega_0)}\boldsymbol{\epsilon}(\mathbf{x}, \omega)\mathbf{E} \cdot \overline{\mathbf{E}})| \geq c_3(\omega_0)\|\mathbf{E}\|^2$, $\forall \mathbf{E} \in \mathbb{C}^3$, and for almost all $\mathbf{x} \in \Omega \setminus O$. Here $B(\omega_0, \delta)$ is the disk of radius δ centered at ω_0.

With these assumptions on $\boldsymbol{\epsilon}(\mathbf{x}, \omega)$, it is shown in [23] that the potential of the total electric field that satisfies the condition $V(\mathbf{x}, \omega) = -\mathbf{E}_0 \cdot \mathbf{x} + O(1/|\mathbf{x}|)$ as $|\mathbf{x}| \to \infty$ is of the form

$$V(\mathbf{x}, \omega) = -\mathbf{E}_0 \cdot \mathbf{x} + \frac{(\boldsymbol{\alpha}(\omega)\mathbf{E}_0) \cdot \mathbf{x}}{4\pi\epsilon_0|\mathbf{x}|^3} + O(1/|\mathbf{x}|^3) \tag{50}$$

for all $\omega \in cl(\mathbb{C}^+) \cup \{\infty\}$, where the complex-valued 3×3 polarizability tensor $\boldsymbol{\alpha}(\omega)$ is given by

$$\boldsymbol{\alpha}(\omega)\mathbf{E}_0 = \int_\Omega (\boldsymbol{\epsilon}(\mathbf{x},\omega) - \epsilon_0\mathbf{I})(\mathbf{E}_0 - \nabla V_s(\mathbf{x},\omega))\, d\mathbf{x}. \tag{51}$$

The key point is that $\boldsymbol{\alpha}(\omega)$ describes the leading term of the far-field scattered field. Hence, the broadband cloaking of the dielectric inclusion O in the frequency interval $[\omega_-, \omega_+]$ is achieved when $\boldsymbol{\alpha}(\omega)$ vanishes for *all* $\omega \in [\omega_-, \omega_+]$. It is proved in [23] that if $\boldsymbol{\epsilon}(\mathbf{x},\omega)$ satisfies the hypotheses $\widetilde{H}1$-$\widetilde{H}7b$, then the function $f(\omega) := \boldsymbol{\alpha}(\omega)\mathbf{E}_0 \cdot \overline{\mathbf{E}_0}$ satisfies hypotheses H1–H4. Consequently,

$$v(\omega) := \omega f(\sqrt{\omega}) = \omega\boldsymbol{\alpha}(\sqrt{\omega})\mathbf{E}_0 \cdot \overline{\mathbf{E}_0} \tag{52}$$

is a Herglotz–Nevanlinna function analytic in $\mathbb{C} \setminus \mathbb{R}^+$ and negative in $\mathbb{R}^-$. Note that v is not a Stieltjes function since it has the "wrong sign" on the negative halfline. Furthermore, $f(\infty) = \boldsymbol{\alpha}(\infty)\mathbf{E}_0 \cdot \overline{\mathbf{E}_0}$ holds, which is positive for any non-zero field $\mathbf{E}_0$. This immediately leads to the conclusion that $\boldsymbol{\alpha}(\omega)$ cannot vanish in any interval $[x_-, x_+]$ with $x_-, x_+ \in \mathbb{R}^+$ and $x_- \neq x_+$. Because if it does vanish, so does f; then the Schwarz reflection principle and the analytic continuation imply f is identically zero in $\mathbb{C}^+$, which contradicts the fact $f(\infty) > 0$. Therefore, broadband cloaking is not possible for a quasi-static passive cloak.

We conclude this section by explaining the main ingredients in the derivation of a more refined quantification of the fundamental limits of broadband passive cloaking in quasi-statics presented in [23].

Since the polarizability tensor $\boldsymbol{\alpha}(\omega)$ with real-valued ω is of interest in physics, the Herglotz–Nevanlinna function setting is applied to extract information from the behavior of $\boldsymbol{\alpha}(\omega)$ as a function in $cl(\mathbb{C}^+)$ to conclude something useful for its behavior on the positive real line. One important tool for making this connection is the sum rule, as stated in Theorem 10 in Part I, which is applied to the composition with an appropriate window function. This same technique is also used in Sect. 3.1 in Part I.

To be able to focus on a finite interval $[-\triangle, \triangle] \subset \mathbb{R}$, $\triangle > 0$, the function h_m is defined as

$$h_m(z) := \int_{-\triangle}^{\triangle} \frac{dm(\xi)}{\xi - z}, \tag{53}$$

where m belongs to $\mathcal{M}_\triangle$, the set of finite positive Borel measure supported in $[-\triangle, \triangle]$ such that $m([-\triangle, \triangle]) = 1$. Obviously, $h_m(z)$ is a Herglotz–Nevanlinna function. By using the theorems in [15] and [23], the following asymptotic behavior can be concluded

$$h_m(z) = -\frac{m(\{0\})}{z} + o\left(\frac{1}{z}\right), \text{ as } |z|\hat{\to}0 \text{ and } h_m(z) = -\frac{1}{z} + o\left(\frac{1}{z}\right), \text{ as } |z|\hat{\to}\infty.$$

Since for any function f that satisfies H1–H4, the corresponding function $v(z) := zf(\sqrt{z})$ is a Herglotz–Nevanlinna function in $\mathbb{C} \setminus \mathbb{R}^+$ and negative in $\mathbb{R}^-$, the composition $v_m(z) := h_m(v(z))$ is again a Herglotz–Nevanlinna function with the following asymptotic expansion:

$$v_m(z) = -\frac{m(\{0\})}{f(0)z} + o\left(\frac{1}{z}\right), \text{ as } |z| \hat{\to} 0 \text{ and } v_m(z) = -\frac{1}{f_\infty z} + o\left(\frac{1}{z}\right), \text{ as } |z| \hat{\to} \infty.$$

Theorem 10 in Part I for $n = 0$ immediately implies that for any given finite interval $[x_-, x_+] \subset \mathbb{R}$ and any $m \in \mathcal{M}_\triangle$, one has

$$\begin{aligned}\lim_{y\to 0^+} \frac{1}{\pi} \int_{x_-}^{x_+} \operatorname{Im} v_m(x+iy)dx &\leq \lim_{\eta\to 0^+} \lim_{y\to 0^+} \frac{1}{\pi} \int_{\eta<|x|<\eta^{-1}} \operatorname{Im} v_m(x+iy)dx \\ &= \frac{1}{f_\infty} - \frac{m(\{0\})}{f(0)} \leq \frac{1}{f_\infty}. \end{aligned} \tag{54}$$

If the cloak is *lossy* in the finite band $[\omega_-, \omega_+]$, i.e., $\operatorname{Im} \boldsymbol{\epsilon}(\mathbf{x}, \omega)$ in (51) is not negligible in $[\omega_-, \omega_+]$, then (52) implies $\operatorname{Im} v(x)$ and hence $\operatorname{Im} v_m$ is not negligible for $x \in [\omega_-^2, \omega_+^2] := [x_-, x_+]$. The choice of $dm(\xi) = \frac{\mathbf{1}_{[-\triangle,\triangle]}(\xi)}{2\triangle} d\xi$ results in $h_m(z) = \frac{1}{2\triangle} \log \frac{z-\triangle}{z+\triangle}$ for all $z \in \mathbb{C}^+$ (branch cut at $\mathbb{R}^+$). Consequently, a lower bound of $\operatorname{Im} h_m(z)$ can be easily derived to be $\operatorname{Im} h_m(z) \geq \frac{\pi}{4\triangle} H(\triangle - |z|)$, where H is the Heaviside function.

Taking into account the sum rule in (54), one has

$$\lim_{y\to 0^+} \frac{\pi}{4\triangle} \int_{x_-}^{x_+} H(\triangle - |v(x+iy)|)dx \leq \lim_{y\to 0^+} \int_{x_-}^{x_+} \operatorname{Im} v_m(x+iy)dx \leq \frac{\pi}{f_\infty}.$$

Applying the Lebesgue Dominated Convergence theorem to the left side leads to $\int_{x_-}^{x_+} H(\triangle - |v(x)|)dx \leq \frac{4\triangle}{f_\infty}$. Finally, letting $\triangle = \max_{x_- \leq x \leq x_+} |v(x)|$ in the previous inequality leads to the bound $\frac{1}{4}(x_+ - x_-) f_\infty \leq \max_{x\in[x_-,x_+]} |v(x)|$. By identifying $x = \omega^2$, the inequality can be directly translated to the following bound on the polarizability tensor in the frequency band

$$\frac{1}{4}(\omega_+^2 - \omega_-^2)\boldsymbol{\alpha}(\infty)\mathbf{E}_0 \cdot \overline{\mathbf{E}_0} \leq \max_{\omega\in[\omega_-,\omega_+]} \left| \omega^2 \boldsymbol{\alpha}(\omega)\mathbf{E}_0 \cdot \overline{\mathbf{E}_0} \right|.$$

Suppose the cloak has a transparent window in the band $[\omega_-, \omega_+]$, i.e., $\operatorname{Im} \boldsymbol{\epsilon}(\mathbf{x}, \omega) = 0$ for $\omega \in [\omega_-, \omega_+]$ for almost all $\mathbf{x} \in \Omega \setminus \mathcal{O}$. Then the corresponding $v(z)$ in (52) is real-valued for $z \in [\omega_-^2, \omega_+^2] := [x_-, x_+]$ because of (49) and (51). In this case, more refined bounds can be derived because first of all, $v(z)$ can be extended to be an analytic function in $D := \mathbb{C} \setminus \{[0, x_-] \cup [x_+, \infty)\}$. By letting the measure used in h_m be a Dirac measure $m = \delta_\zeta$ with $\zeta = v(x_0)$ for some $x_0 \in (x_-, x_+)$ and choosing $\triangle$ so that $-\triangle < \xi < \triangle$, one has

$v_{\delta_\zeta}(z) = \frac{1}{v(x_0)-v(z)}$, which is a Herglotz–Nevanlinna function with a real pole at x_0 and hence must be of multiplicity 1. Therefore, $v'(x_0) \neq 0$. Moreover, this pole must be isolated because $v(x_0) - v(z)$ is analytic in D. Therefore, there must exist a small neighborhood $\mathcal{N}$ around x_0, where v_{δ_ζ} can be expressed as $v_{\delta_\zeta}(z) = \frac{g(z)}{z-x_0}$ with $g(z)$ analytic in $\mathcal{N}$, $g(x_0) = -\frac{1}{v'(x_0)}$ and real-valued in $\mathcal{N} \cap [x_-, x_+] =: (a, b)$. The sum rule (54) implies $\lim_{y\to 0^+} \int_a^b \operatorname{Im} v_{\delta_\zeta}(x + iy)dx \leq \frac{\pi}{f_\infty}$. On the other hand, explicit calculation using Sokhotski-Plemeli formula can be performed to get $\lim_{y\to 0^+} \int_a^b \operatorname{Im} v_{\delta_\zeta}(x + iy)dx = -\pi g(x_0) = \frac{\pi}{v'(x_0)}$. So one has $0 < f_\infty \leq v'(x_0)$ for all $x_0 \in [x_-, x_+]$. This implies $f_\infty \cdot (x_1 - x_2) \leq v(x_1) - v(x_2)$ for any $x_1, x_2 \in [x_-, x_+]$ such that $x_2 < x_1$. Suppose $v(x_2) = 0$, then $v(x_1) \geq f_\infty \cdot x_1$ for all $x_1 > x_2$. Similarly, if $v(x_1) = 0$, then $v(x_2) \leq -f_\infty \cdot x_1$ for all $x_2 < x_1$ in $[x_-, x_+]$. Therefore, even if $\boldsymbol{\alpha}$ is zero at $\omega_0 \in [\omega_-, \omega_+]$, one will have $\boldsymbol{\alpha}(\omega) \leq -\boldsymbol{\alpha}(\infty)\frac{\omega_0^2-\omega^2}{\omega^2}$ if $\omega_- \leq \omega < \omega_0$ and $\boldsymbol{\alpha}(\omega) \geq \boldsymbol{\alpha}(\infty)\frac{\omega_0^2-\omega^2}{\omega^2}$ if $\omega_0 < \omega \leq \omega_+$. Thus it is impossible to achieve the broadband passive quasi-static cloaking (BPQC) in a transparent window.

In conclusion, for the BPQC problem, the Herglotz–Nevanlinna function structure of the function $v(\omega)$ in (52) and the accompanied sum rules not only lead to a proof that BPQC is impossible but also give quantitative limitations of BPQC through providing useful lower bounds.

2.4 Hamiltonian Structure of Time Dispersive and Dissipative Systems

Wave dissipation and dispersion appear in many materials. For example, the dynamic tortuosity describes both the dissipation and dispersion mechanisms for the poroelastic materials. Also, the dispersive nature of the Maxwell's equations is revealed by the frequency-dependent permittivity, permeability, and the susceptibility functions in (46) and (47). These are examples of linear time dispersive and dissipative (TDD) systems. In a series of work, Figotin and Schenker [38–40] developed a framework for studying the Hamiltonian structure of the linear TDD. Specifically, they consider problems of the following form in a Hilbert space setting

$$m\partial_t \mathbf{v}(t) = -i\mathbf{A}\mathbf{v}(t) - \int_0^\infty \mathbf{a}(\tau)\mathbf{v}(t-\tau)d\tau + \mathbf{f}(t), \tag{55}$$

where $m > 0$ is a positive mass operator in a Hilbert space H_0, $\mathbf{A}$ is a self-adjoint operator in H_0, $\mathbf{f}(t) \in H_0$ is a generalized external force, and $\mathbf{a}(t)$ is an operator-valued retarded friction function that satisfies $\mathbf{a}(t) = 0$ for $t < 0$. The total work done by f is $W = \int_{-\infty}^\infty \operatorname{Re}\{(\mathbf{v}(t), \mathbf{f}(t))\}dt$. The term $-i\mathbf{A}\mathbf{v}(t) - \int_0^\infty \mathbf{a}(\tau)\mathbf{v}(t-\tau)d\tau$ is interpreted as the force that v exerts on itself at time t with $-i\mathbf{A}\mathbf{v}(t)$ regarded as the instantaneous term. The time dispersive integral term $\int_0^\infty \mathbf{a}(\tau)\mathbf{v}(t-\tau)d\tau$ is based on two fundamental requirements of *time homogeneity* and *causality*. As a simple example, in [38], the authors consider a non-magnetic medium by setting $\mathbf{J}_B = 0, \mathbf{H} = \mathbf{B}$ (hence $\mathbf{M} = 0$), $\nabla \cdot \mathbf{J} = 0$, $\nabla \cdot \mathbf{E} = 0$ and $\mu_0 = \epsilon_0 = 1$ in (43)–(45). The corresponding TDD for this case is

$$\mathbf{v}(\mathbf{x}, t) = \begin{pmatrix} \mathbf{E}(\mathbf{x}, t) \\ \mathbf{B}(\mathbf{x}, t) \end{pmatrix} \in H_0$$

$$H_0 := \{\mathbf{v} \in L^2(\mathbb{C}^6) | \nabla \cdot \mathbf{E} = 0 = \nabla \cdot \mathbf{B}\}$$

$$\mathbf{A} = \begin{pmatrix} 0 & i\nabla\times \\ -i\nabla\times & 0 \end{pmatrix}, \ \mathbf{a}(t) = \begin{pmatrix} \partial_t \chi_E \mathbf{I}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \end{pmatrix}, \ \mathbf{f}(t) = \begin{pmatrix} \mathbf{J} \\ \mathbf{0} \end{pmatrix}.$$

The difficulty of studying the spectral theory a TDD system can be easily seen by considering the time-frequency Fourier transform $\hat{v}(\omega) := \int_{-\infty}^{\infty} e^{i\omega t} v(t) dt$. Note that here the Fourier transform is denoted in the same way as the Laplace–Fourier transform in the preceding sections. Then the TDD system (55) becomes

$$\omega m \hat{\mathbf{v}}(\omega) = (\mathbf{A} - i\hat{\mathbf{a}}(\omega))\hat{\mathbf{v}}(\omega) + i\hat{\mathbf{f}}(\omega) =: \hat{\mathbf{A}}(\omega)\hat{\mathbf{v}}(\omega) + i\hat{\mathbf{f}}(\omega).$$

The Kramers–Kronig relations imply that $\hat{\mathbf{A}}$ is non-self-adjoint as long as $\hat{\mathbf{a}} \neq 0$, and hence, the eigenvectors of the problem $\omega m \mathbf{e}_\omega = (\mathbf{A} - i\hat{\mathbf{a}}(\omega))\mathbf{e}_\omega$ are not necessarily orthogonal for distinct ω and may not form a basis for H_0. This challenge can be addressed by the method of conservative extension of the TDD system [38] by first noting that $\hat{\mathbf{A}}$ is not an arbitrary non-self-adjoint operator because the friction operator $\hat{\mathbf{a}}$ has to satisfy certain characteristic properties of physical laws. Also, as is pointed out in [38–40], for all DD systems that are *physical*, the frequency dependence of $\hat{\mathbf{a}}$ originates from ignoring its coupling with another system, whose variables are referred to as **the hidden degree of freedom** of the TDD system. Hence by finding the coupling system, there will be a Hamiltonian structure of the *extended* system, which consists of the original TDD system and the coupling system.

Based on this idea, a coupled system is introduced

$$m\partial_t \mathbf{v}(t) = -i\mathbf{A}\mathbf{v}(t) - i\mathbf{\Gamma}\mathbf{w}(t) + \mathbf{f}(t) \tag{56}$$

$$\partial_t \mathbf{w}(t) = -i\mathbf{\Gamma}^\dagger \mathbf{v}(t) - i\mathbf{\Omega}_1 \mathbf{w}(t), \ \mathbf{\Omega}_1 \text{ is self-adjoint in } H_1, \tag{57}$$

where H_1 denotes the Hilbert space of the hidden variables $\mathbf{w}$, and $\mathbf{\Gamma} : H_1 \to H_0$ the coupling operator between the hidden variable $\mathbf{w}$ and the observable variable $\mathbf{v}$. Following the notation in the papers reviewed here, $\mathbf{\Gamma}^\dagger$ denotes the adjoint of $\mathbf{\Gamma}$. This extended system should give the original TDD (55) after eliminating $\mathbf{w}$. Note that the second equation implies $\mathbf{w} = -i \int_0^\infty e^{-i\mathbf{\Omega}_1 \tau} \mathbf{\Gamma}^\dagger \mathbf{v}(t - \tau) d\tau$. Using this to eliminate $\mathbf{w}$ in the first equation leads to the necessary condition

$$\mathbf{a}(t) = \mathbf{\Gamma} e^{-i\mathbf{\Omega}_1 t} \mathbf{\Gamma}^\dagger, \ t > 0. \tag{58}$$

This spectral representation of the friction function $\mathbf{a}(t)$ indicates how the unknowns ($\mathbf{w}$, $\mathbf{\Omega}_1$, H_1) of the desired conservative extension can be recovered from the given $\mathbf{a}(t)$. Suppose $\mathbf{a}(t)$ has the general form

$$\mathbf{a}(t) = \boldsymbol{\alpha}_\infty \delta(t) + \boldsymbol{\alpha}(t),$$

where $\boldsymbol{\alpha}_\infty = \boldsymbol{\alpha}_\infty^\dagger \geq 0$, $\delta(t)$ is the Dirac function, and $\alpha(t)$ is for every $t \geq 0$ a bounded non-negative operator in H_0 such that

$$0 \leq \boldsymbol{\alpha}_\infty \leq C\mathbf{I}_{H_0}, C < \infty, \text{ and } \sup_{t\geq 0} \|\boldsymbol{\alpha}(t)\|_{B(H_0)} < \infty. \tag{59}$$

Note that $\boldsymbol{\alpha}_\infty$ corresponds to the classic and familiar friction constant. Then (58) implies that $\mathbf{\Gamma}$ is unbounded if $\boldsymbol{\alpha}_\infty \neq \mathbf{0}$ because $\mathbf{a}(0) = \mathbf{\Gamma}\mathbf{\Gamma}^\dagger = \boldsymbol{\alpha}_\infty \delta(t)$. Moreover, $\mathbf{a}(t)$ is extended to $t \leq 0$ by

$$\mathbf{a}_e(t) = \mathbf{\Gamma} e^{-i\mathbf{\Omega}_1 t} \mathbf{\Gamma}^\dagger, \; -\infty < t < \infty. \tag{60}$$

Note that $\mathbf{a}_e(-t) = \mathbf{a}_e^\dagger(t)$. As a result, the following power dissipation condition must hold

$$\begin{aligned} \mathcal{W}_{fr}(\mathbf{v}) &:= -\frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\mathbf{v}(t), \mathbf{a}_e(t-\tau)\mathbf{v}(\tau))dtd\tau \\ &= -\frac{1}{2}\int_{-\infty}^{\infty} \|e^{i\Omega_1 t}\mathbf{\Gamma}^\dagger \mathbf{v}(t)dt\|^2 \leq 0. \end{aligned} \tag{61}$$

This power dissipation condition is also a sufficient condition for the existence of a conservative extension for a TDD system [38]. The construction of the conservative extension involves finding the essentially unique triplet $(H_1, \mathbf{\Gamma}, \mathbf{\Omega}_1)$ from the extended friction operator $\mathbf{a}_e$. Reconstruction in the time domain can be carried out by using Bochner's theorem. However, due to the unboundedness of the operator $\mathbf{\Gamma}$ for the general case $\boldsymbol{\alpha}_\infty \neq \mathbf{0}$, the time-domain reconstruction of the triplet involves subtle technicalities for dealing with the unbounded operator; see [38]. On the other hand, as is pointed out also in [38], if one formulates the reconstruction problem in the *complex* frequency form, there will be no unbounded operator involved. The intuition is based on the observation that the Fourier transform of the friction function $\mathbf{a}(t)$ is $\hat{\mathbf{a}}(\zeta) = \boldsymbol{\alpha}_\infty + \hat{\boldsymbol{\alpha}}(\zeta)$, which is an analytic operator function. Assume $\mathbf{v}(t) = 0$ and $\mathbf{f}(t) = 0$ for $t \leq 0$. In this setting, the first step is to Laplace-Fourier transform the TDD problem (55) to obtain the following linear response equation:

$$(\zeta m - \mathbf{A} + i\hat{\mathbf{a}}(\zeta))\hat{\mathbf{v}}(\zeta) =: i\mathfrak{A}(\zeta)^{-1} = i\hat{\mathbf{f}}(\zeta), \; \operatorname{Im}\zeta > 0. \tag{62}$$

The power dissipation condition (61) becomes

$$\operatorname{Re}\hat{\mathbf{a}}(\zeta) := \frac{\hat{\mathbf{a}}(\zeta) + \hat{\mathbf{a}}(\zeta)^\dagger}{2} \geq 0 \text{ for } \operatorname{Im}\zeta > 0,$$

which implies $(\zeta m-[\mathbf{A}-i\hat{\mathbf{a}}(\zeta)])$ is invertible for $\operatorname{Im}\zeta > 0$ because $\mathbf{A}$ is self-adjoint. Define the admittance operator as $\mathfrak{A}(\zeta) := i(\zeta m - \mathbf{A} + i\hat{\mathbf{a}}(\zeta))^{-1}$ for $\operatorname{Im}\zeta > 0$. Note then that both the operator-valued functions $i\hat{\mathbf{a}}(\zeta)$ and $i\mathfrak{A}(\zeta)$ are Herglotz–Nevanlinna functions. The equation above can be written as the admittance equation

$$\hat{\mathbf{v}}(\zeta) = \mathfrak{A}(\zeta)\hat{\mathbf{f}}(\zeta),\ \operatorname{Im}\zeta > 0.$$

From the definition of the admittance operator $\mathfrak{A}$, it is clear that one can recover m, $\mathbf{A}$, and $\hat{\mathbf{a}}$ from $\mathfrak{A}$ as follows:

$$m^{-1} = -\lim_{\eta\to\infty}\eta\mathfrak{A}(i\eta),\ \mathbf{A} = -\lim_{\eta\to\infty}\operatorname{Im}\mathfrak{A}^{-1}(i\eta),\ \hat{\mathbf{a}}(\zeta) = i(\zeta m - \mathbf{A}) + \mathfrak{A}^{-1}(\zeta).$$

To identify the spectral decomposition of $\hat{\mathbf{a}}$, one applies the same transform to the conserved system (57) and eliminates $\hat{\mathbf{w}}$ to obtain $\mathfrak{A}(\zeta) = i[\zeta m - \mathbf{A} - \mathbf{\Gamma}(\zeta\mathbf{I}_{H_1} - \mathbf{\Omega}_1)^{-1}\mathbf{\Gamma}^\dagger]^{-1}$. A comparison with the admittance operator defined by (62) reveals that

$$\hat{\mathbf{a}}(\zeta) = i\mathbf{\Gamma}(\zeta\mathbf{I}_{H_1} - \mathbf{\Omega}_1)^{-1}\mathbf{\Gamma}^\dagger. \tag{63}$$

Besides the power dissipation condition $\operatorname{Re}\hat{\mathbf{a}}(\zeta) \geq 0$ for $\operatorname{Im}\zeta > 0$, the condition (59) implies

$$\hat{\mathbf{a}}(\zeta) = \alpha_\infty + \hat{\alpha}(\zeta),\ \|\hat{\alpha}(\zeta)\|_{B(H_0)} \leq \frac{\sup_{t\geq 0}\|\alpha(t)\|_{B(H_0)}}{\operatorname{Im}\zeta}.$$

This implies that Theorem 14 in Part I can be applied to show the existence of the space of the hidden variables and the operators in the spectral decomposition (63). Below this theorem is formulated as it is in Theorem 3.13 in [38]. Note that $\mathbf{\Omega_1}$ and $\mathbf{\Gamma}^\dagger$ here correspond to A and Γ_0, respectively, in Eq. (24) of Part I.

Theorem 4 *Let $G(\zeta)$ be a $B(H_0)$-valued analytic function in $\mathbb{C}^+$ with $\operatorname{Im}G(\zeta) \geq 0$ for $\zeta \in \mathbb{C}^+$. If G satisfies the growth condition $\limsup_{\eta\to+\infty}\eta\|G(i\eta)\| < \infty$, then G has the following representation:*

$$G(\zeta) = \mathbf{\Gamma}(\mathbf{\Omega}_1 - \zeta\mathbf{I}_{H_1})^{-1}\mathbf{\Gamma}^\dagger \tag{64}$$

with $\mathbf{\Omega}_1$ a self-adjoint operator on a Hilbert space H_1 and $\mathbf{\Gamma} : H_1 \to H_0$ a bounded map such that

$$\mathbf{\Gamma}\mathbf{\Gamma}^\dagger\mathbf{v} = \lim_{\eta\to+\infty} -i\eta G(i\eta)\mathbf{v}\,\textit{for all}\ \mathbf{v} \in H_0.$$

If H_1 is minimal in the sense that $\{f(\mathbf{\Omega}_1)\mathbf{\Gamma}^\dagger\mathbf{v} : f \in C_c(\mathbb{R}),\ v \in H_0\}$ is dense in H_1, then $\{H_1, \mathbf{\Omega}_1, \mathbf{\Gamma}\}$ is uniquely determined up to an isomorphism.

By identifying $G(\zeta) = i\hat{\mathbf{a}}(\zeta)$ in the theorem, we see that (63) has a unique solution up to an isomorphism. Therefore, the conservative extension of (62) exists. In [38], the extended system for dielectric Maxwell's equations with a Lorentzian susceptibility function χ is constructed. The Hamiltonian structure of the TDD can then be studied via the Hamiltonian structure of the conservative extended system.

3 More General Classes of Functions

As we have seen in the preceding sections, there are a wide range of applications where Herglotz–Nevanlinna functions are a valuable tool. However, there are also many situations where Herglotz–Nevanlinna functions do not suffice. If, for instance, a causal system is not passive, then the corresponding analytic function will not have positive imaginary part. Or if a composite material does consist of more than two materials, then the corresponding function will depend on more than only one variable.

On the mathematical side, the class of Herglotz–Nevanlinna functions has been generalized in several directions. To give a short overview, we will concentrate on scalar generalizations only, even if some results do hold for matrix or operator functions as well.

3.1 Quasi-Herglotz Functions

The class of Herglotz–Nevanlinna functions forms a cone (as it is closed under linear combinations with non-negative coefficients) but not a vector space (since multiples with coefficents other than non-negative do not preserve the Herglotz–Nevanlinna property). As also differences of Herglotz–Nevanlinna functions do appear in applications, the class of quasi-Herglotz functions has been introduced, see [48]. For more details concerning this section, see [63].

Definition 2 A function $q : \mathbb{C} \setminus \mathbb{R} \to \mathbb{C}$ is called a quasi-Herglotz function if it can be written in the form $q = h_1 - h_2 + i(h_3 - h_4)$, where h_i for $1 = 1, 2, 3, 4$ are Herglotz–Nevanlinna functions (symmetrically extended to the lower halfplane).

Example 1 Every analytic function $q : \mathbb{C}^+ \to \mathbb{C}$ with $\operatorname{Im} q(z) \geq -c$ for some $c > 0$ is a quasi-Herglotz function, since it can be written in the form $q(z) = (q(z) + ic) - ic$, with both $q + ic$ and ic Herglotz–Nevanlinna functions.

It is obvious from the definition that this class coincides with all linear combinations of Herglotz–Nevanlinna functions. Hence, these functions can also be characterized in terms of an integral representation, however, with complex measures. Recall that complex measures by definition are finite, see, e.g., [81, Chapter 6], and hence, the representation of the form of Eq. (2) in Part I is used.

Proposition 1 *A function q is a quasi-Herglotz function if and only if there exist real numbers a and b and a complex measure σ such that*

$$f(z) = a + bz + \int_{\mathbb{R}} \frac{1+\xi z}{\xi - z} d\sigma(\xi). \tag{65}$$

Moreover, a, b, and σ are unique with this property.

Note that quasi-Herglotz functions by definition are defined in both the upper and lower halfplanes. In contrast to Herglotz–Nevanlinna functions, the values in one halfplane do not determine the values in the other.

Example 2 The functions

$$q_1(z) = \begin{cases} i & \operatorname{Im} z > 0 \\ -i & \operatorname{Im} z < 0 \end{cases} \quad \text{and} \quad q_2(z) = \begin{cases} i & \operatorname{Im} z > 0 \\ i & \operatorname{Im} z < 0 \end{cases}$$

do coincide in the upper halfplane, but not in the lower. Both are quasi-Herglotz functions as they can be written in the form (65) with $(a_1, b_1, \mu_1) = (0, 0, \frac{1}{\pi} d\lambda_{\mathbb{R}})$ for q_1 (as in Example 2 of Part I) and $(a_2, b_2, \mu_2) = (i, 0, 0)$ for q_2.

Considering the difference of the two functions in the example above shows that there are non-trivial quasi-Herglotz functions vanishing identically in one halfplane. All these have been characterized in [63].

Given a function, neither the definition nor the characterization in Proposition 1 is practical to check whether it is a quasi-Herglotz function or not. But these functions can also be characterized by their analytic properties.

Theorem 5 *Let $q\colon \mathbb{C} \setminus \mathbb{R} \to \mathbb{C}$ be a holomorphic function. Then q is a quasi-Herglotz function if and only if the function q satisfies, first, a growth condition, namely, that there exists a number $M \geq 0$ such that for all $z \in \mathbb{C} \setminus \mathbb{R}$*

$$|q(z)| \leq M \frac{1 + |z|^2}{|\operatorname{Im} z|}, \tag{66}$$

and, second, the regularity condition

$$\sup_{y \in (0,1)} \int_{\mathbb{R}} \big| q(x+iy) - q(x-iy) \big| \frac{dx}{1+x^2} < \infty. \tag{67}$$

An important subclass are real quasi-Herglotz functions; these are real linear combinations of Herglotz–Nevanlinna functions or, equivalently, functions that admit an integral representation (65) with a signed (i.e., real) measure σ. It can be shown that these functions are exactly those, which are symmetric with respect to the real line, i.e., $q(\overline{z}) = \overline{q(z)}$.

It can be noted that quasi-Herglotz functions also appear naturally when dealing with Herglotz–Nevanlinna functions only, namely as the off-diagonal elements of matrix-valued Herglotz–Nevanlinna functions.

3.2 Generalized Nevanlinna Functions

In the definition of Herglotz–Nevanlinna functions, the sign of the imaginary part is required to be positive. However, using the integral representation, it can be shown that this is equivalent to the requirement that the so-called Nevanlinna kernel

$$N_f(z,w) := \frac{f(z) - \overline{f(w)}}{z - \overline{w}} \tag{68}$$

is positive. Recall that a kernel $N_f(z,w)$ is said to be positive definite if for any choice of $N \in \mathbb{N}$ and $z_1, \dots, z_N \in \mathcal{D}$, the matrix

$$\big(N(z_i, z_j)\big)_{i,j=1,\dots N}$$

is positive semidefinite.

This view leads to the following generalization by considering kernels with finitely many negative squares [56]. A kernel is said to have κ negative squares if every such matrix above has at most κ negative eigenvalues and κ is minimal with this property.

Definition 3 A function $q : \mathcal{D} \subset \mathbb{C}^+ \to \mathbb{C}$ is called a generalized Nevanlinna function if it is meromorphic in $\mathbb{C}^+$ and the Nevanlinna kernel N_q has finitely many negative squares. If this number is κ, then $q \in \mathcal{N}_\kappa$.

Generalized Nevanlinna functions do also admit an integral representation, but it is much more involved than Eq. (1) in Part I; see [56, Satz 3.1.].

The operator representation, however, carries over quite naturally. The only difference compared to Eq. (23) in Part I is that in this case the space is not a Hilbert space, but a Pontryagin space, that is a vector space equipped with an indefinite inner product, such that any non-positive subspace is finite-dimensional.

Theorem 6 *A function q is a generalized Nevanlinna function if and only if there exist a Pontryagin space $\mathcal{K}$, a self-adjoint linear relation A, a point $z_0 \in \mathbb{C}^+$, and an element $v \in \mathcal{K}$ such that*

$$q(z) = \overline{q(z_0)} + (z - \overline{z_0})\Big[(I + (z - z_0)(A - z)^{-1})v, v\Big]_{\mathcal{K}}. \tag{69}$$

Moreover, if $\mathcal{K} = \overline{span}\{(I + (z - z_0)(A - z)^{-1})v : z \in \varrho(A)\}$, then the representation is called minimal. In this case, $\mathcal{K}$ has κ negative squares if and only if $q \in \mathcal{N}_\kappa$ and the representation is unique up to unitary equivalence.

The conditions on the function q for simplified representations are literally the same as before, and Theorem 7 in Part I holds for generalized Nevanlinna functions as well.

From Theorem 6 and the spectral properties of self-adjoint relations in Pontryagin spaces, it follows directly that a generalized Nevanlinna function $q \in \mathcal{N}_\kappa$ has at most κ poles in the upper halfplane $\mathbb{C}^+$, and there are at most κ real points $\alpha \in \mathbb{R}$ (including ∞) where it does not hold that $\lim_{z \hat{\to} \alpha}(\alpha - z)q(z)$ exists as a non-negative number. These exceptional points (non-real and real) are exactly those eigenvalues, for which the corresponding eigenspace is not a positive subspace. These points are called generalized poles not of positive type. Generalized zeros not of positive type of q are by definition the generalized poles not of positive type of the inverse function $\hat{q}(z) := -\frac{1}{q(z)}$ (which belongs to the same class $\mathcal{N}_\kappa$ as q). The importance of these points becomes visible in the following characterization; see [35] and also [32].

Theorem 7 *A function q is a generalized Nevanlinna function if and only if there is a rational function r and a Herglotz–Nevanlinna function f such that*

$$q(z) = \overline{r(\overline{z})} f(z) r(z). \tag{70}$$

In this case, $q \in \mathcal{N}_\kappa$ if and only if $\deg r = \kappa$. *Moreover, r is of the form $r(z) = \dfrac{\prod_{i=1}^{\ell}(z-\alpha_i)}{\prod_{j=1}^{m}(z-\beta_j)}$, where α_i are the generalized zeros not of positive type and β_j are the generalized poles not of positive type of q and $\kappa = \max\{\ell, m\}$.*

Generalized Nevanlinna functions with polynomial r appear for instance in connection with Sturm–Liouville operators with strongly singular potentials or, more abstractly, with strongly singular perturbations of self-adjoint operators in Hilbert spaces, see, e.g., [34, 36, 57].

Remark 1 A generalized Nevanlinna function q does satisfy $\lim\limits_{z \hat{\to} x_0} \operatorname{Im} q(z) \geq 0$ (as a finite number or $+\infty$) for all but finitely many $x_0 \in \mathbb{R} \cup \{\infty\}$, cf., Remark 2 in Part I.

Remark 2 Also matrix- and operator-valued generalized Nevanlinna functions can be defined via a corresponding kernel condition. The operator representation in Theorem 13 of Part I carries over with the same changes as for scalar functions. The factorization, however, becomes a lot more delicate. The first part holds with rational factors $R(z)$ and $R(\overline{z})^*$, but these are not of a comparably simple form, in particular, since generalized poles and zeros can be at the same points. For details, see [60, 61].

3.3 Pseudo-Nevanlinna Functions

Definition 4 A function g is called pseudo-Nevanlinna if it can be written as the quotient of two bounded analytic functions (defined in $\mathbb{C}^+$) and satisfies $\lim\limits_{z\hat{\to}x_0} \operatorname{Im} g(z) \geq 0$ for almost all $x_0 \in \mathbb{R}$.

Note that every Herglotz–Nevanlinna function belongs to this class since it can be written as a fractional linear transformation of a function mapping $\mathbb{C}^+$ into the closed unit disc $\overline{\mathbb{D}}$. Moreover, by Remark 1, generalized Nevanlinna functions are also pseudo-Nevanlinna functions.

It has been shown in [30, 31] that pseudo-Nevanlinna functions can also be characterized via a factorization, extending Theorem 7. To this end, one needs to introduce the so-called density functions; these are particular pseudo-Nevanlinna functions, which are non-negative (or ∞) on the real line.

Theorem 8 *A function g is a pseudo-Nevanlinna function if and only if there exists a density function I and a Herglotz–Nevanlinna function g_0 such that $g(z) = I(z)g_0(z)$.*

To be precise, in [31], Pseudo-Caratheodory functions are studied; these are corresponding generalizations of Caratheodory functions, i.e., holomorphic functions mapping the open unit disk $\mathbb{D}$ to the closed right halfplane $\mathbb{C}_+ \cup i\mathbb{R}$. However, due to the topic of this text, here we consider the corresponding version for the upper halfplane.

The introduction of Pseudo-Caratheodory functions was motivated by problems arising in digital signal processing and in the theory of circuits and systems.

3.4 Functions in Several Variables

For analytic functions in one variable in the upper halfplane $\mathbb{C}^+$, there are two equivalent ways of defining Herglotz–Nevanlinna functions, either by the requirement that $\operatorname{Im} f(z)$ has to be non-negative or that the Nevanlinna kernel $N_f(z, w)$ has to be positive semidefinite. When considering functions in several variables, however, the generalizations along these two ways lead into different directions, in one case the functions are represented by some kind of resolvents, in the other case by integrals.

In the following, we use the notation $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ and consider analytic functions $H : (\mathbb{C}^+)^n \to \mathbb{C}$, that is H is analytic in each variable z_j for $j = 1, \ldots, n$.

3.4.1 Loewner Functions

Definition 5 A function $H : (\mathbb{C}^+)^n \to \mathbb{C}$ is called a Loewner function if it is holomorphic and there exist positive semidefinite kernels $A_1, \ldots, A_n$ on $(\mathbb{C}^+)^n$ such that

$$H(\mathbf{z}) - \overline{H(\mathbf{w})} = \sum_{j=1}^{n} (z_j - \overline{w_j}) A_j(\mathbf{z}, \mathbf{w}) \tag{71}$$

for all $\mathbf{z}, \mathbf{w} \in (\mathbb{C}^+)^n$.

Loewner functions with $n = 1$ are exactly Herglotz–Nevanlinna functions in one variable.

For $n > 1$, these functions have been characterized in different ways, in particular, as operator monotone functions; see [1]. It has also been shown that functions in this class admit an operator representation [2]. As an example, we give one result, corresponding to Theorem 7 in Part I with $s = 0$, in order to show the flavor of such representations.

Theorem 9 *A function $H : (\mathbb{C}^+)^n \to \mathbb{C}$ is a Loewner function satisfying*

$$\liminf_{y \to \infty} y |\operatorname{Im} H(iy, \ldots, iy)| < \infty$$

if and only if there exists a Hilbert space $\mathcal{H}$, a self-adjoint operator A in $\mathcal{H}$, positive contractions $Y_1, \ldots, Y_n$ with $Y_1 + \ldots + Y_n = I_{\mathcal{H}}$, and an element $v \in \mathcal{H}$ such that

$$H(\mathbf{z}) = \Big((A - z_1 Y_1 - \ldots - z_n Y_n)^{-1} v, v \Big)_{\mathcal{H}}.$$

For Loewner functions, transfer function realizations have also been established; see [8].

3.4.2 Herglotz–Nevanlinna Functions

The other way of considering several variables leads to the following, more general, definition.

Definition 6 A function $F : (\mathbb{C}^+)^n \to \mathbb{C}$ is called Herglotz–Nevanlinna function if it is holomorphic and $\operatorname{Im} F(\mathbf{z}) \geq 0$ for all $\mathbf{z} \in (\mathbb{C}^+)^n$.

It can be shown that not only for $n = 1$ but also for $n = 2$ the class of Herglotz–Nevanlinna functions does coincide with the class of Loewner functions. However, it is known that this is not true for $n > 2$. If $n > 2$, then every Loewner function is a Herglotz–Nevanlinna function, but not conversely.

For the larger class of Herglotz–Nevanlinna functions in several variables, a characterization via an integral representation has been shown. In order to formulate this result, we introduce the following notation. For $\mathbf{z} \in (\mathbb{C}^+)^n$ and $\mathbf{t} \in \mathbb{R}^n$, define

$$K_n(\mathbf{z}, \mathbf{t}) := i \left(\frac{2}{(2i)^n} \prod_{\ell=1}^{n} \left(\frac{1}{t_\ell - z_\ell} - \frac{1}{t_\ell + i} \right) - \frac{1}{(2i)^n} \prod_{\ell=1}^{n} \left(\frac{1}{t_\ell - i} - \frac{1}{t_\ell + i} \right) \right), \tag{72}$$

which for $n = 1$ coincides with the integrand in Eq. (1) of Part I.

Moreover, we say that a Borel measure μ on $\mathbb{R}^n$ satisfies the Nevanlinna condition if for all $\mathbf{z} \in (\mathbb{C}^+)^n$ and all indices $\ell_1, \ell_2 \in \{1, 2, \dots, n\}$ with $\ell_1 < \ell_2$, it holds

$$\int_{\mathbb{R}^n} \frac{1}{(t_{\ell_1} - z_{\ell_1})^2 (t_{\ell_2} - \bar{z_{\ell_2}})^2} \prod_{\substack{j=1 \\ j \neq \ell_1, \ell_2}}^{n} \left(\frac{1}{t_j - z_j} - \frac{1}{t_j - \bar{z}_j} \right) d\mu(\mathbf{t}) = 0. \tag{73}$$

Then the following theorem holds; see [62, Theorem 4.1].

Theorem 10 *A function $F : (\mathbb{C}^+)^n \to \mathbb{C}$ is a Herglotz–Nevanlinna function if and only if there exist a real number $a \in \mathbb{R}$, a vector $\mathbf{b} \in [0, \infty)^n$, and a positive Borel measure μ on $\mathbb{R}^n$ satisfying the Nevanlinna condition and with $\int_{\mathbb{R}^n} \prod_{\ell=1}^n \frac{1}{1+t_\ell^2} d\mu(\mathbf{t}) < \infty$ such that*

$$F(\mathbf{z}) = a + \sum_{\ell=1}^{n} b_\ell z_\ell + \frac{1}{\pi^n} \int_{\mathbb{R}^n} K_n(\mathbf{z}, \mathbf{t}) d\mu(\mathbf{t}). \tag{74}$$

Furthermore, for a given function F, the triple of representing parameters $(a, \mathbf{b}, \mu)$ is unique.

Note that for $n = 1$ the Nevanlinna condition is satisfied for every measure (which satisfies the necessary growth condition), and hence, this theorem becomes Theorem 1 in Part I. However, for $n > 1$, this condition is rather restrictive, and measures satisfying it are rather particular. For example, such a measure cannot have finite total mass and hence, in particular, not compact support. There are also other geometric restrictions on the support; see [64].

4 Summary

In this two-part survey paper, we start with introducing the various forms of Herglotz–Nevanlinna functions. These definitions are very simple to describe but imply many properties that are physically relevant. As can be seen from the diverse set of applications presented here, the Herglotz–Nevanlinna functions indeed provide a clear mathematical language for describing important physical properties such as passivity and causality. From there, a rigorous analysis can be applied to derive useful properties of these physical systems such as the bounds of effective properties of materials or to suggest a way to fabricate materials of desired properties through exploiting the links between some simple forms of Herglotz–Nevanlinna functions and laminated microstructure structure or their links with some simple circuits. Numerically, the Herglotz–Nevanlinna function theory points a way for approximating memory terms that appear very often in a dispersive system but whose descriptions are given only in the frequency domain. Also, it can provide a framework for studying the spectral theory of a TDD system.

Some very interesting results that involve yet another variation of Herglotz function can be found in the paper by Cassier et al. [24], where the Dirichlet-to-Neumann (DtN) map for the time-harmonic Maxwell's equations of a two-component composite is proved to be a Herglotz–Nevanlinna function of the variable $(\omega\mu_1, \omega\mu_2, \omega\epsilon_1, \omega\epsilon_2)$ in $(\mathbb{C}^+)^4$, which represents the electromagnetic properties of the isotropic constituent materials. To extend the result to the general case of anisotropic constituents, which can be spatially piecewise-constant or continuous, the authors define the class of Herglotz–Nevanlinna functions on an open, connected and convex set of *matrices with positive definite imaginary parts*. To preserve the Herglotz function structure, they use the *trajectory method* [14], [71, Section 18.6] to define a trajectory $s(\omega)$ that maps ω to the matrix-valued $(\omega\boldsymbol{\mu}_1, \omega\boldsymbol{\mu}_2, \omega\boldsymbol{\epsilon}_1, \omega\boldsymbol{\epsilon}_2)$ and show that the DtN map is a Herglotz–Nevanlinna function along each trajectory. The implication of this result in electric impedance tomography is yet to be discovered.

With all the applications where Herglotz–Nevanlinna functions have been successfully applied, there are still many open problems that demand further investigations. For example, the IRF for a three-phase dielectric composite has been derived in [42] by using the theory of Herglotz–Nevanlinna functions of two complex variables [55]. Also, for the purpose of separating the influence of contrasts and microstructure, a two-parameter IRF has been derived for composites of isotropic elastic materials in [75] using the results in [55]. However, in these applications, the relations between the moments and the microstructure become much more complicated. Besides, the characterization of extreme sets of measures of two variables is no longer just weak limits of sum of Dirac measures. Also, suppose a set of measurements from a causal and passive system is polluted by noise; how can one design a filter to recover the "nearest" Herglotz–Nevanlinna function that best represents the measured data? With the advance of material sciences, there are materials with negative indices and systems that emit energy; how should the Herglotz–Nevanlinna function be generalized accordingly? As is described in Sect. 3, there have been some generalization on the pure mathematics side. We believe that progress on generalizations can be sped up by collaboration and communication between mathematicians and researchers in various fields of materials sciences through the availability of a set of common mathematical languages and notations.

References

1. J. Agler, J.E. McCarthy, N.J. Young, Operator monotone functions and Löwner functions of several variables. Ann. Math. (2) **176**(3), 1783–1826 (2012)
2. J. Agler, R. Tully-Doyle, N.J. Young, Nevanlinna representations in several variables. J. Funct. Anal. **270**(8), 3000–3046 (2016)

3. N.I Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis* (Hafner Pub. Co., New York, 1965)
4. M. Avellaneda, A.J. Majda, Stieltjes integral representation and effective diffusivity bounds for turbulent transport. Phys. Rev. Lett. **62**, 753 (1989)
5. M. Avellaneda, A.J. Majda, An integral representation and bounds on the effective diffusivity in passive advection by laminar and turbulent flows. Commun. Math. Phys. **138**(2), 339–391 (1991)
6. M. Avellaneda, S. Torquato, Rigorous link between fluid permeability, electrical conductivity, and relaxation times for transport in porous media. Phys. Fluids A Fluid Dyn. **3**, 2529 (1991)
7. M. Avellaneda, M. Vergassola, Stieltjes integral representation of effective diffusivities in time-dependent flows. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top. **52**(3), 3249–3251 (1995)
8. J.A. Ball, D.S. Kaliuzhnyi-Verbovetskyi, Schur-Agler and Herglotz-Agler classes of functions: positive-kernel decompositions and transfer-function realizations. Adv. Math. **280**, 121–187 (2015)
9. A. Bensoussan, J.L. Lions, G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, 2nd edn. (American Mathematical Society, 2011)
10. D.J. Bergman, The dielectric constant of a composite material—a problem in classical physics. Phys. Rep. **43**(9), 377–407 (1978)
11. D.J. Bergman, Exactly solvable microscopic geometries and rigorous bounds for the complex dielectric constant of a two-component composite material. Phys. Rev. Lett. **44**, 1285–1287 (1980)
12. D.J. Bergman, Bounds for the complex dielectric constant of a two-component composite material. Phys. Rev. B **23**, 3058–3065 (1981)
13. D.J. Bergman, Rigorous bounds for the complex dielectric constant of a two-component composite. Ann. Phys. **138**(1), 78–114 (1982)
14. D.J. Bergman, Hierarchies of Stieltjes functions and their application to the calculation of bounds for the dielectric constant of a two-component composite medium. SIAM J. Appl. Math. **53**(4), 915–930 (1993)
15. A. Bernland, A. Luger, M. Gustafsson, Sum rules and constraints on passive systems. J. Phys. A Math. Theor. **44**(14), 145205 (2011)
16. C. Bi, M.J. Yvonne Ou, S. Zhang Integral Representation of Hydraulic Permeability. Proceedings of the Royal Society of Edinburgh: Section A Mathematics, (2022). https://doi.org/10.1017/prm.2022.25
17. M.A. Biot, Theory of propagation of elastic waves in a fluid-saturated porous solid. I. Low-frequency range. J. Acoust. Soc. Am. **28**, 168 (1956)
18. M.A. Biot, Theory of propagation of elastic waves in a fluid-saturated porous solid. II. Higher frequency range. J. Acoust. Soc. Am. **28**(2), 179–191 (1956)
19. C. Bonifasi-Lista, E. Cherkaev, Electrical impedance spectroscopy as a potential tool for recovering bone porosity. Phys. Med. Biol. **54**(10), 3063 (2009)
20. O. Bruno, K. Golden, Interchangeability and bounds on the effective conductivity of the square lattice. J. Stat. Phys. **61**(1), 365–386 (1990)
21. O.P. Bruno, P.H. Leo, On the stiffness of materials containing a disordered array of microscopic holes or hard inclusions. Arch. Rational Mech. Anal. **121**(4), 303–338 (1993)
22. J.M. Carcione, *Wave Fields in Real Media: Wave Propagation in Anisotropic, Anelastic and Porous Media* (Pergamon-Elsevier, Oxford, 2001)
23. M. Cassier, G.W. Milton, Bounds on Herglotz functions and fundamental limits of broadband passive quasistatic cloaking. J. Math. Phys. **58**(7), 071504 (2017)
24. M. Cassier, A. Welters, G.W. Milton, Analyticity of the Dirichlet-to-Neumann map for the time-harmonic Maxwell's equations. arXiv:1512.05838 [math.AP] (2015)
25. E. Charlaix, A.P. Kushnick, J.P. Stokes, Experimental study of dynamic permeability in porous media. Phys. Rev. Lett. **61**, 1595–1598 (1988)
26. E. Cherkaev, Inverse homogenization for evaluation of effective properties of a mixture. Inverse Problems **17**(4), 1203 (2001)

27. E. Cherkaeva, K.M. Golden, Inverse bounds for microstructural parameters of composite media derived from complex permittivity measurements. Waves Random Media **8**(4), 437–450 (1998)
28. E. Cherkaev, M.J.Y. Ou, Dehomogenization: reconstruction of moments of the spectral measure of the composite. Inverse Problems **24**(6), 065008 (2008)
29. G.F. Dell'Antonio, R. Figari, E. Orlandi, An approach through orthogonal projections to the study of inhomogeneous or random media with linear response. Ann. Inst. Henri Poincare (B) **44**(1), 1–28 (1986)
30. P. Delsarte, Y. Genin, Y. Kamp, Canonical factorization of pseudo-Carathéodory functions, in *Computational and Combinatorial Methods in Systems Theory (Stockholm, 1985)* (North-Holland, Amsterdam, 1986), pp. 299–307
31. P. Delsarte, Y. Genin, Y. Kamp, Pseudo-Carathéodory functions and Hermitian Toeplitz matrices. Philips J. Res. **41**(1), 1–54 (1986)
32. V. Derkach, S. Hassi, H. de Snoo, Operator models associated with Kac subclasses of generalized Nevanlinna functions. Methods Funct. Anal. Topol. **5**(1), 65–87 (1999)
33. J.W. Dettman, *Applied Complex Variables* (Dover Publications, Inc, 1965)
34. A. Dijksma, P. Kurasov, Yu. Shondin, High order singular rank one perturbations of a positive operator. Integral Equations Oper. Theory **53**(2), 209–245 (2005)
35. A. Dijksma, H. Langer, A. Luger, Yu. Shondin, A factorization result for generalized Nevanlinna functions of the class $\mathcal{N}_\kappa$. Integral Equations Oper. Theory **36**(1), 121–125 (2000)
36. A. Dijksma, H. Langer, Yu. Shondin, C. Zeinstra, Self-adjoint operators with inner singularities and Pontryagin spaces, in *Operator Theory and Related Topics, Vol. II (Odessa, 1997)*, volume 118 of Oper. Theory Adv. Appl. (Birkhäuser, Basel, 2000), pp. 105–175
37. Yu. Dyukarev, V. Katsnelson, Multiplicative and additive classes of Stieltjes analytic matrix valued functions, and interpolation problems associated with them. Am. Math. Soc. Transl. **131**, 55–70 (1986)
38. A. Figotin, J.H. Schenker, Spectral theory of time dispersive and dissipative systems. J. Stat. Phys. **118**(1), 199–263 (2005)
39. A. Figotin, J.H. Schenker, Hamiltonian treatment of time dispersive and dissipative media within the linear response theory. J. Comput. Appl. Math. **204**(2), 199–208 (2007)
40. A. Figotin, J.H. Schenker, Hamiltonian structure for dispersive and dissipative dynamical systems. J. Stat. Phys. **128**(4), 969–1056 (2007)
41. J. Gelfgren, Multipoint padé approximants used for piecewise rational interpolation and for interpolation to functions of Stieltjes' type. Technical report, Umeå universitet, 1978
42. K. Golden, Bounds on the complex permittivity of a multicomponent material. J. Mech. Phys. Solids **34**(4), 333–358 (1986)
43. K. Golden, G. Papanicolaou, Bounds for effective parameters of heterogeneous media by analytic continuation. Commun. Math. Phys. **90**(4), 473–491 (1983)
44. K.M. Golden, N.B. Murphy, E. Cherkaev, Spectral analysis and connectivity of porous microstructures in bone. J. Biomech. **44**(2), 337–344 (2011)
45. Z. Hashin, S. Shtrikman, A variational approach to the theory of the effective magnetic permeability of multiphase materials. J. Appl. Phys. **33**, 3125 (1962)
46. Z. Hashin, S. Shtrikman, A variational approach to the theory of the elastic behaviour of polycrystals. J. Mech. Phys. Solids **10**(4), 343–352 (1962)
47. Z. Hashin, S. Shtrikman, A variational approach to the theory of the elastic behaviour of multiphase materials. J. Mech. Phys. Solids **11**(2), 127–140 (1963)
48. Y. Ivanenko, M. Nedic, M. Gustafsson, B.L.G. Jonsson, A. Luger, S. Nordebo, Quasi-Herglotz functions and convex optimization. R. Soc. Open Sci. **7**, 191541 (2020)
49. V.V. Jikov, S.M. Kozlov, O.A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, 1st edn. (Springer, Berlin, 1994)
50. D.L. Johnson, J. Koplik, R. Dashen, Theory of dynamic permeability and tortuosity in fluid-saturated porous media. J. Fluid Mech. **176**(1), 379–402 (1987)
51. I.S. Kac, M.G. Krein, R-functions-analytic functions mapping the upper halfplane into itself. AMS Transl. **103**, 1–18 (1974)

52. Y. Kantor, D.J. Bergman, Improved rigorous bounds on the effective elastic moduli of a composite material. J. Mech. Phys. Solids **32**(1), 41–62 (1984)
53. S. Karlin, W.J. Studden, *Tchebycheff Systems, with Applications in Analysis and Statistics*, volume 15 of Pure and Applied Mathematics (Interscience Publishers) (Interscience Publishers, New York, 1966)
54. V. Katsnelson, Stieltjes functions and Hurwitz stable entire functions. Complex Anal. Oper. Theory **5**, 611 (2011)
55. A. Korányi, L. Pukánszky, Holomorphic functions with positive real part on polycylinders. Trans. Am. Math. Soc. **108**(3), 449–456 (1963)
56. M.G. Kreĭn, H. Langer, Über einige Fortsetzungsprobleme, die eng mit der Theorie hermitescher Operatoren im Raume Π_κ zusammenhängen. I. Einige Funktionenklassen und ihre Darstellungen. Math. Nachr. **77**, 187–236 (1977)
57. P. Kurasov, A. Luger, An operator theoretic interpretation of the generalized Titchmarsh-Weyl coefficient for a singular Sturm-Liouville problem. Math. Phys. Anal. Geom. **14**(2), 115–151 (2011)
58. R. Lipton, M. Avellaneda, Darcy's law for slow viscous flow past a stationary array of bubbles. Proc. R. Soc. Edinb. A Math. **114**(1–2), 71–79 (1990)
59. J.F. Lu, A. Hanyga, Wave field simulation for heterogeneous porous media with singular memory drag force. J. Comput. Phys. **208**(2), 651–674 (2005)
60. A. Luger, A factorization of regular generalized Nevanlinna functions. Integral Equations Oper. Theory **43**(3), 326–345 (2002)
61. A. Luger, About generalized zeros of non-regular generalized Nevanlinna functions. Integral Equations Oper. Theory **45**(4), 461–473 (2003)
62. A. Luger, M. Nedic, Herglotz-Nevanlinna functions in several variables. J. Math. Anal. Appl. **472**(1), 1189–1219 (2019)
63. A. Luger, M. Nedic, On quasi-Herglotz functions in one variable. arXiv:1909.10198 (2019)
64. A. Luger, M. Nedic, Geometric properties of measures related to holomorphic functions having positive imaginary or real part. J. Geom. Anal. **31**(3), 2611–2638 (2021)
65. Y.J. Masson, S.R. Pride, Finite-difference modeling of Biot's poroelastic equations across all frequencies. Geophysics **75**(2), N33–N41 (2010)
66. R.C. McPhedran, G.W. Milton, Bounds and exact theories for the transport properties of inhomogeneous media. Appl. Phys. A **26**(4), 207–220 (1981)
67. R.C. Mcphedran, G.W. Milton, Inverse transport problems for composite media. MRS Online Proc. Library **195**(1), 257–274 (1990)
68. G.W. Milton, Bounds on the complex dielectric constant of a composite material. Appl. Phys. Lett. **37**(3), 300–302 (1980)
69. G.W. Milton, Bounds on the transport and optical properties of a two-component composite material. J. Appl. Phys. **52**(8), 5294–5304 (1981)
70. G.W. Milton, Bounds on the complex permittivity of a two-component composite material. J. Appl. Phys. **52**(8), 5286–5293 (1981)
71. G.W. Milton, *The Theory of Composites* (Cambridge University Press, 2002). Cambridge Books Online
72. G.W. Milton, K. Golden, Representations for the conductivity functions of multicomponent composites. Commun. Pure Appl. Math. **43**(5), 647–671 (1990)
73. N.B. Murphy, E. Cherkaev, J. Zhu, J. Xin, K.M. Golden, Spectral analysis and computation for homogenization of advection diffusion processes in steady flows. J. Math. Phys. **61**(1), 013102 (2020)
74. C. Orum, E. Cherkaev, K.M. Golden, Recovery of inclusion separations in strongly heterogeneous composites from effective property measurements. Proc. R. Soc. A Math. Phys. Eng. Sci. **468**(2139), 784–809 (2012)
75. M.J.Y. Ou, Two-parameter integral representation formula for the effective elastic moduli of two-phase composites. Complex Variables Elliptic Equations **57**(2–4), 411–424 (2012)
76. M.J.Y. Ou, On reconstruction of dynamic permeability and tortuosity from data at distinct frequencies. Inverse Probl. **30**(9), 095002 (2014)

77. M.J.Y. Ou, E. Cherkaev, On the integral representation formula for a two-component composite. Math. Methods Appl. Sci. **29**(6), 655–664 (2006)
78. M.J.Y. Ou, H.J. Woerdeman, On the augmented Biot-JKD equations with pole-residue representation of the dynamic tortuosity. Oper. Theory Adv. Appl. **272**, 341–362 (2019). Springer Nature
79. S. Prager, Improved variational bounds on some bulk properties of a two-phase random medium. J. Chem. Phys. **50**(10), 4305–4312 (1969)
80. S.R. Pride, F.D. Morgan, A.F. Gangi, Drag forces of porous-medium acoustics. Phys. Rev. B **47**(9), 4964 (1993)
81. W. Rudin, *Real and Complex Analysis*, 3rd edn. (McGraw-Hill Book Co., New York, 1987)
82. E. Sánchez-Palencia, *Non-homogeneous Media and Vibration Theory*, volume 127 of Lecture Notes in Physics (Springer, 1980)
83. L. Tartar, *The General Theory of Homogenization, A Personalized Introduction*, volume 7 of Lecture Notes of the Unione Matematica Italiana, 1st edn. (Springer, Berlin, 2010)
84. O. Wiener, Die theorie des mischkörpers für das feld der stationären strömung. Abh. Sächs. Akad. Wiss. Leipzig Math.-Naturwiss. Kl. **32**, 509 (1912)
85. J. Xie, M.J.Y. Ou, L. Xu, A discontinuous Galerkin method for wave propagation in orthotropic poroelastic media with memory terms. J. Comput. Phys. **397**, 108825 (2019)
86. D. Zhang, E. Cherkaev, Reconstruction of spectral function from effective permittivity of a composite material using rational function approximations. J. Comput. Phys. **228**(15), 5390–5409 (2009)
87. V. Zhikov, G. Yosifian, Introduction to the theory of two-scale convergence. J. Math. Sci. **197**(3), 325 (2014)
88. M.Y. Zhou, P. Sheng, First-principles calculations of dynamic permeability in porous media. Phys. Rev. B **39**, 12027 (1989)

Rigidity and Flexibility in the Modelling of Shape-Memory Alloys

Angkana Rüland

1 Introduction

Shape-memory alloys are materials displaying a striking thermodynamic behaviour whose modelling gives rise to rich mathematical structures. These materials undergo a first-order, diffusionless, solid–solid phase transformation in which symmetry is reduced upon the passage from the high-temperature phase (*austenite*) to the low-temperature phase (*martensite*), see Fig. 1. This loss of symmetry gives rise to various *variants of martensite* and leads to striking phenomena such as the shape-memory effect and the development of rich and complex microstructures [7].

It is the purpose of this short review paper to survey the recent development of quantitative "wild" convex integration structures and the associated dichotomy between rigidity and flexibility in the mathematical modelling of shape-memory alloys and to formulate various related open problems.

1.1 Shape-Memory Alloys: The Phenomenological Theory, Differential Inclusions, Rigidity, and Flexibility

Following the seminal work [4], in modelling shape-memory alloys, we adopt the variational, continuum framework of the phenomenological theory and view the observed states as minimizers of an energy of the form

A. Rüland (✉)
Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany
e-mail: Angkana.Rueland@uni-heidelberg.de

M. I. Español et al. (eds.), *Research in Mathematics of Materials Science*,
Association for Women in Mathematics Series 31,
https://doi.org/10.1007/978-3-031-04496-0_21

$$\int_{\Omega} W(\nabla u, \theta)dx, \tag{1}$$

possibly with prescribed boundary conditions. Here, $\Omega \subset \mathbb{R}^3$ denotes the *reference configuration* that is chosen as the austenite at the critical temperature, $u : \Omega \to \mathbb{R}^3$ denotes the *deformation* describing the material deformation with respect to the austenite state at the critical temperature, $\theta : \Omega \to [0, \infty)$ represents *temperature*, and $W : \mathbb{R}^{3\times 3} \times [0, \infty) \to [0, \infty)$ is the *stored energy function*. Seeking to model the physical properties of shape-memory alloys, we assume that:

- The stored energy function is *frame-indifferent*, i.e., $W(QF, \theta) = W(F, \theta)$ for all $F \in \mathbb{R}^{3\times 3}$, $Q \in SO(3)$, where $SO(3)$ denotes the group of 3×3 rotation matrices, $\theta \in [0, \infty)$.
- The stored energy function respects the *material symmetries*, i.e., $W(FH, \theta) = W(F, \theta)$ for all $F \in \mathbb{R}^{3\times 3}$, $H \in \mathcal{P}$, where $\mathcal{P}$ denotes the (discrete) point group of the material, $\theta \in [0, \infty)$.

Furthermore, the minima of W reflect the properties of the phase transformation. More precisely, we assume that $W(F, \theta) = 0$, iff

$$F \in K(\theta) := \begin{cases} \alpha(\theta)SO(3)Id & \text{for } \theta > \theta_c, \\ SO(3) \cup \bigcup\limits_{j=1}^{m} SO(3)U_j(\theta_c) & \text{for } \theta = \theta_c, \\ \bigcup\limits_{j=1}^{m} SO(3)U_j(\theta) & \text{for } \theta < \theta_c. \end{cases}$$

Here $\theta_c \in (0, \infty)$ denotes the critical temperature, $\alpha : [0, \infty) \to [0, \infty)$ models the thermal expansion of the austenite, and the positive-definite, symmetric, matrix-valued functions $U_j(\theta) : [0, \infty) \to \mathbb{R}^{3\times 3}_{sym,+}$, $j \in \{1, \dots, m\}$, model the variants of martensite. They can be obtained from $U_1(\theta)$ by the action of an element of the symmetry group, i.e., for each $j \in \{2, \dots, m\}$, there exists an element $P_j \in \mathcal{P}$ such that $U_j(\theta) = P_j U_1(\theta) P_j^t$. The described properties render the variational problem (1) highly non-quasi-convex and thus lead to the absence of weak lower semi-continuity. In turn, this gives rise to finely twinned, highly complex microstructures and the notion of very weak, gradient Young measure-valued solutions [2, 4].

Due to their outlined complexity, in seeking to study minimizers of (1), it is instructive (and common [37, 42, 55]) to first consider *exactly stress-free deformations*, i.e., solutions to the associated differential inclusion at a fixed temperature $\theta \in (0, \infty)$

$$\nabla u \in K(\theta) \text{ a.e. in } \Omega \tag{2}$$

for $u \in W^{1,\infty}(\Omega, \mathbb{R}^3)$. Already here a striking dichotomy between strong *rigidity* and extreme *flexibility* arises which we discuss in the following for the model case

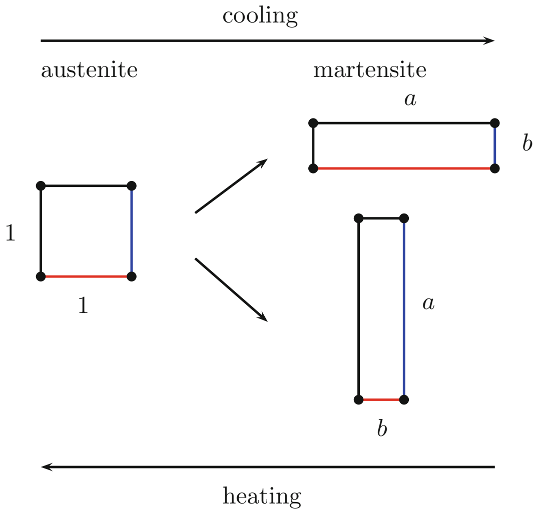


Fig. 1 Schematic illustration of the thermodynamic behaviour in shape-memory alloys for the two-dimensional square-to-rectangular transformation. Upon cooling, the typically highly symmetric high-temperature phase, austenite (in this schematic illustration with a square lattice for its unit cell), is transformed into the less symmetric low-temperature phase (martensite). The reduction of symmetry gives rise to various variants of martensite (here the two rectangular lattice structures for the low-temperature unit cell)

of the two-well problem and which microscopically can be viewed as a square-to-rectangular phase transformation (see Fig. 1). By a reduction outlined in [38], various questions on it can be analysed in the two-dimensional setting on which we will focus in the following for presentation purposes.

Studying solutions to the (two-dimensional) square-to-rectangular transformation for $\theta < \theta_c$ fixed (see Fig. 1) leads to the investigation of the following two-well problem after normalization (see [38, Section 5]): For $\Omega \subset \mathbb{R}^2$ open, bounded, smooth, classify all deformations $u : \Omega \to \mathbb{R}^2$ such that

$$\nabla u \in K_2 := SO(2)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cup SO(2)\begin{pmatrix} \mu & 0 \\ 0 & \lambda \end{pmatrix} \tag{3}$$

for some given (material-dependent) parameters $0 < \mu < 1 < \lambda$ with $\mu\lambda = 1$.

Now, on the one hand, if u is such that additionally to solving (3) it also holds that $\nabla u \in BV(\Omega, \mathbb{R}^2)$, the following rigidity result was proved in [38]:

Theorem 1 ([38, Theorem 1.1]) *Let $u : \Omega \to \mathbb{R}^2$ be a solution to (3) with $\nabla u \in BV(\Omega, \mathbb{R}^2)$. Then, locally, u is a simple laminate; more precisely, there exist two normals $\nu_1, \nu_2 \in \mathbb{S}^1 := \{v \in \mathbb{R}^2 : |v| = 1\}$ such that for some $j \in \{1, 2\}$ fixed (up to boundary effects), it holds*

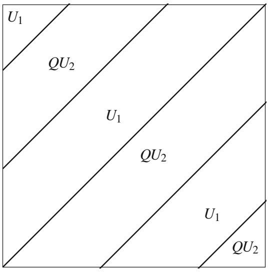


Fig. 2 Simple laminate microstructures. These structures display a one-dimensional dependence with deformation gradients that alternate along bands between two values (in this illustration between $\nabla u = U_1$ and $\nabla u = QU_2$). The orientation of the normals is determined by the rank-one connections between the wells (here between $SO(2)U_1$ and $SO(2)U_2$)

$$u(x) = h(x \cdot \nu_j) \, for \, x \in \Omega,$$

with $h : \mathbb{R} \to \mathbb{R}^2$.

In other words, under the described assumptions, any solution to (3) is *rigid* in the sense that (up to boundary effects) simple laminate solutions arise whose gradients only alternate between two possible values (up to boundary effects), with only one-dimensional directional dependences (see Fig. 2). Here, the condition $\nabla u \in BV(\Omega, \mathbb{R}^2)$ should be read as a mild, analytical form of imposing a natural "surface energy" constraint for the deformation, ruling out too irregular oscillations between the two wells.

On the other hand, using the method of convex integration [56] or the Baire category theorem [29] (see also [43] for a unification of these ideas), if the "surface energy condition" $\nabla u \in BV(\Omega, \mathbb{R}^2)$ is dropped, the rigidity from Theorem 1 is lost and a plethora of "wild" solutions exist:

Theorem 2 ([56, Corollary 1.4]) *For any* $M \in int(K_2^{lc})$*, there exist solutions* $u \in W^{1,\infty}_{loc}(\mathbb{R}^2, \mathbb{R}^2)$ *to* (3) *with the property that* $\nabla u = M$ *in* $\mathbb{R}^2 \setminus \overline{\Omega}$.

Here, K_2^{lc} denotes the *lamination convex hull* of the set K_2 that is obtained by convexification along rank-one directions. In general, the lamination convex hull of a set is substantially smaller than the convex hull (see, for instance, [29, 55] for a precise definition of this notion and a hierarchy of notions of convexification in the context of the calculus of variations). In our context, the specific set K_2^{lc} is known with an explicit characterization (see [55, Theorem 4.12]). In particular, since the boundary conditions from the explicitly known set $\text{int}(K_2^{lc})$ [62] are in general *not* compatible with simple laminates, the solutions from Theorem 2 do *not* coincide with those from Theorem 1 and thus do *not* obey the surface energy constraints from above. They are expected to be quite fractal and thus of rather low regularity. Moreover, they are highly non-unique: With respect to the L^∞ norm, it is possible to approximate up to any desired precision any affine function with gradient in $\text{int}(K_2^{lc})$

by a solution to (3) in Ω with matching boundary conditions. Hence, there exist a plethora of solutions as in Theorem 2. In what follows, we will refer to these possibly quite irregular and highly non-unique solutions as “wild” solutions.

We remark that the two-dimensional two-well problem (3) is to be considered as a model case for the dichotomy between rigidity and flexibility, but that this dichotomy also arises in various other settings in the modelling of shape-memory alloys [24, 42, 44, 63]. Similar phenomena are also observed in other problems in the calculus of variations and PDEs, ranging from elliptic and parabolic systems [57, 59, 72], origami constructions [21, 32], isometric embeddings [25], and to fluid mechanics [9, 33, 73].

1.2 Main Questions

The presence of the outlined dichotomy thus gives rise to various natural questions both for the general setting involving *m* rotation invariant wells as in (2) and for the specific example from (3):

(Q1) *Physical role of wild solutions*: While simple laminates as in Theorem 1 are indeed observed experimentally, the physical role of the wild, non-unique solutions is less evident. Do they have a physical significance? Or are they “only” mathematical constructions?

(Q2) *Thresholds*: Are there sharp (regularity) thresholds between the rigid and flexible regimes (see Fig. 3)? Can these be identified?

(Q3) *Selection mechanisms*: Among the solutions from Theorem 2 in which there is a high degree of non-uniqueness, is there a (physical or mathematical) selection mechanism for certain classes of “appropriate solutions”?

Seeking to make progress towards understanding these long-term goals, in the past years, the study of both the rigidity and flexibility properties of the associated differential inclusions was initiated from a *quantitative* point of view. Here the

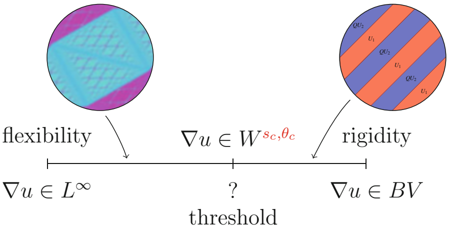


Fig. 3 The dichotomy between rigidity and flexibility. Theorems 1 and 2 correspond to the borderline cases. Natural questions thus deal with the intermediate region on Sobolev and Besov scales. In particular, a key challenge consists in identifying and proving the presence of regularity thresholds separating these two regimes

following, more modest questions were considered, which we will review in the subsequent sections:

(Q4) *Persistence of flexibility at positive Sobolev regularity*: Does the flexibility from Theorem 2 persist at higher Sobolev regularities?
(Q5) *Bounds on the maximal regularity of wild solutions by scaling*: Are there natural limitations, e.g. in terms of scaling, to the possible regularity of "wild solutions"?
(Q6) *Numerical implementations*: What type of structures arise in these constructive higher regularity schemes? Can numerical implementations of these be compared to experimental structures? Can they be related to intermittent nucleation behaviour?

These questions are also to be viewed in the context of other rigidity and flexibility results in related systems (see [8, 46, 57, 73]), where similar questions are studied on the corresponding, problem-specific regularity scales (e.g., in Hölder spaces for Onsager's conjecture in fluids). Shape-memory alloys should thus be regarded as a particular instance of this phenomenon whose understanding would lead to a further facet of the relevant mechanisms at play, complementing the already known observations, tools, and results from fluid mechanics and geometry and thus potentially contributing to inferring a broader perspective on these phenomena.

2 Flexibility

Shape-memory alloys are physical systems for which first important mathematical results on the *qualitative* dichotomy between rigidity and flexibility were already derived in the 1990s with fundamental contributions by Dacorogna and Marcellini [30], Dolzmann [38], Kirchheim [43], Müller and Šverák [57], and Sychev [58], building on their fundamental contributions due to Gromov [40]. This study also partially motivated the investigation of the m-well problem in which the rotation invariant differential inclusion from (2) is replaced by the inclusion

$$\nabla u \in \{A_1, \dots, A_m\} \text{ in } \Omega \subset \mathbb{R}^n \tag{4}$$

with $A_j \in \mathbb{R}^{n\times n}$ being pairwise not rank-one connected [3, 70] and which was fully resolved by Chlebík–Kirchheim [18] and Kirchheim–Preis [43]. In view of the dichotomy established in Theorems 1 and 2 and similar progress in other continuum mechanical settings—most prominently in fluids—the *quantitative* question $(Q2)$ on the existence of *thresholds* between the rigid and flexible settings arises naturally. This in turn consists of two directions: The study of the *persistence of flexibility* at higher (Sobolev) regularity, formulated in question $(Q3)$, and the *identification of obstructions* to the existence of wild solutions as, for instance, formulated in question $(Q4)$. In the context of shape-memory alloys, Sobolev or Besov scales pose natural classes in which to consider the problems. Indeed, since (sharp) interfaces between different deformation gradients are expected (see, for instance, Fig. 2) and are in fact observed in these materials [7], other common function space scales such as Hölder spaces, for instance, are ruled out.

In what follows, we discuss these two directions individually, focussing on the persistence of flexibility at higher Sobolev regularity in this section and presenting first results on obstructions and possible means of identifying thresholds in the next section.

Returning to the model case of the geometrically nonlinear, two-dimensional square-to-rectangular phase transformation, in recent work the following persistence of flexibility result could be proved:

Theorem 3 ([34, Theorem 1.1]) *Let Ω be an open, connected domain that can be covered by finitely many, up to null-set disjoint triangles. Then there is $\theta_0 > 0$ such that for all $s \in (0,1)$, $p \in (1,\infty)$, and $sp < \theta_0$ and for any $M \in \operatorname{int}(K_2^{lc})$, there exists a deformation $u \in W^{1,\infty}_{loc}(\mathbb{R}^2, \mathbb{R}^2)$ such that ∇u obeys (3) and $\nabla u = M$ in $\mathbb{R}^2 \setminus \overline{\Omega}$, and $\nabla u \in W^{s,p}_{loc}(\mathbb{R}^2, \mathbb{R}^2)$.*

In particular, the flexibility from Theorem 2 persists also at positive Sobolev regularity. We remark that in [34] the result is formulated with a slightly different set K_2. This, however, can be transformed into the one from above by affine transformations and thus implies that the results from [34] hold in the framework introduced above.

Let us comment on this result: Theorem 3 provides a first example in the literature on shape-memory alloys of a *geometrically nonlinear* phase transformation in which the persistence of higher regularity on Sobolev scales is proved. Moreover, the given Sobolev regularity threshold is *uniform* in the boundary data. Similar regularity results are also proved for the phase indicators for these solutions, i.e., for the characteristic functions that at each point assign the closest well to the solution. As a consequence, certain bounds on the box dimension of the interfaces separating the wells can be obtained [34, Section 7]. Key ingredients in the proof of Theorem 3 are:

(i) A (explicit) quantitative convex integration scheme, i.e., an iterative (gradient) improvement scheme that simultaneously quantitatively controls the improvement in the distance to the wells measured in the L^1 norm and the growth of the BV norm. In this way, a *subsolution* (see [73] for an outline on how this fits into a more general framework à la Tartar) whose gradient attains values in the substantially larger set K_2^{lc} is gradually and *quantitatively* transformed into a *solution* of the differential equation. The competition between fast convergence in L^1 and the divergence of the BV norm then, by interpolation, yields the desired Sobolev scale up to which flexibility of the given construction persists. This construction makes use of certain building blocks that were introduced in [20], see also [22], which allow to replace a given constant gradient by a successively more favourable, finite gradient distribution.

(ii) A good understanding of the lamination convex hull of the two-well problem. In contrast to earlier results on the quantitative persistence of flexibility from

[67, 68], Theorem 3 is the first example of a phase transformation in which the relevant lamination convex hull does *not* coincide with the (usual) convex hull. In particular, carefully chosen coordinates have to be considered for which the two available rank-one connections are exploited and which is inspired by the seminal work [4].

We remark that the interpolation-based scheme from (i) had already been developed in the context of other quantitative persistence of flexibility results for the modelling of shape-memory alloys: In [68], based on [61], a first scheme in which part of the deformation was "directly pushed into the wells" was analysed quantitatively for the two-dimensional, geometrically linearized hexagonal-to-rhombic phase transformation. In this scheme, the gradient distribution of a given subsolution was transformed in each iteration step in such a way that always at least on a certain volume fraction of the domain the modified gradient distribution already attained values within the wells by using Conti-type constructions as in [20]. On this part of the domain, the gradient was not modified any further; on the complement, the described procedure was iterated. Strongly using the geometrically linearized character of the specific hexagonal-to-rhombic phase transformation and the two-dimensional, infinitesimally volume-preserving structure of the wells, this scheme could however not directly be applied to more general transformations. To this end, in [67], a further, more general scheme was developed with a geometric improvement of the gradient distribution towards the wells but without pushing directly into the wells. The latter, in particular, provided a framework for constructions within both the geometrically linear and the nonlinear theories of elasticity and also allowed to treat physically relevant, three-dimensional problems such as the geometrically linearized cubic-to-orthorhombic phase transformation. An application to the geometrically nonlinear theory however required the verification of certain geometric properties for the relevant hulls, which was achieved in [34] by a careful study of item (ii) outlined above. A major obstruction to applying the outlined quantitative ideas to other geometrically nonlinear settings such as the geometrically nonlinear, three-dimensional cubic-to-tetragonal phase transformation is thus exactly an analogue of the analysis from (ii) since the quantitative construction requires a good understanding of the corresponding lamination convex hulls that are not known in many physically interesting cases.

A major question that remains open at present is the problem of finding *optimal* constructions. While for more flexible systems such as origamis explicit "threshold constructions" are known in two dimensions [31], in the context of shape-memory alloys this is only the case in very particular, highly symmetric settings with very specific boundary conditions. For these, it is possible to prove that (in specific geometries) flexibility persists up to BV regularity. Some of the most notable constructions with this property are two-dimensional nucleation mechanisms such as the "star-type" deformations in the hexagonal-to-rhombic phase transformation (see [45, 53, 54] for experimental observations and [14, 68] for the geometrically linearized constructions). In [15, 27], these nucleation constructions were extended to a related family of highly symmetric, two-dimensional, geometrically nonlinear transformations with austenite boundary conditions, connecting shape-memory alloys and nematic elastomers. We also refer to [36] for a first article treating these physical systems simultaneously. For generic boundary conditions and phase transformations, and in particular in three dimensions, however this question remains open and a key challenge:

(Q7) *Optimal constructions.* Determine optimal constructions for flexibility. Are these only possible under very high symmetry conditions?

3 Rigidity

Pushing towards a threshold between rigidity and flexibility, a major question concerns the presence of *obstructions* to flexibility. In fluid systems, this is given by conserved quantities (e.g., the kinetic energy at sufficient regularity); in isometric immersions, this role is taken by weak notions of curvature [73]. In both settings, scaling plays a crucial role in proving rigidity and in obtaining candidates of possible thresholds, see, for example, [19]. Contrary to equations from fluid mechanics, the differential inclusions (2) do not directly provide thresholds since they are scaling invariant with respect to the $W^{1,\infty}$ rescaling; they do not select a particular length scale.

In order to nevertheless introduce scaling as a mechanism to detect a possible threshold behaviour, in [66] we adopted a singular perturbation point of view and added an "artificial viscosity" in the form of surface energy penalization:

$$E_\epsilon(u) := \int\limits_{\Omega} \operatorname{dist}^2(\nabla u, K)dx + \epsilon^2 \int\limits_{\Omega} |\nabla^2 u|^2 dx. \tag{5}$$

Here $K \subset \mathbb{R}^{n\times n}_{sym,+}$ denotes a set of the type $K(\theta)$ from above for some fixed $\theta \in (0,\infty)$. Hence, the first contribution in (5) denotes a piecewise quadratic energy as in (1), and the second contribution penalizes fast and rapid oscillations between (and within) the different wells. It can be interpreted as a form of a surface energy (or a "viscous" contribution). Such a penalization point of view represents a common approach in analysing and predicting length scales for microstructures in shape-memory alloys (and other phase transforming problems modelled within the calculus of variations); we refer to [6, 10, 11, 16, 23, 26, 28, 47–52, 55, 64] and the references therein. We further remark that the particular choice of the surface energy (diffuse, Ginzburg–Landau energies as in (5) versus, for instance, sharp interface energies) does not provide major conceptual differences (although, of course, requiring technical adaptations in many proofs).

Seeking to employ (5) in the investigation of (2) and to detect regularity thresholds for (2), in particular the scaling behaviour of the minimal energy

$$E_\epsilon(M) := \inf_{\nabla u = M \text{ in } \mathbb{R}^n \setminus \overline{\Omega}} E_\epsilon(u)$$

becomes relevant. Heuristically, the scaling behaviour of $E_\epsilon(M)$ in terms of $\epsilon > 0$ as $\epsilon \to 0$ is expected to provide upper bounds on the possible regularity of wild structures. Indeed, consider a given convex integration solution with ∇u of a certain H^s_{loc} regularity for some $s \in (0, 1)$ and boundary data M. Then, while, due to its potentially too low regularity, u itself may not be an admissible candidate in the minimization problem for $E_\epsilon(M)$, a regularized version u_δ indeed yields a competitor for (5). Now, the penalization term roughly behaves as

$$\|\nabla^2 u_\delta\|^2_{L^2(\Omega)} \sim \delta^{2s-2}$$

in the regularization parameter δ. The elastic energy on the other hand converges towards a solution of the differential inclusion and thus

$$\int_\Omega \operatorname{dist}^2(\nabla u_\delta, K)dx \lesssim \|\nabla u_\delta - \nabla u\|^2_{L^2(\Omega)} \lesssim \delta^{2s}.$$

Hence, optimizing (5) in δ, one obtains that for these solutions the energy is expected to behave as ϵ^{2s}. This was made rigorous in [66]:

Theorem 4 ([66, Theorem 1]) *Let $M \in \mathbb{R}^{n\times n}_{sym,+}$, and let $E_\epsilon(M)$ be as above. Assume that there exists a solution $u \in H^{1+s}_{loc}(\mathbb{R}^n, \mathbb{R}^n)$ to*

$$\nabla u \in K \text{ in } \Omega,$$
$$\nabla u = M \text{ in } \mathbb{R}^n \setminus \overline{\Omega}.$$

Then there exists a constant $C = C(\Omega, M, n, s) > 0$ such that for $\epsilon > 0$ small enough

$$E_\epsilon(M) \leq C\epsilon^{2s}.$$

As a consequence, if for given boundary data $M \in \mathbb{R}^{n\times n}_{sym,+}$ there are wild solutions whose gradients are of $H^s_{loc}(\mathbb{R}^n, \mathbb{R}^n)$ regularity, this yields an *upper* bound on $E_\epsilon(M)$. Conversely, if *lower* bounds in terms of the ϵ scaling for $E_\epsilon(M)$ are known, then this yields an *upper* bound on the *maximal regularity* of wild solutions. It is this latter interpretation that we view as a possible means of detecting and identifying thresholds for the regularity of wild solutions that replaces the usual scaling arguments for other systems. We remark that the specific forms of the surface energies here do not play a major role; in [66], it is proved that analogous results also hold for other measures of surface energy penalizations. However, to the best of the author's knowledge, none of the currently existing lower scaling bounds for models on shape-memory alloys hold in a setting in which there is a dichotomy between rigidity and flexibility.

Only in very recent work [65, Theorem 1], a rather unusual, slower than any power scaling bound was obtained for a singular perturbation of the Tartar square. The Tartar square, also denoted as T_4, is a well-known set in matrix space that was discovered in different contexts (see [1, 12, 60, 69, 74]). It falls into the framework of (4) with $m = 4$, i.e., $T_4 = \{A_1, \dots, A_4\}$ with four specific, pairwise not rank-one connected matrices $A_1, \dots, A_4 \in \mathbb{R}^{2\times 2}_{sym}$. For the Tartar square, a dichotomy between rigidity of exact solutions and approximate solutions, which is related to the dichotomy outlined in Sect. 1.1, is known to exist (see, for instance, the survey on this in [55]): While exact solutions to $\nabla u \in T_4$ are rigid in the sense that they must already have constant gradients, approximate solutions, i.e., sequences for which $\text{dist}(\nabla u_k, T_4) \to 0$ in measure, need not be rigid in the sense that there exist sequences of approximate solutions $(\nabla u_k)_k$ such that no subsequences converge to one of the constant gradient deformations $\nabla u \in T_4$. In [56], it was even proved that any open perturbation of the Tartar square permits wild convex integration solutions. In addition to these aspects directly related to the Tartar square, the Tartar square also plays a major, auxiliary role in convex integration constructions [33, 57, 69, 72], the analysis of Morrey's conjecture, and the calculus of variations in general [39, 44, 71] and in the study of compensated compactness [74]. Furthermore, siblings of it, such as certain geometrically linearized T_3 structures, were discovered in the context of the physically relevant cubic-to-monoclinic phase transformation [17]. These consist of three symmetric, pairwise not symmetrized rank-one connected 3×3 matrices and enjoy similar properties as the Tartar square. While the scaling result from [65] thus does not directly fit into the exact framework from [66], it could provide an important first step into such a direction.

In general, however, the scaling for more general sets in matrix space displaying a dichotomy between rigidity and flexibility remains an outstanding problem at the moment.

(Q8) *Lower scaling bounds.* Determine lower scaling bounds for models for shape-memory alloys displaying the dichotomy between rigidity and flexibility.

4 Simulations

In order to eventually make progress towards the physically most relevant question $(Q1)$ from Sect. 1.2 and in order to compare the mathematically constructed solutions with experiments, in [66] also numerical implementations of the "wild" convex integration solutions were provided and discussed. These strongly display fractal behaviour with (up to numerically evaluable scales) power law length distributions, see Fig. 4 and the discussion in [66, Section 3].

Also various experimental systems display very complex structures upon nucleation. Here observations as in [41] are promising systems of comparison. Moreover,

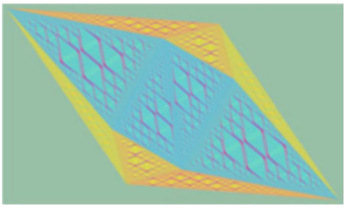


Fig. 4 A simulation of a (finite iteration of a) wild microstructure for the geometrically linearized, two-dimensional hexagonal-to-rhombic phase transformation (within the diamond-shaped domain) based on the analysis and numerics from [66]. This phase transformation has three variants of martensite that here are colour-coded as yellow, magenta, and cyan. One clearly observes fractal behaviour. The greenish colour at the exterior of the diamond corresponds to the chosen boundary datum that lies in the interior of the lamination convex hull and thus in none of the wells. Colour versions of the figure are available online

in the context of (self-organized) nucleation dynamics that display strongly intermittent behaviour, convex integration could potentially play a physically relevant role (see [5, 13, 35, 75] for models on this). The latter models may further be of intrinsic mathematical interest in understanding "probabilistic (average) variants" of convex integration algorithms. However, in spite of these first promising initial experimental and theoretical results, major further analysis and experiments seem necessary.

Acknowledgments A.R. gratefully acknowledges that part of the research was funded through SPP 2256 of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project ID 441068247. She is also supported by the Heidelberg STRUCTURES Excellence Cluster, which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948.

References

1. R.J. Aumann, S. Hart, Bi-convexity and bi-martingales. Israel J. Math. **54**(2), 159–180 (1986)
2. J.M. Ball, A version of the fundamental theorem for young measures, in *PDEs and Continuum Models of Phase Transitions* (Springer, 1989), pp. 207–215
3. J.M. Ball, R.D. James, Fine phase mixtures as minimizers of energy, in *Analysis and Continuum Mechanics* (Springer, 1989), pp. 647–686
4. J.M. Ball, R.D. James, Proposed experimental tests of a theory of fine microstructure and the two-well problem. Philos. Trans. R. Soc. Lond. A **338**(1650), 389–450 (1992)
5. J.M. Ball, P. Cesana, B. Hambly, A probabilistic model for martensitic avalanches, in *MATEC Web of Conferences*, vol. 33 (EDP Sciences, 2015), p. 02008
6. P. Bella, M. Goldman, Nucleation barriers at corners for a cubic-to-tetragonal phase transformation. Proc. R. Soc. Edinb. A Math. **145**(4), 715–724 (2015)
7. K. Bhattacharya, *Microstructure of Martensite: Why It Forms and How It Gives Rise to the Shape-memory Effect Oxford Series on Materials Modeling* (Oxford University Press, 2003)
8. T. Buckmaster, V. Vicol, Nonuniqueness of weak solutions to the Navier-Stokes equation. Ann. Math. **189**(1), 101–144 (2019)

9. T. Buckmaster, V. Vicol, Convex integration and phenomenologies in turbulence. EMS Surv. Math. Sci. **6**(1), 173–263 (2020)
10. A. Capella, F. Otto, A rigidity result for a perturbation of the geometrically linear three-well problem. Commun. Pure Appl. Math. **62**(12), 1632–1669 (2009)
11. A. Capella, F. Otto, A quantitative rigidity result for the cubic-to-tetragonal phase transition in the geometrically linear theory with interfacial energy. Proc. R. Soc. Edinb. A Math. **142**, 273–327 (2012). https://doi.org/10.1017/S0308210510000478
12. E. Casadio-Tarabusi, An algebraic characterization of quasi-convex functions. Ricerche Mat. **42**(1), 11–24 (1993)
13. P. Cesana, B. Hambly, A probabilistic model for interfaces in a martensitic phase transition. arXiv preprint arXiv:1810.04380 (2018)
14. P. Cesana, M. Porta, T. Lookman, Asymptotic analysis of hierarchical martensitic microstructure. J. Mech. Phys. Solids **72**, 174–192 (2014)
15. P. Cesana, F. Della Porta, A. Rüland, C. Zillinger, B. Zwicknagl, Exact constructions in the (non-linear) planar theory of elasticity: From elastic crystals to nematic elastomers. Arch. Rational Mech. Anal. **237**(1), 383–445 (2020)
16. A. Chan, S. Conti, Energy scaling and domain branching in solid-solid phase transitions, in *Singular Phenomena and Scaling in Mathematical Models* (Springer, 2014), pp. 243–260
17. I.V. Chenchiah, A. Schlömerkemper, Non-laminate microstructures in monoclinic-I martensite. Arch. Rational Mech. Anal. **207**(1), 39–74 (2013)
18. M. Chlebík, B. Kirchheim, Rigidity for the four gradient problem. J. Reine Angew. Math. **551**, 1–9 (2002)
19. P. Constantin, E.S. Titi, F. Weinan, Onsager's conjecture on the energy conservation for solutions of Euler's equation. Commun. Math. Phys. **165**(1), 207 (1994)
20. S. Conti, Quasiconvex functions incorporating volumetric constraints are rank-one convex. J. Math. Pures Appl. **90**(1), 15–30 (2008)
21. S. Conti, F. Maggi, Confining thin elastic sheets and folding paper. Arch. Rational Mech. Anal. **187**(1), 1–48 (2008)
22. S. Conti, F. Theil, Single-slip elastoplastic microstructures. Arch. Rational Mech. Anal. **178**(1), 125–148 (2005)
23. S. Conti, B. Zwicknagl, Low volume-fraction microstructures in martensites and crystal plasticity. Math. Models Methods Appl. Sci. **26**(07), 1319–1355 (2016)
24. S. Conti, G. Dolzmann, B. Kirchheim, Existence of Lipschitz minimizers for the three-well problem in solid-solid phase transitions. Ann. Inst. Henri Poincare (C) Non Linear Anal. **24**(6), 953–962 (2007)
25. S. Conti, C. De Lellis, L. Székelyhidi, h-principle and rigidity for $C^{1,\alpha}$ isometric embeddings, in *Nonlinear Partial Differential Equations* (Springer, 2012), pp. 83–116
26. S. Conti, J. Diermeier, B. Zwicknagl, Deformation concentration for martensitic microstructures in the limit of low volume fraction. Calc. Variations Partial Differential Equations **56**(1), 16 (2017)
27. S. Conti, M. Klar, B. Zwicknagl, Piecewise affine stress-free martensitic inclusions in planar nonlinear elasticity. Proc. R. Soc. A Math. Phys. Eng. Sci. **473**(2203), 20170235 (2017)
28. S. Conti, J. Diermeier, D. Melching, B. Zwicknagl, Energy scaling laws for geometrically linear elasticity models for microstructures in shape memory alloys. ESAIM Control Optim. Calc. Variations **26**, 115 (2020)
29. B. Dacorogna, *Direct Methods in the Calculus of Variations*, vol. 78 (Springer, 2007)
30. B. Dacorogna, P. Marcellini, *Implicit Partial Differential Equations*, vol. 37 (Springer Science & Business Media, 2012)
31. B. Dacorogna, P. Marcellini, E. Paolini, Lipschitz-continuous local isometric immersions: rigid maps and origami. J. Math. Pures Appl. **90**(1), 66–81 (2008)
32. B. Dacorogna, P. Marcellini, E. Paolini, Origami and partial differential equations. Not. AMS **57**(5), 598–606 (2010)
33. C. De Lellis, L. Székelyhidi Jr, The Euler equations as a differential inclusion. Ann. Math., **170**(3), 1417–1436 (2009)

34. F. Della Porta, A. Rüland, Convex integration solutions for the geometrically nonlinear two-well problem with higher Sobolev regularity. Math. Models Methods Appl. Sci. **30**(03), 611–651 (2020)
35. F. Della Porta, A. Rüland, J.M. Taylor, C. Zillinger, On a probabilistic model for martensitic avalanches incorporating mechanical compatibility. Nonlinearity **34**(7), 4844–4896 (2021)
36. A. DeSimone, Energetics of fine domain structures. Ferroelectrics **222**(1), 275–284 (1999)
37. G. Dolzmann, S. Müller, The influence of surface energy on stress-free microstructures in shape memory alloys. Meccanica **30**, 527–539 (1995). https://doi.org/10.1007/BF01557083
38. G. Dolzmann, S. Müller, Microstructures with finite surface energy: the two-well problem. Arch. Rational Mech. Anal. **132**, 101–141 (1995)
39. D. Faraco, L. Székelyhidi, Tartar's conjecture and localization of the quasiconvex hull in $\mathbb{R}^{2\times 2}$. Acta Math. **200**(2), 279–305 (2008)
40. M.L. Gromov, Convex integration of differential relations. I. Izvestiya Math. **7**(2), 329–343 (1973)
41. T. Inamura, Personal communication, manuscript in preparation.
42. B. Kirchheim, Lipschitz minimizers of the 3-well problem having gradients of bounded variation. MPI preprint (1998)
43. B. Kirchheim, Rigidity and geometry of microstructures, in *MPI-MIS Lecture Notes* (2003)
44. B. Kirchheim, S. Müller, V. Šverák, Studying nonlinear PDE by geometry in matrix space, in *Geometric Analysis and Nonlinear Partial Differential Equations* (Springer, 2003), pp. 347–395
45. Y. Kitano, K. Kifune, HREM study of disclinations in MgCd ordered alloy. Ultramicroscopy **39**(1–4), 279–286 (1991)
46. S. Klainerman, On Nash's unique contribution to analysis in just three of his papers. Bull. Am. Math. Soc. **54**(2), 283–305 (2017)
47. H. Knüpfer, R.V. Kohn, Minimal energy for elastic inclusions. Proc. R. Soc. A Math. Phys. Eng. Sci. **467**(2127), 695–717 (2011)
48. H. Knüpfer, F. Otto, Nucleation barriers for the cubic-to-tetragonal phase transformation in the absence of self-accommodation. ZAMM J. Appl. Math. Mech. [Z. Angew. Math. Mech.] **99**(2), e201800179 (2019)
49. H. Knüpfer, R.V. Kohn, F. Otto, Nucleation barriers for the cubic-to-tetragonal phase transformation. Commun. Pure Appl. Math. **66**(6), 867–904 (2013)
50. R.V. Kohn, Energy-driven pattern formation, in *International Congress of Mathematicians*, vol. 1 (European Mathematical Society, Zürich, 2007), pp. 359–383
51. R.V. Kohn, S. Müller, Branching of twins near an austenite—twinned-martensite interface. Philos. Mag. A **66**(5), 697–715 (1992)
52. R.V. Kohn, S. Müller, Surface energy and microstructure in coherent phase transitions. Commun. Pure Appl. Math. **47**(4), 405–435 (1994)
53. C. Manolikas, S. Amelinckx, Phase transitions in ferroelastic lead orthovanadate as observed by means of electron microscopy and electron diffraction. I. Static observations. Physica Status Solidi (A) **60**(2), 607–617 (1980)
54. C. Manolikas, S. Amelinckx, Phase transitions in ferroelastic lead orthovanadate as observed by means of electron microscopy and electron diffraction. II. Dynamic Observations. Physica Status Solidi (A) **61**(1), 179–188 (1980)
55. S. Müller, Variational models for microstructure and phase transitions, in *Calculus of Variations and Geometric Evolution Problems* (Springer, 1999), pp. 85–210
56. S. Müller, V. Šverák, Convex integration with constraints and applications to phase transitions and partial differential equations. J. Eur. Math. Soc. **1**, 393–422 (1999). https://doi.org/10.1007/s100970050012
57. S. Müller, V. Šverák, Convex integration for Lipschitz mappings and counterexamples to regularity. Ann. Math. **157**(3), 715–742 (2003)
58. S. Müller, M.A. Sychev, Optimal existence theorems for nonhomogeneous differential inclusions. J. Funct. Anal. **181**(2), 447–475 (2001)

59. S. Müller, M.O. Rieger, V. Šverák, Parabolic systems with nowhere smooth solutions. Arch. Rational Mech. Anal. **177**(1), 1–20 (2005)
60. V. Nesi, G.W. Milton, Polycrystalline configurations that maximize electrical resistivity. J. Mech. Phys. Solids **39**(4), 525–542 (1991)
61. F. Otto, Pattern formation and scaling laws in materials science. https://www.ima.umn.edu/2011-2012/SW6.21-29.12/12380. Lecture at the NSF PIRE Summer School for Graduate Students: New frontiers in multiscale analysis and computing for materials, Minneapolis (2012)
62. P. Pedregal, *Parametrized Measures and Variational Principles*, vol. 30 (Birkhauser, Basel, 1997)
63. A. Rüland, The cubic-to-orthorhombic phase transition: Rigidity and non-rigidity properties in the linear theory of elasticity. Arch. Rational Mech. Anal. **221**(1), 23–106 (2016)
64. A. Rüland, A rigidity result for a reduced model of a cubic-to-orthorhombic phase transition in the geometrically linear theory of elasticity. J. Elasticity **123**(2), 137–177 (2016)
65. A. Rüland, A. Tribuzio, On the energy scaling behaviour of a singularly perturbed Tartar square. Arch. Ration. Mech. Anal. **243**(1), 401-431 (2022)
66. A. Rüland, J.M. Taylor, C. Zillinger, Convex integration arising in the modelling of shape-memory alloys: some remarks on rigidity, flexibility and some numerical implementations. J. Nonlinear Sci., **29**(5), 2137–2184 (2019)
67. A. Rüland, C. Zillinger, B. Zwicknagl, Higher Sobolev regularity of convex integration solutions in elasticity: The Dirichlet problem with affine data in int(K^{lc}). SIAM J. Math. Anal. **50**(4), 3791–3841 (2018)
68. A. Rüland, C. Zillinger, B. Zwicknagl, Higher Sobolev regularity of convex integration solutions in elasticity: The planar geometrically linearized hexagonal-to-rhombic phase transformation. J. Elasticity (2019). https://doi.org/10.1007/s10659-018-09719-3
69. V. Scheffer, Regularity and irregularity of solutions to nonlinear second-order elliptic systems of partial differential-equations and inequalities. Thesis (Ph.D.)-Princeton University (1974), 116 pp.
70. V. Šverák, New examples of quasiconvex functions. Arch. Rational Mech. Anal. **119**(4), 293–300 (1992)
71. V. Šverák, On Tartar's conjecture, in *Annales de l'IHP Analyse non linéaire*, vol. 10 (1993), pp. 405–412
72. L. Székelyhidi Jr, The regularity of critical points of polyconvex functionals. Arch. Rational Mech. Anal. **172**(1), 133–152 (2004)
73. L. Székelyhidi Jr, From isometric embeddings to turbulence, in *HCDTE Lecture Notes. Part II. Nonlinear Hyperbolic PDEs, Dispersive and Transport Equations*, vol. 7, 63 (2012)
74. L. Tartar, Some remarks on separately convex functions, in *Microstructure and Phase Transition* (Springer, 1993), pp. 191–204
75. G. Torrents, X. Illa, E. Vives, A. Planes, Geometrical model for martensitic phase transitions: Understanding criticality and weak universality during microstructure growth. Phys. Rev. E **95**(1), 013001 (2017)