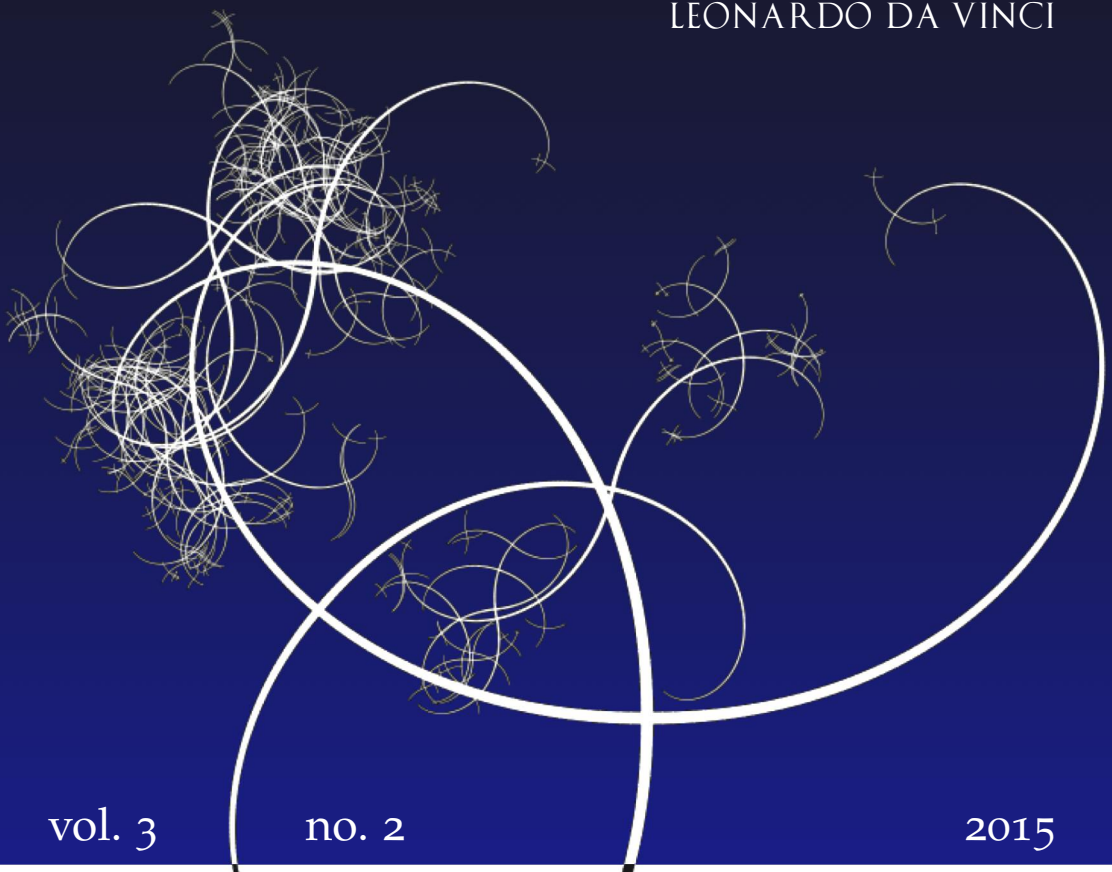
NISSUNA UMANA INVESTIGAZIONE SI PUO DIMANDARE
VERA SCIENZIA S'ESSA NON PASSA PER LE
MATEMATICHE DIMOSTRAZIONI
LEONARDO DA VINCI



vol. 3 no. 2 2015

Mathematics and Mechanics *of* Complex Systems

msp

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS

msp.org/memocs

EDITORIAL BOARD

ANTONIO CARCATERRA	Università di Roma "La Sapienza", Italia
ERIC A. CARLEN	Rutgers University, USA
FRANCESCO DELL'ISOLA	(CO-CHAIR) Università di Roma "La Sapienza", Italia
RAFFAELE ESPOSITO	(TREASURER) Università dell'Aquila, Italia
ALBERT FANNJIANG	University of California at Davis, USA
GILLES A. FRANCFORT	(CO-CHAIR) Université Paris-Nord, France
PIERANGELO MARCATI	Università dell'Aquila, Italy
JEAN-JACQUES MARIGO	École Polytechnique, France
PETER A. MARKOWICH	DAMTP Cambridge, UK, and University of Vienna, Austria
MARTIN OSTOJA-STARZEWSKI	(CHAIR MANAGING EDITOR) Univ. of Illinois at Urbana-Champaign, USA
PIERRE SEPPECHER	Université du Sud Toulon-Var, France
DAVID J. STEIGMANN	University of California at Berkeley, USA
PAUL STEINMANN	Universität Erlangen-Nürnberg, Germany
PIERRE M. SUQUET	LMA CNRS Marseille, France

MANAGING EDITORS

MICOL AMAR	Università di Roma "La Sapienza", Italia
CORRADO LATTANZIO	Università dell'Aquila, Italy
ANGELA MADEO	Université de Lyon–INSA (Institut National des Sciences Appliquées), France
MARTIN OSTOJA-STARZEWSKI	(CHAIR MANAGING EDITOR) Univ. of Illinois at Urbana-Champaign, USA

ADVISORY BOARD

ADNAN AKAY	Carnegie Mellon University, USA, and Bilkent University, Turkey
HOLM ALTENBACH	Otto-von-Guericke-Universität Magdeburg, Germany
MICOL AMAR	Università di Roma "La Sapienza", Italia
HARM ASKES	University of Sheffield, UK
TEODOR ATANACKOVIĆ	University of Novi Sad, Serbia
VICTOR BERDICHEVSKY	Wayne State University, USA
GUY BOUCHITTÉ	Université du Sud Toulon-Var, France
ANDREA BRAIDES	Università di Roma Tor Vergata, Italia
ROBERTO CAMASSA	University of North Carolina at Chapel Hill, USA
MAURO CARFORE	Università di Pavia, Italia
ERIC DARVE	Stanford University, USA
FELIX DARVE	Institut Polytechnique de Grenoble, France
ANNA DE MASI	Università dell'Aquila, Italia
GIANPIETRO DEL PIERO	Università di Ferrara and International Research Center MEMOCS, Italia
EMMANUELE DI BENEDETTO	Vanderbilt University, USA
BERNOLD FIEDLER	Freie Universität Berlin, Germany
IRENE M. GAMBA	University of Texas at Austin, USA
DAVID Y. GAO	Federation University and Australian National University, Australia
SERGEY GAVRILYUK	Université Aix-Marseille, France
TIMOTHY J. HEALEY	Cornell University, USA
DOMINIQUE JEULIN	École des Mines, France
ROGER E. KHAYAT	University of Western Ontario, Canada
CORRADO LATTANZIO	Università dell'Aquila, Italy
ROBERT P. LIPTON	Louisiana State University, USA
ANGELO LUONGO	Università dell'Aquila, Italia
ANGELA MADEO	Université de Lyon–INSA (Institut National des Sciences Appliquées), France
JUAN J. MANFREDI	University of Pittsburgh, USA
CARLO MARCHIORO	Università di Roma "La Sapienza", Italia
GÉRARD A. MAUGIN	Université Paris VI, France
ROBERTO NATALINI	Istituto per le Applicazioni del Calcolo "M. Picone", Italy
PATRIZIO NEFF	Universität Duisburg-Essen, Germany
ANDREY PIATNITSKI	Narvik University College, Norway, Russia
ERRICO PRESUTTI	Università di Roma Tor Vergata, Italy
MARIO PULVIRENTI	Università di Roma "La Sapienza", Italia
LUCIO RUSSO	Università di Roma "Tor Vergata", Italia
MIGUEL A. F. SANJUAN	Universidad Rey Juan Carlos, Madrid, Spain
PATRICK SELVADURAI	McGill University, Canada
ALEXANDER P. SEYRANIAN	Moscow State Lomonosov University, Russia
MIROSLAV ŠILHAVÝ	Academy of Sciences of the Czech Republic
GUIDO SWEERS	Universität zu Köln, Germany
ANTOINETTE TORDESILLAS	University of Melbourne, Australia
LEV TRUSKINOVSKY	École Polytechnique, France
JUAN J. L. VELÁZQUEZ	Bonn University, Germany
VINCENZO VESPRI	Università di Firenze, Italia
ANGELO VULPIANI	Università di Roma La Sapienza, Italia

MEMOCS (ISSN 2325-3444 electronic, 2326-7186 printed) is a journal of the International Research Center for the Mathematics and Mechanics of Complex Systems at the Università dell'Aquila, Italy.

Cover image: "Tangle" by © John Horigan; produced using the *Context Free* program (contextfreeart.org).

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

© 2015 Mathematical Sciences Publishers

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 3, No. 2, 2015
dx.doi.org/10.2140/memocs.2015.3.101

DERIVATION OF NONLINEAR SHELL MODELS COMBINING SHEAR AND FLEXURE: APPLICATION TO BIOLOGICAL MEMBRANES

OLIVIER PANTZ AND KARIM TRABELSI

Biological membranes are often idealized as incompressible elastic surfaces whose strain energy only depends on their mean curvature and possibly on their shear. We show that this type of model can be derived using a formal asymptotic method by considering biological membranes to be thin, strongly anisotropic, elastic, locally homogeneous bodies.

1. Introduction

Shells, plates and membranes are solid deformable bodies having one characteristic dimension small by comparison with the other two dimensions. Their behavior is fully described by standard three-dimensional laws of continuum mechanics. Nevertheless, it is tempting, at least from the modeling viewpoint, to consider them as two-dimensional structures and to replace the genuine mechanical laws by two-dimensional reduced versions. This immediately raises two questions: (1) What is the correct model? and (2) How can it be mathematically justified? To this end, we consider the thickness ε of the plate/shell/membrane as a parameter and identify the limit behavior of the structure as ε goes to zero. According to the dependence of the elasticity moduli on the thickness of the shell, a full zoology of models may be derived. Membrane, isometric bending and von Kármán theories have been justified (amongst others), first formally (see Fox, Raoult and Simo [Fox et al. 1993]), then by means of Γ-convergence (see [Le Dret and Raoult 1995; 1996; Pantz 2003], Müller, Friesecke and James [Friesecke et al. 2002], Friesecke, James and Mora [Friesecke et al. 2003]; see also [Conti et al. 2006]). In those works, elasticity coefficients are assumed to scale like a power of the thickness ε of the plate or shell, that is, like $\varepsilon^{-\alpha}$. Membrane theory corresponds to the case $\alpha = 1$,

Communicated by Gilles A. Francfort.
The authors would like to thank the referees for their valuable comments, which helped to improve the manuscript.
This work was partly supported by the ANR Geometrya.
O. Pantz is a member of the DEFI team at INRIA Saclay Île-de-France.
MSC2010: 74K20, 74K25, 74B20.
Keywords: shell, nonlinear elasticity, Helfrich, red blood cell, vesicle.

isometric bending to the case $\alpha = 3$ and von Kármán to $\alpha = 4$. Intermediate values of α have also been considered, and an almost exhaustive hierarchy of models has thus been produced (see Müller, Friesecke and James [Friesecke et al. 2006]). Some cases remain to be treated; Conti and Maggi [2008], for instance, investigate the scaling of the energy corresponding to folds. The initial motivation for this article was the study of the mechanical behavior of red blood cells (RBCs), and our aim was to determine whether the classical RBC model could be derived by the above procedure.

Mature anucleate RBCs[1] are made of two mechanical structures: the cytoskeleton — a two-dimensional network of protein filaments that extends throughout the interior of the cell — and a lipid bilayer. Both are bound together by proteins linking the nodes of the mesh of the cytoskeleton to the lipid bilayer via transmembrane proteins. Lipid bilayers are self-assembled structures of phospholipids, which are small molecules containing a negatively charged phosphate group (called the head), and two highly hydrophobic fatty acid chains (called the tails). In an aqueous environment, phospholipids spontaneously form a double layer whose configuration isolates the hydrophobic tails from the surrounding water molecules. Modifying the area of such a lipid bilayer is energy-costly because it exposes some of the tails to the environment.

A bilayer that supports no other mechanical structure, is connected and has no boundary is called a vesicle. Vesicles are massively studied because they are easy to obtain experimentally. Moreover, they partially mimic the behavior of RBCs. Roughly speaking, they are RBCs without cytoskeleton (even if the RBC bilayer does embed a lot of different proteins responsible for different functions of the cell). They similarly resist bending. However, vesicles show no resistance to shear stress, while RBCs do, owing to their cytoskeletons.

A widely used model consists in considering that a lipid bilayer may be endowed with an elastic energy depending solely on the mean curvature of the vesicle. This energy is usually known as the Helfrich functional (named after Willmore in other contexts). It was introduced independently, as far as we know, by Canham [1970] and Helfrich [1973] some forty years ago. Evans [1974] has shown that the Helfrich functional can be derived by assuming a vesicle to be made of two interconnected elastic fluid membranes, each of them resistant to change of local area but not to bending itself. Jenkins [1977a] has extended the analysis of Helfrich to general two-dimensional liquid crystals [Singer and Nicolson 1972]. In particular, he derives the Euler–Lagrange equations satisfied by the equilibrium states, and examines the consequences of fluidity on the form of the strain energy (see also [Steigmann

[1] Every mention to RBCs in this article will implicitly refer to anucleate mature RBCs without further notice.

1999]). As a means to take into account the various vesicle shapes observed, it is common to presume that the vesicle is endowed with a nonzero spontaneous curvature. The origin of this spontaneous curvature is usually attributed to different compositions of the outer and inner layers. Several refinements to this basic model have since been proposed as the so-called bilayer-couple model [Svetina and Žekš 1989], that consists in allowing the two lipid layers to slip on one another, and imposing that the total area of each layer remains constant (see also [Seifert et al. 1991] for a comparison between the two models). Miao, Seifert, Wortis and Döbereiner [Miao et al. 1994] proposed an intermediate model, called the area-difference elasticity model, where slight total area changes of each layer are allowed but still penalized.

As previously mentioned, the mechanical structure of the RBC is not only imputable to its bilayers. Their cytoskeleton endows them with resistance to shear stress. In most models, only the deformation of the RBC membrane is considered (that is, of the bilayer). To take into account the presence of the cytoskeleton an additional term is added to the total energy depending on the change of the metric of the membrane. Krishnaswamy [1996] proposed another model for which the deformations of the cytoskeleton and the fluid bilayer may differ.

The aforementioned models for vesicles and RBCs are backed up by numerous numerical studies that reproduce various shapes observed experimentally. Amongst others, Deuling and Helfrich [1976] (see also [Jenkins 1977b; Luke 1982; Luke and Kaplan 1979]) have computed axisymmetric vesicle shapes of minimum energy with respect to the values of the reduced volume and spontaneous curvature. Seifert, Berndl and Lipowsky [Seifert et al. 1991] have compared the axisymmetric solutions obtained using the spontaneous curvature model and the bilayer-couple model, whereas Agrawal and Steigmann [2009] have included contact conditions between the vesicle and a substrate. Full three-dimensional simulations have been performed by Feng and Klug [2006], Bonito, Nochetto and Pauletti [Bonito et al. 2010; 2011], Dziuk [2008] using a finite element method. Peng et al. [2013] use a dissipative particle dynamic approach and focus on the interaction between the lipid bilayer and the cytoskeleton. Du, Chun and Xiaoqiang perform numerical computations based on a phase field method [Du et al. 2006; 2004; Du and Zhang 2008]. Boundary integral methods have been used by Veerapaneni, Gueyffier and Zorin [Veerapaneni et al. 2009], Sohn, Tseng, Li, Voigt and Lowengrub [Sohn et al. 2010]. Another approach based on the immersed boundary method has been investigated by Kim and Lai [2010], Liu et al. [2004; 2006] and, together with a lattice Boltzmann approach, by Crowl and Fogelson [2010]. Finally, level set methods have also been implemented in this context by Salac and Miksis [2011], and Maitre, Milcent, Cottet, Raoult and Usson [Maitre et al. 2009] (see also [Doyeux et al. 2013]).

We prove in this article that the classical mechanical model of the RBCs can be recovered by means of a formal asymptotic analysis assuming that the RBC's membrane is made of a locally homogeneous, albeit strongly anisotropic, nonlinearly elastic material. The main difference with previous works on the justification of thin structures is that we assume different scalings for the elastic moduli in the tangential and normal directions to the midsection. Let us underline that our work cannot be considered as a justification of the classical RBC mechanical model. Indeed, the RBC is not a locally homogeneous elastic membrane, firstly because it is made of two different structures: the lipid bilayer (responsible for the resistance to bending) and a cytoskeleton (responsible for resistance to shear). Even the lipid bilayer could hardly be considered as made of a homogeneous material, the scale of the phospholipids it contains being of the same order as the thickness of the membrane. The cytoskeleton, being a two-dimensional spectrin network, is no more a homogeneous elastic body. Even if it is not overt at first glance, our work is strongly related to the justification, already mentioned, proposed in [Evans 1974].

We have chosen to consider a rather general setting (presented in Section 2) for which the modeling of the RBCs is obtained as a particular case (see Section 6). The asymptotic analysis is performed in Section 3. Assuming that the minimizers of the energy admit an asymptotic expansion with respect to the thickness (Section 3.2), they converge toward the solutions of a two-dimensional problem (see Section 3.3). The limit energy, computed in Section 3, contains membrane and flexural terms. In Section 4, we prove that, under invariance assumptions on the stored energy of the material, the flexural term depends only upon the second fundamental form, or even only upon the mean curvature of the shell. The isometric bending shell, RBC and vesicle models are obtained as particular applications in Section 6. The last section is devoted to some general remarks, in particular on the relaxation of the formal energy limit.

Finally, let us specify some notation. If M is a differentiable manifold, we denote by TM and T^*M its tangent and cotangent bundles. Moreover, $T^*(M;\mathbb{R}^3)$ will stand for the triple Whitney sum $T^*M \oplus T^*M \oplus T^*M$. The tangent spaces of a product of manifolds will be implicitly identified with the product of the tangent spaces, so that if M_1 and M_2 are differentiable manifolds and $M = M_1 \times M_2$, the bundle TM will be implicitly identified with $TM_1 \times TM_2$. If M is an open subset of $\mathbb{R}^N$, TM will be identified with $M \times \mathbb{R}^N$. The corresponding identifications will also be made for T^*M and $T^*(M;\mathbb{R}^3)$. The set of reals $\mathbb{R}$ and its dual $\mathbb{R}'$ will also be often implicitly identified. Sets will always be displayed with capital letters (for instance, the set of deformations ψ^ε will be denoted Ψ^ε). Sequences of terms of an asymptotic expansion are denoted using bold letters (for instance, $\boldsymbol{\psi} = (\psi_k)_{k\in\mathbb{N}}$ stands for the asymptotic expansion of ψ^ε). Accordingly, the sets of asymptotic expansions use both bold and capitalized letters (for instance, $\boldsymbol{\psi} \in \boldsymbol{\Psi}$). Moreover,

calligraphic letters will be exclusively used for fiber spaces. Two different reference configurations are used throughout our article; one is qualified to be abstract and the other geometrical. The same notations are used for both configurations, the only distinction being that a tilde is added over variables, sets and functionals defined on the geometric configuration (for instance, $\tilde{\psi}^\varepsilon$ is the deformation defined on the geometric configuration, whereas ψ^ε stands for the deformation over the abstract one). All of the notation introduced is recalled at the end of the article for convenience.

2. Elastic shells — three-dimensional modeling

We consider a thin nonlinearly elastic shell of midsurface S' and constant half-thickness $\varepsilon > 0$, and choose $S^\varepsilon = S' \times (-\varepsilon, \varepsilon)$ to be the reference configuration of this elastic body. We assume S' to be a regular two-dimensional orientable submanifold of $\mathbb{R}^3$ with or without boundary. In the following, S' is implicitly endowed with the metric induced by the Euclidean metric in $\mathbb{R}^3$. Let ψ^ε be the deformation of the shell, that is, a map from S^ε into $\mathbb{R}^3$. The differential $D\psi^\varepsilon(x^\varepsilon)$ of ψ^ε at x^ε is a linear map from $T_{x^\varepsilon}S^\varepsilon$ into $T_{\psi^\varepsilon(x^\varepsilon)}\mathbb{R}^3$. Since $T_{\psi^\varepsilon(x^\varepsilon)}\mathbb{R}^3$ is canonically isomorphic to $\mathbb{R}^3$, $D\psi^\varepsilon(x^\varepsilon)$ is identified with an element of the Whitney sum $T^*S^\varepsilon \oplus T^*S^\varepsilon \oplus T^*S^\varepsilon$ denoted by $T^*(S^\varepsilon; \mathbb{R}^3)$. We denote by $J_\varepsilon(\psi^\varepsilon)$ the elastic energy of the shell under the deformation ψ^ε. We assume that the elastic energy is local and depends only on the first derivatives of the deformation. In other words, there exists a map W^ε from $T^*(S^\varepsilon; \mathbb{R}^3)$ into $\overline{\mathbb{R}}^+$ such that

$$J_\varepsilon(\psi^\varepsilon) := \int_{S^\varepsilon} W^\varepsilon(D\psi^\varepsilon)\, dx^\varepsilon,$$

where $dx^\varepsilon = dx' \wedge dx_3^\varepsilon$ and dx' is the two-dimensional Hausdorff measure restricted to S', whereas $D\psi^\varepsilon(x^\varepsilon)$ stands for the differential of ψ^ε at $x^\varepsilon \in S^\varepsilon$. Note that this representation enables us to consider inhomogeneous shells. The shell is assumed to be subjected to volumic dead-body loads $f_\varepsilon \in L^2(S^\varepsilon)^3$, and we set

$$L_\varepsilon(\psi^\varepsilon) := \int_{S^\varepsilon} f_\varepsilon \cdot \psi^\varepsilon\, dx^\varepsilon.$$

The total energy of the system is accordingly given by

$$I_\varepsilon(\psi^\varepsilon) := J_\varepsilon(\psi^\varepsilon) - L_\varepsilon(\psi^\varepsilon).$$

Finally, boundary conditions may also be added. We set $\Gamma^\varepsilon = \gamma \times (-\varepsilon, \varepsilon)$, where $\gamma \subset \partial S'$ is the — possibly empty — part of the boundary where the shell is clamped, and we denote by ϕ^ε the imposed deformation on this set. Our aim is to determine the behavior of the minimizers φ^ε of I_ε over

$$\Psi^\varepsilon := \{\psi^\varepsilon \in W^{1,\infty}(S^\varepsilon)^3 : \psi^\varepsilon(x^\varepsilon) = \phi^\varepsilon(x^\varepsilon) \text{ for every } x^\varepsilon \in \Gamma^\varepsilon\}$$

as ε goes to zero. Note that the minimization problem of I_ε over Ψ^ε without any growth and polyconvex or quasiconvex assumptions on the stored energy function is generally not well posed. Here, we implicitly assume this problem to have a regular solution. Various assumptions have to be made regarding the dependence of the energy on the thickness for the needs of our analysis. These mainly concern the stored energy W^ε (see Section 2.1), but also the applied loads (see Section 2.2).

2.1. ***Dependence of the stored energy functions with respect to the thickness.*** We set $S = S^1$, and assume the stored energy W^ε to be of the form

$$W^\varepsilon(F) = \varepsilon^{-1}(\varepsilon^{-2}W_2(F) + W_0(F)) \tag{1}$$

for every $F \in T^*(S^\varepsilon; \mathbb{R}^3)$ and $\varepsilon \leq 1$, where W_0 and W_2 are continuous nonnegative maps from $T^*(S; \mathbb{R}^3)$ into $\overline{\mathbb{R}}^+$. Note that we implicitly use the injection of

$$T^*(S^\varepsilon; \mathbb{R}^3) = T^*(S' \times (-\varepsilon, \varepsilon); \mathbb{R}^3) = T^*(S'; \mathbb{R}^3) \times T^*((-\varepsilon, \varepsilon); \mathbb{R}^3)$$

into

$$T^*(S; \mathbb{R}^3) = T^*(S' \times (-1, 1); \mathbb{R}^3) = T^*(S'; \mathbb{R}^3) \times T^*((-1, 1); \mathbb{R}^3)$$

in the definition (1) of W^ε. Standard analysis focuses on the case where only one element of this expansion is not zero. For instance, if $W_2 = 0$, we recover a nonlinear membrane model [Le Dret and Raoult 1996], and if $W_0 = 0$ we obtain the isometric bending one [Friesecke et al. 2003].

Behavior of strongly extended fibers. We assume that the stored energy W_2 is bounded from below by a positive constant for strongly extended fibers, namely, there exist $\delta, c > 0$ such that

$$W_2(F', F_3) \geq c \quad \text{for all } F' \in T^*(S'; \mathbb{R}^3),\ F_3 \in T^*((-1, 1); \mathbb{R}^3) \text{ such that } |F_3| \geq \delta. \tag{2}$$

Note that for every element F of a vector bundle endowed with a Riemann metric, the notation $|F|$ should be understood as the norm of the vectorial part of F. In particular, in (2), $|F_3| = |v|$ if $F_3 = (x_3, v) \in (-1, 1) \times \mathbb{R}^3 = T^*((-1, 1); \mathbb{R}^3)$.

Regularity and zero set of W_2. We assume that W_2 is a nonnegative C^2 function and denote by $\mathcal{M}$ the restriction of its zero set to the midsection, that is,

$$\mathcal{M} := \{F \in T^*(S'; \mathbb{R}^3) \times T_0^*((-1, 1); \mathbb{R}^3) : W_2(F) = 0\}.$$

Let $\mathcal{M}'$ be the projection of $\mathcal{M}$ onto $T^*(S'; \mathbb{R}^3)$, that is,

$$\mathcal{M}' := \{F' \in T^*(S'; \mathbb{R}^3) : \text{there exists } n_0 \in T_0^*((-1, 1); \mathbb{R}^3) \text{ with } (F', n_0) \in \mathcal{M}\}.$$

We assume that the projection of $\mathcal{M}$ onto $\mathcal{M}'$ is one-to-one. We denote by $n_0 : \mathcal{M}' \to T_0^*((-1,1);\mathbb{R}^3)$ the function that maps every element F' of $\mathcal{M}'$ to the corresponding element F_3 of $T_0^*((-1,1);\mathbb{R}^3)$, so that

$$\mathcal{M} = \{(F', n_0(F')) \in T^*(S;\mathbb{R}^3) : F' \in \mathcal{M}'\}. \tag{3}$$

We recall that $S = S' \times (-1,1)$ and that $T^*(S;\mathbb{R}^3)$ is identified with $T^*(S';\mathbb{R}^3) \times T^*((-1,1);\mathbb{R}^3)$. The vectorial part of $n_0(F') \in T_0^*((-1,1);\mathbb{R}^3) = \{0\} \times \mathbb{R}'^3$ will be denoted $n(F')$, so that $n_0(F') = (0, n(F'))$.

Local interpenetration. To avoid local interpenetration of matter, it is geometric to expect $D\psi^\varepsilon$ to be invertible. To this end, we require that $W_0(F) = \infty$ for every $F \in \mathcal{M}$ such that $\det F < 0$, and that

$$W_0(F) \to \infty \quad \text{if } F \in \mathcal{M} \text{ and } \det(F) \to 0. \tag{4}$$

2.2. ***Dependence of the applied loads on the thickness.*** The volumic loads are assumed to scale as the inverse of the thickness of the shell; more precisely, we assume that there exists $f : S \to \mathbb{R}^3$ such that, for every $\varepsilon \le 1$,

$$f_\varepsilon(x) = \varepsilon^{-1} f(x) \quad \text{for every } x \in S^\varepsilon. \tag{5}$$

3. From 3D to 2D: a formal asymptotic analysis

3.1. ***Rescaling.*** We set $\psi(\varepsilon)(x', x_3) = \psi^\varepsilon(x', \varepsilon x_3)$, and define rescaled energies

$$J(\varepsilon)(\psi(\varepsilon)) := J_\varepsilon(\psi^\varepsilon) \quad \text{and} \quad I(\varepsilon)(\psi(\varepsilon)) := I_\varepsilon(\psi^\varepsilon).$$

The minimization problem of I_ε over Ψ^ε is then equivalent to the minimization problem of $I(\varepsilon)$ over

$$\Psi(\varepsilon) := \{\psi(\varepsilon) \in W^{1,\infty}(S)^3 : \psi(\varepsilon)(x) = \phi(\varepsilon)(x) \text{ for every } x \in \Gamma\},$$

where $\phi(\varepsilon)(x) = \phi^\varepsilon(x', \varepsilon x_3)$.

For every map $\psi^\varepsilon : S^\varepsilon \to \mathbb{R}^3$, we denote by $(D'\psi^\varepsilon, D_3\psi^\varepsilon)$ the decomposition of the differential ψ^ε along the sections of the cylinder S^ε and along its fibers, respectively. In other words, for every $x^\varepsilon = (x', x_3) \in S^\varepsilon$, $D'\psi^\varepsilon(x^\varepsilon)$ and $D_3\psi^\varepsilon(x^\varepsilon)$ stand for the elements of $T^*_{x'}(S';\mathbb{R}^3)$ and $T^*_{x_3}((-1,1);\mathbb{R}^3)$ such that $D\psi^\varepsilon(x^\varepsilon) = (D'\psi^\varepsilon(x^\varepsilon), D_3\psi^\varepsilon(x^\varepsilon))$.

For every deformation $\psi(\varepsilon)$ of S, we define its partial derivative $\partial_3\psi(\varepsilon)$ with respect to the normal direction as

$$\partial_3\psi(\varepsilon)(x', x_3) = \lim_{t\to 0} \frac{\psi(\varepsilon)(x', x_3 + t) - \psi(\varepsilon)(x', x_3)}{t}.$$

Performing a simple change of variable, we get

$$\begin{aligned} J(\varepsilon)(\psi(\varepsilon)) &= \varepsilon^{-1}\int_{S^\varepsilon}(\varepsilon^{-2}W_2+W_0)\big(D'\psi^\varepsilon(x^\varepsilon),\, D_3\psi^\varepsilon(x^\varepsilon)\big)\,dx^\varepsilon \\ &= \varepsilon^{-1}\int_{S^\varepsilon}(\varepsilon^{-2}W_2+W_0)\big(D'\psi^\varepsilon(x^\varepsilon),\, (x_3^\varepsilon,\, \partial_3\psi^\varepsilon(x^\varepsilon))\big)\,dx^\varepsilon \\ &= \int_S(\varepsilon^{-2}W_2+W_0)\big(D'\psi(\varepsilon)(x),\, (\varepsilon x_3,\, \varepsilon^{-1}\partial_3\psi(\varepsilon)(x))\big)\,dx. \end{aligned}$$

3.2. ***Ansatz.*** In order to perform our formal analysis, we assume that the minimizers $\varphi(\varepsilon)(x', x_3) = \varphi^\varepsilon(x', \varepsilon x_3)$ of the energy admit an asymptotic expansion

$$\varphi(\varepsilon)(x) = \sum_{k\geq 0}\varepsilon^k\varphi_k(x) \quad \text{for every } x\in S, \tag{6}$$

with $(\varphi_k)\in \ell^1(W^{1,\infty}(S)^3)$. Obviously, the same assumption has to be made on the applied Dirichlet boundary condition, and we let $\boldsymbol{\phi} = (\phi_k)\in \ell^1(W^{1,\infty}(S)^3)$ be the terms of the asymptotic expansion of the deformation $\phi(\varepsilon)(x) = \phi^\varepsilon(x', \varepsilon x_3)$ imposed on $\Gamma := \gamma\times(-1,1)$, that is

$$\phi(\varepsilon)(x) = \sum_{k\geq 0}\varepsilon^k\phi_k(x) \quad \text{for every } x\in S. \tag{7}$$

The condition $\varphi^\varepsilon\in\Psi^\varepsilon$ reads as $\varphi_k(x) = \phi_k(x)$ for every $x\in\Gamma$. Consequently, we introduce the admissible set

$$\boldsymbol{\Psi} := \{\boldsymbol{\psi} = (\psi_k)\in\ell^1(W^{1,\infty}(S)^3) : \psi_k = \phi_k \text{ for every } x\in\Gamma\},$$

and the rescaled energies $\boldsymbol{J}(\varepsilon)$ and $\boldsymbol{I}(\varepsilon)$ from $\boldsymbol{\Psi}$ into $\overline{\mathbb{R}}$ defined by

$$\boldsymbol{J}(\varepsilon)(\boldsymbol{\psi}) := J(\varepsilon)\bigg(\sum_{k\geq 0}\varepsilon^k\psi_k\bigg) \quad \text{and} \quad \boldsymbol{I}(\varepsilon)(\boldsymbol{\psi}) := I(\varepsilon)\bigg(\sum_{k\geq 0}\varepsilon^k\psi_k\bigg). \tag{8}$$

3.3. ***Limit of the total energy.*** The first step of our analysis consists in computing the limit of $\boldsymbol{J}(\varepsilon)(\boldsymbol{\psi})$ as ε goes to zero for $\boldsymbol{\psi}\in\boldsymbol{\Psi}$. As we shall see in Proposition 1, the limit of $\boldsymbol{J}(\varepsilon)$ contains two terms. Roughly speaking, one term measures the elastic energy due to the change of the metric of the midsection of the shell. It depends only on W_0. The second term measures the elastic energy due to the variations of the orientation of its fibers. It depends on the second derivative of the stored energy function W_2 through a quadratic form $Q_{D'\psi_0}$.

In order to enhance the readability of the sequel, we introduce a practical notation. We recall that a section F of a vector bundle $\mathcal{F}$ is a map from its base into $\mathcal{F}$ such that $\pi_B(F)$ is the identity, where π_B stands for the projection of $\mathcal{F}$ onto its

base B. Given such a section, we define the bundle map

$$\mathscr{F} \to T\mathscr{F}, \quad G \mapsto G_F = \frac{d}{dt}(F(\pi_B(G)) + tG)|_{t=0}. \tag{9}$$

Roughly speaking, G_F is the element G of $T_{F(\pi_B(G))}\mathscr{F}$. Similarly, for every $(x, v) \in \mathbb{R}^N \times \mathbb{R}^N$, we will sometimes denote $(x, v) \in T_x\mathbb{R}^N$ by v_x. For a section F' of $\mathscr{M}'$, for every $(G', s, v) \in T^*(S'; \mathbb{R}^3) \times \mathbb{R} \times (\mathbb{R}')^3$ we set

$$Q_{F'}(G', s, v) := D^2 W_2[G'_{F'}, s_0, v_{n(F')}]^2, \tag{10}$$

where $(G'_{F'}, s_0, v_{n(F')})$ is the element of $T_{(F',\, n_0(F'))}(T^*(S; \mathbb{R}^3))$ defined in (9) based on the decomposition of $T^*(S; \mathbb{R}^3) = T^*(S'; \mathbb{R}^3) \times (-1, 1) \times (\mathbb{R}')^3$, while $D^2 W_2$ stands for the Hessian of W_2. Namely, we have

$$(G'_{F'}, s_0, v_{n(F')}) = \frac{d\gamma}{dt}(0), \quad \text{where } \gamma(t) = (F'_{\pi_{S'}(G')} + tG', ts, n(F') + tv). \tag{11}$$

At first glance, the meaning of $D^2 W_2[\dot\gamma(0)]^2$ is unclear, considering that the Hessian of a map defined on a manifold is not, in general, intrinsically defined. Nevertheless, it is well known that this is consistent on the set of critical points, which is precisely what is considered here. Indeed, $\gamma(0)$ is equal to the value of the section $(F', n_0(F'))$ at $\pi_{S'}(G)$. Yet F' is a section of $\mathscr{M}'$, hence $W_2(F', n_0(F)) = 0$, $W_2(\gamma(0)) = 0$ and $DW_2(\gamma(0)) = 0$. As a result, $D^2 W_2[\dot\gamma(0)]^2$ is well defined, and, accordingly,

$$D^2 W_2[\dot\gamma(0)]^2 = 2 \lim_{t \to 0} t^{-2} W_2(\gamma(t)). \tag{12}$$

Note that the right-hand side of (12) only depends on $\dot\gamma(0)$, so that the particular choice of the representative $\gamma(t)$ of $\dot\gamma(0)$ is irrelevant, as already mentioned.

We are now in a position to state the main result of this section.

Proposition 1. *Let $\mathbf{\Phi}$ be the subset of the admissible set $\mathbf{\Psi}$ defined by*

$$\mathbf{\Phi} := \{\boldsymbol{\psi} \in \mathbf{\Psi} : \partial_3 \psi_0 = 0,\ D'\psi_0(x) \in \mathscr{M}',\ \textit{and } \partial_3 \psi_1(x) = n(D'\psi_0(x)) \textit{ for every } x \in S\}. \tag{13}$$

Let $\boldsymbol{\psi} \in \mathbf{\Psi}$. Then

$$\lim_{\varepsilon \to 0} \boldsymbol{I}(\varepsilon)(\boldsymbol{\psi}) = \begin{cases} I_0\big(\psi_0, \frac{1}{2}\int_{-1}^{1} \psi_1\, dx_3, \partial_3 \psi_2\big) & \textit{if } \boldsymbol{\psi} \in \mathbf{\Phi}, \\ +\infty & \textit{if } \boldsymbol{\psi} \notin \mathbf{\Phi}, \end{cases}$$

where

$$I_0(\psi_0, u, v) := J_0(\psi_0, u, v) - 2\int_{S'} f_0 \cdot \psi_0\, dx,$$

$$J_0(\psi_0, u, v) := \frac{1}{2}\int_S Q_{D'\psi_0}(D'u + x_3 D'n, x_3, v)\, dx + 2\int_{S'} W_0(D'\psi_0, n_0)\, dx',$$

where n and n_0 stand for $n(D'\psi_0)$ and $n_0(D'\psi_0)$, $Q_{D'\psi_0}$ is defined by (10), *and* $f_0(x') = f(x', 0)$.

Proof. We proceed in two steps. First, we prove that every sequence of deformations $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ of finite elastic energy, that is, those that satisfy

$$\liminf_{\varepsilon\to 0} \boldsymbol{J}(\varepsilon)(\boldsymbol{\psi}) < +\infty,$$

belongs to $\boldsymbol{\Phi}$. In particular, this implies that $\boldsymbol{J}(\varepsilon)(\boldsymbol{\psi})$ converges to infinity as ε goes to zero, for every $\boldsymbol{\psi}$ that is not in $\boldsymbol{\Phi}$. In a second step, we compute the limit of $\boldsymbol{J}(\varepsilon)(\boldsymbol{\psi})$ for every $\boldsymbol{\psi}$ in $\boldsymbol{\Phi}$.

Let $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ be the asymptotic expansion of a deformation of finite elastic energy. From Fatou's lemma, we deduce

$$\begin{aligned}
&\int_S \liminf_{\varepsilon\to 0}(\varepsilon^{-2}W_2)\Big(\sum_{k\geq 0}\varepsilon^k(D'\psi_k, (\varepsilon x_3, \varepsilon^{-1}\partial_3\psi_k))\Big)\,dx \\
&\qquad \leq \liminf_{\varepsilon\to 0}\int_S (\varepsilon^{-2}W_2 + W_0)\Big(\sum_{k\geq 0}\varepsilon^k(D'\psi_k, (\varepsilon x_3, \varepsilon^{-1}\partial_3\psi_k))\Big)\,dx \\
&\qquad = \liminf_{\varepsilon\to 0} \boldsymbol{J}(\varepsilon)(\boldsymbol{\psi}) < \infty.
\end{aligned}$$

Hence, we have

$$\liminf_{\varepsilon\to 0} W_2\Big(\sum_{k\geq 0}\varepsilon^k(D'\psi_k, (\varepsilon x_3, \varepsilon^{-1}\partial_3\psi_k))\Big) = 0 \quad \text{a.e.}$$

From the assumption (2) made on the behavior of strongly extended fibers, it follows that, for almost every $x \in S$, $\sum_{k\geq 0}\varepsilon^{k-1}\partial_3\psi_k$ remains bounded (up to a subsequence), that is, $\partial_3\psi_0 = 0$. Now, since

$$\sum_{k\geq 0}\varepsilon^k(D'\psi_k, (\varepsilon x_3, \varepsilon^{-1}\partial_3\psi_k)) \xrightarrow[\varepsilon\to 0]{L^\infty} (D'\psi_0, (0, \partial_3\psi_1)), \tag{14}$$

and W_2 is assumed to be continuous, we have $W_2(D'\psi_0, (0, \partial_3\psi_1)) = 0$ almost everywhere. From the hypothesis (3), we get $D'\psi_0 \in \mathcal{M}'$ and $\partial_3\psi_1 = n(D'\psi_0)$. As a conclusion, every sequence of deformations of finite elastic energy belongs to $\boldsymbol{\Phi}$, as announced.

For the next step, let us consider an element $\boldsymbol{\psi} \in \boldsymbol{\Phi}$ and its associated energy

$$\begin{aligned}
\lim_{\varepsilon\to 0} \boldsymbol{J}(\varepsilon)(\boldsymbol{\psi}) = \lim_{\varepsilon\to 0} \varepsilon^{-2}\int_S W_2\Big(\sum_{k\geq 0}\varepsilon^k(D'\psi_k, (\varepsilon x_3, \varepsilon^{-1}\partial_3\psi_k))\Big)\,dx \\
+ \int_S W_0(D'\psi_0, (0, \partial_3\psi_1))\,dx.
\end{aligned}$$

Since W_2 is a C^2 function and $W_2(D'\psi_0, (0, \partial_3\psi_1)) = 0$, using (12) with $\gamma(t) = (D'\psi_0 + tD'\psi_1, tx_3, \partial_3\psi_1 + t\partial_3\psi_2)$, Lebesgue's theorem implies

$$\begin{aligned}
&\lim_{\varepsilon\to 0} \varepsilon^{-2} \int_S W_2\Big(\sum_{k\geq 0} \varepsilon^k \big(D'\psi_k, (\varepsilon x_3, \varepsilon^{-1}\partial_3\psi_k)\big)\Big)\, dx \\
&\quad = \frac{1}{2}\int_S D^2 W_2\big[(D'\psi_{1\,D'\psi_0(x)}, x_{30}, \partial_3\psi_{2\partial_3\psi_1})\big]^2\, dx \\
&\quad = \frac{1}{2}\int_S D^2 W_2\big[(D'\psi_{1\,D'\psi_0(x)}, x_{30}, \partial_3\psi_{2n(D'\psi_0)})\big]^2\, dx \\
&\quad = \frac{1}{2}\int_S Q_{D'\psi_0}(D'\psi_1, x_3, \partial_3\psi_2)\, dx.
\end{aligned}$$

The limit of the elastic energy $\boldsymbol{J}(\varepsilon)$ falls out from the fact that ψ_1 may be written as

$$\psi_1 = \frac{1}{2}\int_{-1}^{1} \psi_1\, dx_3 + x_3 n(D'\psi_0).$$

Finally, according to the definition (5) of f, we have

$$\boldsymbol{I}(\varepsilon)(\boldsymbol{\psi}) = \boldsymbol{J}(\varepsilon)(\boldsymbol{\psi}) - \int_S f(x', \varepsilon x_3)\cdot\Big(\sum_{k\geq 0}\varepsilon^k\psi_k\Big)\, dx.$$

The second term on the right-hand side converges to $2\int_{S'} f\cdot\psi_0\, dx$ as $\varepsilon\to 0$. □

Since the limit energy is finite only for elements $\boldsymbol{\psi}$ in $\boldsymbol{\Phi}$, $\boldsymbol{\phi}$ has to be equal to an element of $\boldsymbol{\Phi}$ on the subset Γ of the boundary where clamping conditions are imposed.

Corollary 2. *If the minimizers $\varphi(\varepsilon)$ of the total rescaled energy $I(\varepsilon)$ over $\Psi(\varepsilon)$ admit an asymptotic expansion as in* (6), *and if their total energy $I(\varepsilon)(\varphi(\varepsilon))$ remains bounded, then $\phi_0(x', x_3)$ depends only on $x' \in \Gamma$. In addition, there exist $u_\gamma, n_\gamma \in W^{1,\infty}(\gamma)^3$ such that*

$$\phi_1(x', x_3) = u_\gamma(x') + x_3 n_\gamma(x') \quad \text{for } x \in \Gamma.$$

Note: since ϕ_0 depends only on x', we shall write $\phi_0(x')$ instead of $\phi_0(x', x_3)$ henceforth.

3.4. *Convergence of the minimizers.*

Lemma 3. *If the minimizers $\varphi(\varepsilon)$ of the total rescaled energy $I(\varepsilon)$ over $\Psi(\varepsilon)$ admit an asymptotic expansion as in* (6), *and if their total energy $I(\varepsilon)(\varphi(\varepsilon))$ remains bounded, then*

$$I_0\Big(\varphi_0, \frac{1}{2}\int_{-1}^{1}\varphi_1\, dx_3, \partial_3\varphi_2\Big) \leq \inf_{(\psi_0,u,v)\in\Phi_0} I_0(\psi_0, u, v),$$

where Φ_0 is the set of all

$$(\psi_0, u, v) \in W^{1,\infty}(S')^3 \times W^{1,\infty}(S')^3 \times W^{1,\infty}(S'; L^\infty(-1,1)^3)$$

such that $D'\psi_0 \in \mathcal{M}'$ a.e., $n(D'\psi_0) \in W^{1,\infty}(S')^3$, $\psi_0(x') = \phi_0(x')$, $u(x') = u_\gamma(x')$, $n(D'\psi_0(x')) = n_\gamma(x')$ for every $x' \in \gamma$, and $v(x) = \partial_3\phi_2(x)$ for $x \in \Gamma$.

Proof. Let (φ_k) be the asymptotic expansion of a minimizer $\varphi(\varepsilon)$ of the total energy $I(\varepsilon)$. For every $(\psi_0, u, v) \in \Phi_0$, we set

$$\begin{aligned} \psi_1 &= u + x_3 n(D'\psi_0), \\ \psi_2(x) &= \phi_2(x', 0) + \int_0^{x_3} v(x', s)\, ds, \\ \psi_k &= \phi_k \quad \text{for all } k \geq 3. \end{aligned}$$

Using the Leibniz integral rule, we get that $D'\psi_2 = D'\phi_2 + \int_0^{x_3} D'v(x', s)\, ds$ and $\partial_3\psi_2 = v$. It follows that ψ_2 belongs to $W^{1,\infty}(S)^3$. As u and $n(D'\psi_0)$ are assumed to be Lipschitzian, so is ψ_1, and $\boldsymbol{\psi} \in \ell^1(W^{1,\infty}(S)^3)$. From Corollary 2, ϕ_0 depends only on x' and $\phi_1(x', x_3) = u_\gamma(x') + x_3 n_\gamma(x')$ on Γ. As $(\psi_0, u, v) \in \Phi_0$, we infer that $\psi_0 = \phi_0$ and $\psi_1 = \phi_1$ on Γ. Similarly, $\psi_2 = \phi_2$ on Γ. Thus, $\boldsymbol{\psi}$ belongs to $\boldsymbol{\Phi}$ and $\boldsymbol{I}(\varepsilon)(\boldsymbol{\varphi}) \leq \boldsymbol{I}(\varepsilon)(\boldsymbol{\psi})$. Letting ε goes to zero, we get from Proposition 1 that

$$I_0\left(\varphi_0, \frac{1}{2}\int_{-1}^{1} \varphi_1\, dx_3, \partial_3\varphi_2\right) \leq I_0(\psi_0, u, v). \tag{15}$$

This completes the proof. □

Lemma 4. *If the minimizers $\varphi(\varepsilon)$ of the total rescaled energy $I(\varepsilon)$ over $\Psi(\varepsilon)$ admit an asymptotic expansion as in* (6)*, and if their total energy $I(\varepsilon)(\varphi(\varepsilon))$ remains bounded, then*

$$\left(\varphi_0, \frac{1}{2}\int_{-1}^{1} \varphi_1\, dx_3\right) = \arg\min_{(\psi_0,u)\in\Phi_1} I_1(\psi_0, u),$$

where Φ_1 is the set of all (ψ_0, u) in $W^{1,\infty}(S')^3 \times W^{1,\infty}(S')^3$ such that $D'\psi_0 \in \mathcal{M}'$ a.e., $n(D'\psi_0) \in W^{1,\infty}(S')^3$, $\psi_0(x') = \phi_0(x')$, $u(x') = u_\gamma(x')$, and $n(D'\psi_0(x')) = n_\gamma(x')$ for $x' \in \gamma$, I_1 is given by

$$\begin{aligned} I_1(\psi_0, u) := \int_{S'} Q^0_{D'\psi_0}(D'u, 0)\, dx' + \frac{1}{3}\int_{S'} Q^0_{D'\psi_0}(D'n, 1)\, dx' \\ + 2\int_{S'} W_0(D'\psi_0, n_0(D'\psi_0))\, dx' - 2\int_{S'} f_0 \cdot \psi_0\, dx', \end{aligned}$$

$f_0(x') = f(x', 0)$ for every $x' \in S'$, and

$$Q^0_{F'}(G', x_3) = \inf_{v\in\mathbb{R}^3} Q_{F'}(G', x_3, v). \tag{16}$$

Proof. From Lemma 3, we have

$$I_0\left(\varphi_0, \frac{1}{2}\int_{-1}^{1}\varphi_1\,dx_3, \partial_3\varphi_2\right) \le \inf_{(\psi_0,u,v)\in\Phi_0} I_0(\psi_0, u, v).$$

Moreover, from Proposition 1, we have $\boldsymbol{\varphi} \in \boldsymbol{\Phi}$, which implies that

$$\left(\varphi_0, \frac{1}{2}\int_{-1}^{1}\varphi_1\,dx_3\right) \in \Phi_1.$$

Since $\Phi_0 = \Phi_1 \times V$ with

$$V := \{v \in W^{1,\infty}(S'; L^\infty(-1,1)^3) : v(x) = \partial_3\phi_2(x) \text{ for } x \in \Gamma\},$$

it follows that

$$I_1\left(\varphi_0, \frac{1}{2}\int_{-1}^{1}\varphi_1\,dx_3\right) = \inf_{(\psi_0,u)\in\Phi_1}\inf_{v\in V} I_0(\psi_0, u, v).$$

To complete the proof, we need to show that, for every $(\psi_0, u) \in \Phi_1$, we have

$$\inf_{v\in V} I_0(\psi_0, u, v) = I_1(\psi_0, u). \tag{17}$$

We recall that for every $(\psi_0, u) \in \Phi_1$ and every $v \in V$, we have

$$I_0(\psi_0, u, v) = \frac{1}{2}\int_S Q_{D'\psi_0}(D'u + x_3D'n, x_3, v)\,dx + 2\int_{S'} W_0(D'\psi_0, n_0(D'\psi_0))\,dx' - 2\int_{S'} f_0\cdot\psi_0\,dx'.$$

Next, the definition of $Q^0_{D'\psi_0}$ entails that

$$I_0(\psi_0, u, v) \ge \frac{1}{2}\int_S Q^0_{D'\psi_0}((D'u, 0) + x_3(D'n, 1))\,dx + 2\int_{S'} W_0(D'\psi_0, n_0(D'\psi_0))\,dx' - 2\int_{S'} f_0\cdot\psi_0\,dx'.$$

Furthermore, for every $x' \in S'$ and every $F \in T^*_{(x',0)}(S; \mathbb{R}^3)$, the quadratic form $Q^0_{F'}$ derives from a bilinear form. Hence,

$$\begin{aligned}\int_{-1}^{1} Q^0_{D'\psi_0}\big((D'u, 0) + x_3(D'n, 1)\big)\,dx_3 &= \int_{-1}^{1}\big(Q^0_{D'\psi_0}(D'u, 0) + x_3^2 Q^0_{D'\psi_0}(D'n, 1)\big)\,dx_3\\ &= 2Q^0_{D'\psi_0}(D'u, 0) + \tfrac{2}{3}Q^0_{D'\psi_0}(D'n, 1).\end{aligned}$$

Accordingly, we obtain that $I_0(\psi_0, u, v) \geq I_1(\psi_0, u)$, so that $\inf_{v\in V} I_0(\psi_0, u, v) \geq I_1(\psi_0, u)$. It remains to prove the reverse inequality to establish (17). For every $\delta \geq 0$, we have

$$I_0(\psi_0, u, v) \leq I_0(\psi_0, u, v) + \int_S \delta|v|^2\, dx.$$

As a consequence,

$$\inf_{v\in V} I_0(\psi_0, u, v) \leq \inf_{v\in V}\left(I_0(\psi_0, u, v) + \int_S \delta|v|^2\, dx\right).$$

From the assumptions made on W_2, it follows that the quadratic form $Q_{D'\psi_0}$ is positive semidefinite almost everywhere. As a consequence, the map

$$v \mapsto Q_{D'\psi_0}(D'u + x_3 D'n, x_3, v) + \delta|v|^2$$

admits a unique minimizer for almost every $x \in S$. Let $v_\delta : S \to \mathbb{R}^3$ be the map such that $v_\delta = \arg\min_v Q_{D'\psi_0}(D'u + x_3 D'n, x_3, v) + \delta|v|^2$.

Since W_2 is assumed to be of class C^2 and $D'\psi_0$ is bounded, the norm of the quadratic form $Q_{D'\psi_0}$ is uniformly bounded. As a result, v_δ is measurable and belongs to $L^\infty(S)^3$. Also, there exists a sequence v_δ^k in V converging to v_δ in $L^2(S)^3$ as k goes to infinity, due to the density of V in $L^2(S)^3$. For every k, we have

$$\inf_{v\in V} I_0(\psi_0, u, v) \leq \frac{1}{2}\int_S \left(Q_{D'\psi_0}(D'u + x_3 D'n, x_3, v_\delta^k) + \delta|v_\delta^k|^2\right) dx \\ + 2\int_{S'} W_0(D'\psi_0, n_0(D'\psi_0))\, dx' - 2\int_{S'} f_0 \cdot \psi_0\, dx'.$$

Taking the limit with respect to k, we infer that

$$\inf_{v\in V} I_0(\psi_0, u, v) \leq \frac{1}{2}\int_S \left(Q_{D'\psi_0}(D'u + x_3 D'n, x_3, v_\delta) + \delta|v_\delta|^2\right) dx \\ + 2\int_{S'} W_0(D'\psi_0, n_0(D'\psi_0))\, dx' - 2\int_{S'} f_0 \cdot \psi_0\, dx'.$$

Note that $Q_{D'\psi_0}(D'u + x_3 D'n, x_3, v_\delta) + \delta|v_\delta|^2$ is a decreasing sequence (as δ goes to zero) of nonnegative functions. Therefore, its integral over S converges to its pointwise limit $Q^0_{D'\psi_0}(D'u + x_3 D'n, x_3)$, and the intended inequality follows:

$$\inf_{v\in V} I_0(\psi_0, u, v) \leq \frac{1}{2}\int_S Q^0_{D'\psi_0}(D'u + x_3 D'n, x_3)\, dx \\ + 2\int_{S'} W_0(D'\psi_0, n)\, dx' - 2\int_{S'} f_0 \cdot \psi_0\, dx' = I_1(\psi_0, u). \quad \square$$

3.5. ***Boundary conditions.*** An interesting feature of the limit energy is that it depends on both ψ_0 and $u = \frac{1}{2}\int \psi_1\, dx_3$. For general boundary conditions, it implies a coupling between both quantities through the term $\int_{S'} Q^0_{D'\psi_0}(D'u, 0)\, dx'$ of $I_1(\psi_0, u)$. Hence, small perturbations scaling as the thickness of the shell may have an influence on the deformation ψ_0 of the midsection. In the literature, the boundary conditions are usually chosen to satisfy $u_\gamma = 0$, that is,

$$\phi_0(x) = \phi_0(x') \quad \text{and} \quad \phi_1(x) = x_3 n_\gamma(x') \quad \text{for every } (x', x_3) \in \Gamma^\varepsilon, \tag{18}$$

where n_γ is a unit vector. In this case, the minimization of $I_1(\psi_0, u)$ with respect to u is trivial and the limit energy can be expressed solely in terms of ψ_0.

Proposition 5. *If the minimizers $\varphi(\varepsilon)$ of the total rescaled energy $I(\varepsilon)$ over $\Psi(\varepsilon)$ admit an asymptotic expansion as in* (6)*, and if their total energy $I(\varepsilon)(\varphi(\varepsilon))$ remains bounded with $u_\gamma = 0$ on γ, then*

$$\varphi_0 = \arg\min_{\psi_0 \in \Psi_0} I_0(\psi_0),$$

where

$$I_0(\psi_0) := \frac{1}{3}\int_{S'} Q^0_{D'\psi_0}(D'n, 1)\, dx' + 2\int_{S'} W_0(D'\psi_0, n_0(D'\psi_0))\, dx' - 2\int_{S'} f_0 \cdot \psi_0\, dx', \tag{19}$$

$f_0(x') = f(x', 0)$ for every $x' \in S'$, and

$$\Psi_0 := \{\psi_0 \in W^{1,\infty}(S')^3 : n = n(D'\psi_0) \in W^{1,\infty}(S')^3,\ D'\psi_0 \in \mathcal{M}',\ \psi_0(x') = \phi_0(x'),\ \text{and } n = n_\gamma(x') \text{ for every } x' \in \gamma\}.$$

4. Invariance and flexural energy

Under several assumptions on the stored energy function W_2, the expression of the flexural part

$$I_{\text{flex}}(\psi_0) := \frac{1}{3}\int_{S'} Q^0_{D'\psi_0}(D'n, 1)\, dx \tag{20}$$

of the total limit energy $I_0(\psi_0)$ may be reduced. More precisely, we shall consider the implications of homogeneity along the fibers, frame-indifference (left invariance under SO(3)), planar isotropy (right invariance under in-plane rotations), and finally right invariance of the stored energy under the special linear group of TS'.

4.1. ***Homogeneity along the fibers.*** We say that the shell is homogeneous along the fibers if, for every

$$(F', x_3, v) \in T^*(S; \mathbb{R}^3) = T^*(S'; \mathbb{R}^3) \times (-1, 1) \times (\mathbb{R}')^3,$$

we have $W_2(F', x_3, v) = W_2(F', 0, v)$. In this case, for every

$$(G', s, v) \in T^*(S'; \mathbb{R}^3) \times \mathbb{R} \times (\mathbb{R}')^3$$

and every section F' of $\mathcal{M}'$, we have $D^2 W_2[G'_{F'}, s_0, v_{n(F')}]^2 = D^2 W_2[G'_{F'}, 0, v_{n(F')}]^2$. It follows that $Q^0_{F'}(G', s)$ is independent of s, and hence we denote it by $Q^0_{F'}(G')$, so that

$$I_{\text{flex}}(\psi_0) = \frac{1}{3} \int_{S'} Q^0_{D'\psi_0}(D'n)\, dx.$$

4.2. ***Frame-indifference.*** The principle of frame-indifference states that the space is invariant under rotation, which translates in our case to the following condition on the stored energy function W^ε:

$$W^\varepsilon(F) = W^\varepsilon(RF) \quad \text{for every rotation } R \in \mathrm{SO}(3).$$

This is assumed in the sequel. Accordingly, W_2 satisfies the same property, i.e., $W_2(F) = W_2(RF)$ for every $R \in \mathrm{SO}(3)$.

In the following, we denote by $\mathscr{E}_{S'}$ the set of symmetric bilinear forms on TS', that is, the fiber bundle with base space S' whose fiber $(\mathscr{E}_{S'})_{x'}$ at $x' \in S'$ is the set of symmetric bilinear forms on $T_{x'}S'$. The fiber bundle $\mathscr{E}_S$ is defined in a similar way, and $\mathscr{E}'_S$ stands for its restriction to S'. Moreover, if $F \in T^*_x(S; \mathbb{R}^3)$, $F^T F$ stands for the element of $(\mathscr{E}_{S'})_x$ that maps every element (u, v) of $(T_x S)^2$ to the scalar product between Fu and Fv. A similar notation is used to define $(F')^T F' \in \mathscr{E}_{S'}$ for every $F' \in T^*(S'; \mathbb{R}^3)$.

Lemma 6. *If the stored energy W_2 is frame-indifferent, then for every $F' \in \mathcal{M}'$ of maximum rank and every $R \in \mathrm{SO}(3)$, we have*

$$RF' \in \mathcal{M}' \quad \textit{and} \quad n(RF') = Rn(F').$$

Moreover, there exists a bundle map $\tau' : \mathcal{M}' \to TS'$ and a map $\tau_3 : \mathcal{M}' \to \mathbb{R}$ such that, for every $F' \in \mathcal{M}'$ of maximal rank,

$$n(F') = F'\tau'(F') + n_{F'}\tau_3(F'),$$

where both $\tau'(F')$ and $\tau_3(F')$ depend only on $C' = (F')^T F'$, and $n_{F'} \in \mathbb{R}^3$ is defined by

$$n_{F'} \cdot w = \det(F', w) \quad \textit{for every } w \in \mathbb{R}^3. \tag{21}$$

Lastly, $C = (F', n_0(F'))^T (F', n_0(F'))$ depends only on C'.

Proof. The first part of the proposition is obvious. Next, since F' is of maximum rank, $(F', n_{F'})$ is invertible, so we can set $(\tau'(F'), \tau_3(F')) = (F', n_{F'})^{-1} n(F')$.

Moreover, we can check that

$$\begin{aligned}(\tau'(RF'), \tau_3(RF')) &= (RF', n_{RF'})^{-1} n(RF) \\ &= (RF', R(n_{F'}))^{-1} Rn(F) = (R(F', n_{F'}))^{-1} Rn(F) \\ &= (F', n_{F'})^{-1} n(F') = (\tau'(F'), \tau_3(F')),\end{aligned}$$

whence both τ' and τ_3 only depend on $(F')^T F'$. Finally, it is readily verified that both $n(F')^T n(F')$ and $n(F')^T F'$ are invariant under rotations of F'. As a result, $C = (F', n_0(F'))^T (F', n_0(F'))$ depends only on $(F')^T F'$ as well. □

Since $T_{(x',x_3)}S = T_{x'}S' \times T_{x_3}(-1, 1) = T_{x'}S' \times \mathbb{R}$, every element $C \in (\mathscr{C}_S)_x$ can be decomposed uniquely into $(C', C_3, C_{33}) \in (\mathscr{C}_{S'})_{x'} \times T^*_{x'}S' \times \mathbb{R}$ such that, for every (u', u_3) and every (v', v_3) in $T_{x'}S' \times \mathbb{R} = T_x S$,

$$C((u', u_3), (v', v_3)) = C'(u', v') + u_3 C_3(v') + v_3 C_3(u') + C_{33} u_3 v_3.$$

In addition, we write this decomposition as

$$C = \begin{pmatrix} C' & C_3^T \\ C_3 & C_{33} \end{pmatrix}.$$

Let us introduce the fiber bundle $\mathscr{P}$ with base space S' whose fiber at $x' \in S'$ is the set of polynomials of degree less than or equal to two on $(\mathscr{C}_{S'})_{x'}$.

Proposition 7. *If the stored energy function W_2 is frame-indifferent, then there exists a bundle map $P : C' \mapsto P_{C'}$ over S' from $\mathscr{C}_{S'}$ into $\mathscr{P}$ such that, for every deformation ψ_0 of finite limit energy $I_0(\psi_0)$ and every $G' \in T^*(S'; \mathbb{R}^3)$, we have*

$$Q^0_{D'\psi_0}(G', 1) = P_{C'}(D'\psi_0{}^T G' + G'^T D'\psi_0),$$

where $C' = D'\psi_0{}^T D'\psi_0$ for short. Moreover, if W_2 is homogeneous along the fibers, then $P_{C'}$ is homogeneous of degree two.

Proof. Let $\mathscr{M}^+ = \{F \in \mathscr{M} : \det F > 0\}$. Since W_2 is assumed to be frame-indifferent, there exists a map $W_S : \mathscr{C}_S \to \mathbb{R}$ such that, for every F in a neighborhood of $\mathscr{M}^+$,

$$W_2(F) = W_S \circ m(F), \tag{22}$$

where $m : T^*(S; \mathbb{R}^3) \to \mathscr{C}_S$ is the bundle map defined by $m(G) = G^T G$. Let F' be a section of $\mathscr{M}'$, $s \in \mathbb{R}$ and $G' \in T^*(S'; \mathbb{R}^3)$ such that $(F', n_0(F')) \in \mathscr{M}^+$ a.e. Then definition (10) combined with (22) gives

$$Q^0_{F'}(G', s) = \inf_{v \in \mathbb{R}^3} D^2 W_2[G'_{F'}, s_0, v_{n(F')}]^2 = \inf_{v \in \mathbb{R}^3} D^2 W_S[Dm(G'_{F'}, s_0, v_{n(F')})]^2.$$

Since $\mathscr{E}_S = (-1, 1) \times \mathscr{E}'_S$, we can identify $T\mathscr{E}_S$ with $T(-1, 1) \times T\mathscr{E}'_S$. Doing so, we obtain that

$$Dm(G'_{F'}, s_0, v_{n(F')}) = \frac{d}{dt}(ts, C + tE)(t = 0) = (s_0, E_C),$$

where

$$E = \begin{pmatrix} (F')^T G' + (G')^T F' & (F')^T v + (G')^T n \\ v^T F' + n^T G' & n^T v + v^T n \end{pmatrix},$$

$n(F')$ is denoted by n for short, and $C = m(F)$. Since $(F', n(F'))(x')$ is assumed to be invertible for every $x' \in S'$, setting $w' = {F'}^T v + {G'}^T n$ and $w_3 = n^T v + v^T n$, we get

$$Q^0_{F'}(G', 1) = P_{C'}((F')^T G' + (G')^T F'), \tag{23}$$

where $P_{C'}$ is the section of $\mathscr{P}$ defined by

$$P_{C'}(M) = \inf_{w \in \mathbb{R}^3} D^2 W_S \left[1_0, \begin{pmatrix} M & w' \\ (w')^T & w_3 \end{pmatrix}_C \right]^2 \tag{24}$$

for every $M \in \mathscr{E}_{S'}$, and $C = (F', n_0(F'))^T (F', n_0(F'))$, which depends only on C', according to Lemma 6. Finally, if ψ_0 is a deformation satisfying $I_0(\psi_0) < \infty$, owing to the noninterpenetration assumptions made, we know that $D\psi_0 \in \mathscr{M}^+$ a.e. As a result, $Q^0_{D'\psi_0}(D'n, 1) = P_{C'}(M)$ a.e. on S', with $M = F^T D'n + D'n^T F'$, as claimed. Moreover, if the shell is homogeneous along its fibers, (24) reduces to

$$P_{C'}(M) = \inf_{w \in \mathbb{R}^3} D^2 W_S \left[0, \begin{pmatrix} M & w' \\ (w')^T & w_3 \end{pmatrix}_C \right]^2, \tag{25}$$

which is homogeneous of degree two with respect to M. □

Remark. Note that $D'\psi_0{}^T D'n + D'n^T D'\psi_0$ is not, in general, the second fundamental form of the deformation, except in the case where $n(D'\psi_0)$ is the normal vector to $D'\psi_0$.

4.3. ***Planar isotropy.*** We say that the material is isotropic along the midsection of the shell if, for every planar rotation R in the set $\mathrm{SO}(T_{x'}S')$ of rotations of $T_{x'}S'$ and every $F = (F', F_3) \in T^*_{x'}(S^\varepsilon; \mathbb{R}^3)$, we have

$$W^\varepsilon(F', F_3) = W^\varepsilon(F'R, F_3). \tag{26}$$

As a consequence, for deformations of finite energy, the fibers of the shell remain normal to its section.

Lemma 8. *Assume that the shell is isotropic along its midsection. Then there exists a map $\tau_3 : \mathscr{M}' \to \mathbb{R}$ such that*

$$n(F') = n_{F'} \tau_3(F')$$

for every $F' \in \mathcal{M}'$ of maximal rank, where $n_{F'}$ is defined by (21). *Moreover, $\tau_3(F')$ depends only on the metric $C' = F'^T F'$.*

Proof. Let F' be an element of $\mathcal{M}'$ of maximal rank. By definition, we have

$$n_0(F') = \arg \min_{F_3 \in T_0^*((-1,1);\mathbb{R}^3)} W_2(F', F_3).$$

For every rotation $R \in \mathrm{SO}(TS')$, the isotropy property yields

$$\begin{aligned} n_0(F'R) &= \arg \min_{F_3 \in T_0^*((-1,1);\mathbb{R}^3)} W_2(F'R, F_3) \\ &= \arg \min_{F_3 \in T_0^*((-1,1);\mathbb{R}^3)} W_2(F', F_3) = n_0(F'). \end{aligned}$$

In particular, this entails that $n(-F') = n(F')$. What is more, by frame-indifference, we have from Lemma 6 that

$$n(F') = F'\tau'(F') + n_{F'}\tau_3(F'),$$

where τ' is a bundle map from $\mathcal{M}'$ into TS' and τ_3 is a map from $\mathcal{M}'$ into $\mathbb{R}$, both of them depending only on the metric $C' = F'^T F'$. Thus,

$$\begin{aligned} F'\tau'(F') + n_{F'}\tau_3(F') = n(F') = n(-F') &= -F'\tau'(-F') + n_{-F'}\tau_3(-F') \\ &= -F'\tau'(F') + n_{F'}\tau_3(F'). \end{aligned}$$

Consequently, $F'\tau'(F') = 0$ and $n(F') = n_{F'}\tau_3(F')$, as claimed. □

Proposition 9. *Assume that the shell is isotropic along its midsection. Then the flexural energy $I_{\mathrm{flex}}(\psi_0)$ depends only on the metric and the second fundamental form of the deformed surface. Namely, we have*

$$I_{\mathrm{flex}}(\psi_0) = \frac{1}{3} \int_{S'} P_{C'}(|n(D'\psi_0)|b_{D'\psi_0})\, dx',$$

where $b_{D'\psi_0}$ is the second fundamental form of ψ_0, i.e.,

$$b_{F'} = D'N^T F' + F'^T D'N, \tag{27}$$

where $N = n_{F'}/|n_{F'}|$, and $P_{C'}$ is defined by Proposition 7.

Proof. Since $n(D'\psi_0)$ (denoted n for short) is normal collinear to $n_{D'\psi_0}$ and thus to the normal N to the deformed surface, we get

$$D'n^T D'\psi_0 + D'\psi_0{}^T D'n = (n \cdot N)(D'N^T D'\psi_0 + D'\psi_0{}^T D'N). \qquad \square$$

4.4. ***Right invariance under the special linear group.*** We denote by $\mathrm{SL}(TS')$ the special linear group over TS', that is the fiber bundle over S' whose fiber at x' is the linear diffeomorphisms of $T_{x'}S'$ with determinant equal to one. In this section, we consider the case where the energy W_2 is right invariant under the special linear group, that is, $W_2(F', F_3) = W_2(F'U, F_3)$ for every $x' \in S'$, $U \in \mathrm{SL}(T_{x'}S')$ and $(F', F_3) \in T^*_{x'}(S'; \mathbb{R}^3) \times T^*((-1, 1); \mathbb{R}^3)$.

Proposition 10. *Assume that W_2 is right invariant under* $\mathrm{SL}(TS')$. *Then the flexural energy $I_{\mathrm{flex}}(\psi_0)$ depends only on the metric and on the mean curvature*

$$H = \mathrm{Tr}(C'^{-1/2} b_{D'\psi} C'^{-1/2})$$

of the deformation. More precisely, we have

$$I_{\mathrm{flex}}(\psi_0) = \frac{1}{3} \int_{S'} K_{x', \det(C')}(H)\, dx', \tag{28}$$

where $K : S' \times \mathbb{R} \to \Pi_2$; here Π_2 is the set of polynomials of degree at most two. Moreover, if the shell is homogeneous along its fibers, then

$$I_{\mathrm{flex}}(\psi_0) = \frac{1}{3} \int_{S'} \kappa_{x', \det(C')} |H|^2\, dx', \tag{29}$$

where κ is a map from $S' \times \mathbb{R}^+$ into $\mathbb{R}^+$.

Proof. Let $\mathbb{O}$ be the fiber bundle over S' whose fibers are the maps from $T_{x'}S'$ into itself of zero trace. For every $O \in \mathbb{O}$ and $x' = \pi_{S'}(O)$, there exists a regular map $U : (0, 1) \to \mathrm{SL}(T_{x'}S')$ such that $\dot{U}(0) = O$ and $U(0) = \mathrm{Id}$. Let F be a section of $\mathcal{M}$, $(G', s, v) \in T^*_{x'}(S'; \mathbb{R}^3) \times \mathbb{R} \times (\mathbb{R}')^3$, and let $\gamma(t) = (\gamma'(t), \gamma_3(t))$ be a curve in $T^*(S; \mathbb{R}^3)$ such that $\dot{\gamma}(0) = (G'_{F'}, s_0, v_{n(F')})$, as in (11). From the right invariance under the special linear group, for every $U_0 \in \mathrm{SL}(T_{x'}S')$, we have

$$W_2(\gamma(t)) = W_2(\gamma'(t)U_0U(t), \gamma_3(t)).$$

As a consequence,

$$D^2W_2[\dot{\gamma}(0)]^2 = D^2W_2\left[\frac{d}{dt}(\gamma' U_0 U, \gamma_3)|_{t=0}\right]^2. \tag{30}$$

Then a simple computation yields

$$\frac{d}{dt}(\gamma' U_0 U, \gamma_3)|_{t=0} = ((G' U_0 + F' U_0 O)_{F' U_0}, s_0, v_{n(F')}),$$

which, owing to (30) and (10), leads to $Q^0_{F'}(G', s) = Q^0_{F'U_0}(G'U_0 + F'U_0O, s)$. From (23), recalling that $C' = F'^T F'$, we get

$$\begin{aligned}
P_{C'}(F'^T G' + G'^T F') &= Q^0_{F'}(G', 1) = Q^0_{F'U_0}(G'U_0 + F'U_0O, 1)\\
&= P_{U_0^T F'^T F' U_0}((F'U_0)^T(G'U_0 + F'U_0O) + (G'U_0 + F'U_0O)^T(F'U_0))\\
&= P_{U_0^T C' U_0}(U_0^T(F'^T G' + G'^T F')U_0 + U_0^T C' U_0 O + O^T U_0^T C' U_0)\\
&= P_{C'_0}\big((C'_0)^{1/2}[C_0'^{-1/2}U_0^T(F'^T G' + G'^T F')U_0 C_0'^{-1/2}\\
&\qquad + C_0'^{1/2} O C_0'^{-1/2} + C_0'^{-1/2} O^T C_0'^{1/2}]C_0'^{1/2}\big),
\end{aligned}$$

where $C' = F'^T F'$ and $C'_0 = U_0^T C' U_0$. Since the map

$$O \mapsto C_0'^{1/2} O C_0'^{-1/2} + C_0'^{-1/2} O^T C_0'^{1/2}$$

is a diffeomorphism over the set of symmetric trace-free matrices, the above expression leads to

$$P_{C'}(F'^T G' + G'^T F') = P_{C'_0}\big(\tfrac{1}{2}\operatorname{Tr}\big(C_0'^{-1/2}U_0^T(F'^T G' + G'^T F')U_0 C_0'^{-1/2}\big)C'_0\big).$$

In addition,

$$\operatorname{Tr}\big(C_0'^{-1/2}U_0^T(F'^T G' + G'^T F')U_0 C_0'^{-1/2}\big) = \operatorname{Tr}\big(C'^{-1/2}(F'^T G' + G'^T F')C'^{-1/2}\big),$$

so that we may write

$$P_{C'}(F'^T G' + G'^T F') = P_{C'_0}\big(\tfrac{1}{2}\operatorname{Tr}\big(C'^{-1/2}(F'^T G' + G'^T F')C'^{-1/2}\big)C'_0\big).$$

Since C' is symmetric and nonnegative, there exists a rotation $R \in \mathrm{SO}(T_{x'}S')$ and nonnegative reals λ_1, λ_2 such that

$$C' = R^T \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} R.$$

Let us choose $U_0 \in \mathrm{SL}(T_{x'}S')$ such that

$$U_0 = (\det C')^{1/4} R^T \begin{pmatrix} \lambda_1^{-1/2} & 0 \\ 0 & \lambda_2^{-1/2} \end{pmatrix}.$$

Hence, $C'_0 = U_0^T C' U_0 = (\det C')^{1/2}\,\mathrm{Id}$, so that

$$\begin{aligned}
&P_{C'}(F'^T G' + G'^T F') =\\
&\qquad P_{(\det C')^{1/2}\,\mathrm{Id}}\big(\tfrac{1}{2}\operatorname{Tr}\big(C'^{-1/2}(F'^T G' + G'^T F')C'^{-1/2}\big)(\det C')^{1/2}\,\mathrm{Id}\big). \quad (31)
\end{aligned}$$

Using the definition of I_{flex} and $D'\psi_0{}^T D'n + D'n^T D'\psi_0 = |n(D'\psi_0)| b_{D'\psi_0}$, we infer that

$$I_{\text{flex}}(\psi_0) = \frac{1}{2}\int_{S'} P_{(\det C')^{1/2}\,\mathrm{Id}}\big(\tfrac{1}{2}|n(D'\psi_0)|\,\mathrm{Tr}(C'^{-1/2} b_{D'\psi} C^{-1/2})(\det C')^{1/2}\,\mathrm{Id}\big)\,dx'$$
$$= \frac{1}{2}\int_{S'} P_{(\det C')^{1/2}\,\mathrm{Id}}\big(\tfrac{1}{2}|n(D'\psi_0)| H (\det C')^{1/2}\mathrm{Id}\big)\,dx'.$$

Finally, the right invariance of W_2 with respect to $\mathrm{SL}(TS')$ implies that $|n(F')|$ depends only on $\det C'$. Setting

$$K_{x',\det C'}(H) = P_{(\det C')^{1/2}\,\mathrm{Id}}\big(\tfrac{1}{2}|n(F')| H (\det C')^{1/2}\,\mathrm{Id}\big),$$

we get (28). Moreover, if the shell is homogeneous along its fibers, then $P_{(\det C')^{1/2}\,\mathrm{Id}}$ is homogeneous of degree two and, accordingly, $K_{x',\det C'}(H)$ is a monomial. □

5. Geometric configuration

Classically, the energy of an elastic body is not written in terms of the deformation ψ^ε of S^ε, but in terms of the deformation $\tilde{\psi}^\varepsilon$ of the *geometric* configuration $\widetilde{S}^\varepsilon := g_\varepsilon(S^\varepsilon)$, where

$$g_\varepsilon : S' \times (-\varepsilon, \varepsilon) \to \mathbb{R}^3$$
$$(x', x_3) \mapsto x' + x_3 n'(x'),$$

$n' : S' \to \mathbb{R}^3$ being the normal to S'. We set $\widetilde{S} := \widetilde{S}^{\varepsilon_0}$, for a small enough ε_0 such that g_{ε_0} is one-to-one. In the following, we will always assume that $\varepsilon \le \varepsilon_0$. We intend to recast our results in this geometric configuration. This is easily achieved by a mere change of variables. To begin with, we have to recast our initial three-dimensional problem in the geometric configuration.

5.1. ***Recast of the problem.*** We denote by $\tilde{J}_\varepsilon(\tilde{\psi}^\varepsilon)$ the elastic energy of a deformation $\tilde{\psi}^\varepsilon$ of $\widetilde{S}^\varepsilon$, and assume that it has the form

$$\tilde{J}_\varepsilon(\tilde{\psi}^\varepsilon) := \int_{\widetilde{S}^\varepsilon} \widetilde{W}^\varepsilon(\tilde{x}, \nabla\tilde{\psi}^\varepsilon(\tilde{x}))\,d\tilde{x},$$

where $\widetilde{W}^\varepsilon$ stands for the stored energy function of the solid. Furthermore, we assume the shell to be subjected to dead-body loads $\tilde{f}_\varepsilon$, so that the total energy of the system is given by

$$\tilde{I}_\varepsilon(\tilde{\psi}^\varepsilon) := \tilde{J}_\varepsilon(\tilde{\psi}^\varepsilon) - \int_{\widetilde{S}^\varepsilon} \tilde{f}_\varepsilon \cdot \tilde{\psi}^\varepsilon\,d\tilde{x}. \tag{32}$$

Finally, clamping boundary conditions are added on a part of the boundary: $g_\varepsilon(\Gamma^\varepsilon) = g_\varepsilon(\gamma \times (-\varepsilon, \varepsilon))$, where $\gamma \subset \partial S'$.

Our aim is to determine the behavior of the minimizers $\tilde{\varphi}^\varepsilon$ of $\tilde{I}_\varepsilon$ over

$$\widetilde{\Psi}^\varepsilon := \{\tilde{\psi}^\varepsilon \in W^{1,\infty}(\widetilde{S}^\varepsilon)^3 : \tilde{\psi}^\varepsilon \circ g_\varepsilon(x^\varepsilon) = \phi^\varepsilon(x^\varepsilon),\ x^\varepsilon \in \Gamma^\varepsilon\}$$

as ε goes to zero, under the assumptions on the stored energy and the applied loads made hereunder.

In order to apply our results, several assumptions, similar to the ones we made on W^ε and φ^ε, have to be imposed on $\widetilde{W}^\varepsilon$ and on the minimization sequences $\tilde{\varphi}^\varepsilon$.

Dependence of the stored energy on the thickness. We assume that there exist continuous nonnegative maps $\widetilde{W}_2$ and $\widetilde{W}_0$ such that, for every $(\tilde{x}, F) \in \widetilde{S} \times \mathbb{R}^{3\times 3}$, we have

$$\widetilde{W}^\varepsilon(\tilde{x}, F) = \varepsilon^{-1}\big(\varepsilon^{-2}\widetilde{W}_2(\tilde{x}, F) + \widetilde{W}_0(\tilde{x}, F)\big).$$

Behavior of strongly extended fibers. We assume that the stored energy $\widetilde{W}_2$ is bounded from below by a positive constant for strongly extended fibers, that is, there exist $\delta, c > 0$ such that

$$\widetilde{W}_2\big(\tilde{x}, F' \circ \pi'_{x'} + F_3 \otimes n'(x')\big) \geq c \quad \text{for all } \tilde{x} \in \widetilde{S},\ F' \in T^*_{x'}(S'; \mathbb{R}^3),\ F_3 \in \mathbb{R}^3 \text{ such that } |F_3| \geq \delta, \quad (33)$$

where x' is the projection of $\tilde{x}$ onto S' and $\pi'_{x'}$ is the projection of $\mathbb{R}^3$ onto $T_{x'}(S'; \mathbb{R}^3)$.

Regularity and zero set of $\widetilde{W}_2$. We assume that $\widetilde{W}_2$ is a nonnegative C^2 function, and denote by $\widetilde{\mathcal{M}}$ the restriction of its zero set to $S' \times \mathbb{R}^{3\times 3}$, that is,

$$\widetilde{\mathcal{M}} := \{(x', F) \in S' \times \mathbb{R}^{3\times 3} : \widetilde{W}_2(x', F) = 0\}.$$

Let $\mathcal{M}'$ be the projection of $\widetilde{\mathcal{M}}$ onto $T^*(S'; \mathbb{R}^3)$, that is,

$$\mathcal{M}' := \bigcup_{x' \in S'} \{F' \in T^*_{x'}(S'; \mathbb{R}^3) : \text{there exists } n \in \mathbb{R}^3 \text{ with } (x', F' \circ \pi'_{x'} + n \otimes n'(x')) \in \widetilde{\mathcal{M}}\}.$$

Once again, we assume that the projection of $\widetilde{\mathcal{M}}$ onto $\mathcal{M}'$ is one-to-one, that is, there exists a map $n : \mathcal{M}' \to \mathbb{R}^3$ such that

$$\widetilde{\mathcal{M}} = \{(x', F' \circ \pi'_{x'} + n(F') \otimes n'(x')) \in S' \times \mathbb{R}^{3\times 3} : F' \in \mathcal{M}'\}. \quad (34)$$

Interpenetration. To avoid interpenetration of matter, it is geometric to expect $D\tilde{\psi}^\varepsilon$ to be invertible. To this end, we require that $\widetilde{W}_0(x', F) = \infty$ for every $(x', F) \in \widetilde{\mathcal{M}}$ such that $\det F < 0$, and that

$$\widetilde{W}_0(x', F) \to \infty \quad \text{if } (x', F) \in \widetilde{\mathcal{M}} \text{ and } \det F \to 0.$$

Applied loads. The volumic loads are assumed to scale as the inverse of the thickness of the shell; more precisely, we assume that there exists $\tilde{f} : \widetilde{S} \to \mathbb{R}^3$ such that

$$\tilde{f}_\varepsilon(\tilde{x}) = \varepsilon^{-1}\tilde{f}(\tilde{x}) \quad \text{for every } \tilde{x} \in \widetilde{S}.$$

Ansatz. We assume that the minimizers of the energy admit an asymptotic expansion in the form

$$\tilde{\varphi}^\varepsilon(x' + \varepsilon x_3 n') = \sum_{k\geq 0} \varepsilon^k \varphi_k(x', x_3). \tag{35}$$

5.2. ***Change of variable.*** In order to apply our result, we first have to rewrite the energy in terms of the associated deformation $\psi^\varepsilon = \tilde{\psi}^\varepsilon \circ g_\varepsilon$ of S^ε. We have

$$\tilde{J}(\tilde{\psi}^\varepsilon) = J(\psi^\varepsilon) = \int_{S^\varepsilon} W^\varepsilon(D\psi^\varepsilon)\, dx, \tag{36}$$

with $W^\varepsilon = \varepsilon^{-1}(\varepsilon^{-2}W_2 + W_0)$, and for every $F \in T^*_x(S; \mathbb{R}^3)$,

$$W_k(F) = \widetilde{W}_k(F \circ (Dg_\varepsilon(x))^{-1}) \det(Dg_\varepsilon(x)), \quad k = 0, 2. \tag{37}$$

Note that W_2 and W_0 are independent of ε since $Dg_\varepsilon = (\mathrm{Id}', n') + x_3(D'n', 0)$ (which is denoted by Dg hereafter). In addition, these energies satisfy the assumptions made in Section 2.1. Finally, the minimizers $\varphi^\varepsilon = \tilde{\varphi}^\varepsilon \circ g_\varepsilon$ admit the same asymptotic expansion as $\tilde{\varphi}^\varepsilon$. Thus, all of the results of Sections 3 and 4 apply and may be expressed in terms of $\widetilde{W}_0$ and $\widetilde{W}_2$ up to a change of variable. Moreover, the definitions of $\mathcal{M}'$ and of the map $n : \mathcal{M}' \to \mathbb{R}^3$ are independent of the chosen approach.

Lemma 11. *If the function W_2 is defined by* (37)*, and F' is a section of $\mathcal{M}'$, then, for every $G' \in T^*_{x'}(S'; \mathbb{R}^3)$ and $s \in \mathbb{R}$, we have*

$$Q^0_{F'}(G', s) = \widetilde{Q}^0_{F'}(G' - sF'D'n', s),$$

where

$$\widetilde{Q}^0_{F'}(G', s) := \inf_{v\in\mathbb{R}^3} D^2\widetilde{W}_2(x', F)[sn', G'\pi'_{x'} + v \otimes n']^2,$$

where $\pi'_{x'}$ is the projection of $\mathbb{R}^3$ onto $T_{x'}S'$ and $F = F'(x') \circ \pi'_{x'} + n(F'(x')) \otimes n'(x')$.

Proof. Let F' be a section of $\mathcal{M}'$, x' be an element of S' and $G' \in T^*_{x'}(S'; \mathbb{R}^3)$. From the definition of $Q^0_{F'}$, we have

$$Q^0_{F'}(G', s) = \inf_{v\in\mathbb{R}^3} D^2 W_2[G'_{F'}, s_0, v_{n(F')}]^2,$$

where $(G'_{F'}, s_0, v_{n(F')}) = \dot{\gamma}(0)$, and

$$\gamma(t) = (F'(x') + tG', ts, n(F'(x')) + vt)$$

is an element of

$$T^*(S'; \mathbb{R}^3) \times (-1, 1) \times (\mathbb{R}')^3 = T^*(S'; \mathbb{R}^3) \times T^*((-1, 1); \mathbb{R}^3) = T^*(S; \mathbb{R}^3).$$

From (37), we deduce

$$Q^0_{F'}(G', s) = \det(Dg(x', 0)) \inf_{v\in\mathbb{R}^3} D^2\widetilde{W}_2[\dot{\tilde{\gamma}}(0)]^2 = \inf_{v\in\mathbb{R}^3} D^2\widetilde{W}_2[\dot{\tilde{\gamma}}(0)]^2, \qquad (38)$$

where $\tilde{\gamma}(t) = \gamma(t) \circ Dg(x', ts)^{-1}$. On the other hand, $Dg = (\mathrm{Id}', n') + x_3(D'n', 0)$, and hence

$$\begin{aligned}(Dg)^{-1} &= ((\mathrm{Id}', n') + x_3(D'n', 0))^{-1}\\ &= \big((\mathrm{Id}', n')(\mathrm{Id} + x_3(\mathrm{Id}', n')^{-1}(D'n', 0))\big)^{-1}\\ &= \big(\mathrm{Id} - x_3(\mathrm{Id}', n')^{-1}(D'n', 0)\big)(\mathrm{Id}', n')^{-1} + o(x_3).\end{aligned}$$

Since $(\mathrm{Id}' + n' \otimes e_3)^{-1} = \begin{pmatrix} \pi' \\ n'^T \end{pmatrix}$, the above identity reads

$$\begin{aligned}(Dg)^{-1} &= \begin{pmatrix} \pi' \\ n'^T \end{pmatrix} - x_3 \begin{pmatrix} \pi' \\ n'^T \end{pmatrix} (D'n', 0) \begin{pmatrix} \pi' \\ n'^T \end{pmatrix} + o(x_3)\\ &= \begin{pmatrix} \pi' \\ n'^T \end{pmatrix} - x_3 \begin{pmatrix} \pi' D'n' \pi' \\ 0 \end{pmatrix} + o(x_3).\end{aligned}$$

It follows that $\tilde{\gamma}(t) = (x(t), F(t)) + o(t)$ with $x(t) = x' + tsn'$, and, using the notation F' for $F'(x')$ for short,

$$\begin{aligned}F(t) &= ((F', n(F')) + t(G', v)) \left(\begin{pmatrix} \pi' \\ n'^T \end{pmatrix} - ts \begin{pmatrix} \pi' D'n' \pi' \\ 0 \end{pmatrix} \right)\\ &= (F', n(F')) \begin{pmatrix} \pi' \\ n'^T \end{pmatrix} + t(G'\pi' + v \otimes n') - ts F' D'n' \pi'.\end{aligned}$$

Consequently,

$$\begin{aligned}\dot{\tilde{\gamma}}(0) &= \left[\left(x', (F', n(F')) \begin{pmatrix} \pi' \\ n'^T \end{pmatrix} \right), (sn', G'\pi' + v \otimes n' - s(F'D'n'\pi')) \right]\\ &= \big[(x', F'\pi' + n(F') \otimes n'), (sn', (G' - sF'D'n')\pi' + v \otimes n') \big]\\ &= \big[(x', F), (sn', (G' - sF'D'n')\pi' + v \otimes n') \big]\end{aligned}$$

The conclusion follows from (38). □

From now on, we limit our analysis to the case where standard boundary conditions (18) are applied. From Proposition 5, we immediately infer the next result.

Proposition 12. *Assume that the standard boundary conditions* (18) *apply to the shell. Let $\tilde{\varphi}^\varepsilon$ be the minimizer of the total energy $\tilde{I}_\varepsilon(\tilde{\varphi}^\varepsilon)$ over the space of admissible deformations. If $\tilde{\varphi}^\varepsilon$ admits an asymptotic expansion as in* (35), *and if the total energy $\tilde{I}_\varepsilon(\tilde{\varphi}^\varepsilon)$ remains bounded, then*

$$\varphi_0 = \arg\min_{\psi_0\in\Psi_0} \tilde{I}_0(\psi_0),$$

where

$$\tilde{I}_0(\psi_0) = \frac{1}{3}\int_{S'} \widetilde{Q}^0_{D'\psi_0}(D'n - D'\psi_0 D'n', 1)\,dx' + 2\int_{S'} \widetilde{W}_0(x', (D'\psi_0, n))\,dx' - 2\int_{S'} \tilde{f}_0 \cdot \psi_0\,dx',$$

$\tilde{f}_0(x') = \tilde{f}(x', 0)$ *for every* $x' \in S'$, $n = n(D'\psi_0)$, *and*

$$\Psi_0 := \{\psi_0 \in W^{1,\infty}(S')^3 : D'\psi_0 \in \mathcal{M}',\ n = n(D'\psi_0) \in W^{1,\infty}(S')^3 \text{ such that } \psi_0(x') = \phi_0(x') \text{ and } n(x') = n_\gamma(x') \text{ for every } x' \in \gamma\}.$$

Note that more general Dirichlet conditions could have been considered in the same fashion as in Lemma 4.

5.3. ***Homogeneity along the fibers.*** We say that the shell is homogeneous along its fibers in the geometric configuration if, for every $x' \in S'$, $s \in (-1, 1)$ and $F \in \mathbb{R}^{3\times 3}$, we have $\widetilde{W}_2(x' + sn', F) = \widetilde{W}_2(x', F)$.

Proposition 13. *If the shell is homogeneous along its fibers in the geometric configuration, then* $\widetilde{Q}^0_{F'}(G', s)$ *is independent of* s, *and is denoted by* $\widetilde{Q}^0_{F'}(G')$.

5.4. ***Frame-indifference.*** In the following, we assume the stored energy to be frame-indifferent, that is, $\widetilde{W}^\varepsilon(\tilde{x}, RF) = \widetilde{W}^\varepsilon(\tilde{x}, F)$ for every $(\tilde{x}, F) \in \widetilde{S}^\varepsilon \times \mathbb{R}^{3\times 3}$ (with $\varepsilon > 0$ small enough), and every rotation $R \in \mathrm{SO}(3)$. Note that this is equivalent to the frame-indifference of W^ε.

Proposition 14. *If the stored energy function* $\widetilde{W}_2$ *is frame-indifferent, then there exists a bundle map* $\widetilde{P} : C' \mapsto \widetilde{P}_{C'}$ *over* S' *from* $\mathscr{E}_{S'}$ *into* $\mathscr{P}$ *such that, for every deformation* ψ_0 *of finite energy* $\tilde{I}_0(\psi_0)$ *and for every* $G' \in T^*(S'; \mathbb{R}^3)$, *we have*

$$\widetilde{Q}^0_{D'\psi_0}(G', 1) = \widetilde{P}_{C'}(D'\psi_0{}^T G' + G'^T D'\psi_0),$$

with $C' = D'\psi_0{}^T D'\psi_0$ *and* $n = n(D'\psi_0)$. *Moreover, if the shell is homogeneous along its fibers in the geometric configuration, then* $\widetilde{P}_{C'}$ *is homogeneous of degree two.*

Proof. The proof is similar to the one devised for the abstract configuration. Once again, there exists a map $\widetilde{W}_S$ such that, at least in a neighborhood of $\widetilde{\mathcal{M}}^+ = \{F \in \widetilde{\mathcal{M}} : \det F > 0\}$, we may write $\widetilde{W}_2(x, F) = \widetilde{W}_S(x, F^T F)$. After some computations, we derive the claimed result with

$$\widetilde{P}_{C'}(M) = \inf_{\substack{w' \in T_{x'}S' \\ w_3 \in \mathbb{R}}} D^2\widetilde{W}_S(x', \widetilde{C})\big[n', \pi'^T M\pi' + n' w'^T \pi' + \pi'^T w' n'^T + n' w_3 n'^T\big]^2, \quad (39)$$

where $\widetilde{C} = (D'\psi_0\pi' + n(F)\otimes n')^T(D'\psi_0\pi' + n(F)\otimes n')$. Moreover, frame-indifference implies also that $\widetilde{C}$ depends only on C'. Finally, if the shell is homogeneous along its fibers in the geometric configuration, we have

$$\widetilde{P}_{C'}(M) = \inf_{\substack{w'\in T_{x'}S' \\ w_3\in\mathbb{R}}} \frac{\partial^2\widetilde{W}_S}{\partial\widetilde{C}^2}(x', \widetilde{C})\big[\pi'^T M\pi' + n'w'^T\pi' + \pi'^T w'n'^T + n'w_3n'^T\big]^2. \quad (40)$$

This completes the proof. □

5.5. ***Planar isotropy.*** We say the shell is isotropic along its midsection if, for every $x'\in S'$ and $(F', F_3)\in T^*_{x'}(S';\mathbb{R}^3)\times\mathbb{R}^3$, we have

$$\widetilde{W}^\varepsilon(x', F'R\pi'_{x'} + F_3\otimes n') = \widetilde{W}^\varepsilon(x', F'\pi'_{x'} + F_3\otimes n')$$

for every $R\in \mathrm{SO}(T_{x'}S')$. This is equivalent to the definition used in the abstract configuration. We investigate the consequences of planar isotropy on the flexural part of the energy

$$\tilde{I}_{\mathrm{flex}}(\psi_0) := \frac{1}{3}\int_{S'} \widetilde{Q}^0_{D'\psi_0}(D'n - D'\psi_0 D'n', 1)\,dx'.$$

Proposition 15. *If the shell is isotropic along its midsection, then*

$$\tilde{I}_{\mathrm{flex}}(\psi_0) = \frac{1}{3}\int_{S'} \widetilde{P}_{C'}\big(|n(D'\psi_0)|b_{D'\psi_0} - (C'D'n' + (D'n')^T C')\big)\,dx',$$

where $b_{D'\psi_0}$ is the second fundamental form of the deformed surface, given by (27).

Proof. The proof is similar to Proposition 9. □

5.6. ***Right invariance under the special linear group.*** We say that the stored energy $\widetilde{W}_2$ is invariant under the special linear group if, for every $\tilde{x} = x' + x_3n'$ and $(F', F_3)\in T^*_{x'}(S';\mathbb{R}^3)\times\mathbb{R}^3$, we have

$$\widetilde{W}_2(\tilde{x}, F'U\pi'_{x'} + F_3\otimes n') = \widetilde{W}_2(\tilde{x}, F'\pi'_{x'} + F_3\otimes n')$$

for every $U\in \mathrm{SL}(T_{x'}S')$. Note that this definition is equivalent to the one given in the abstract configuration.

Proposition 16. *If $\widetilde{W}_2$ is right invariant under the special linear group, then*

$$\tilde{I}_{\mathrm{flex}}(\psi_0) = \frac{1}{3}\int_{S'} \widetilde{K}_{x',\det(C')}\big(|n(D'\psi_0)|H - H_0\big)\,dx', \quad (41)$$

where H and H_0 are the mean curvatures of the deformed shell $\psi_0(S')$ and undeformed shell S', respectively. Moreover, if the shell is homogeneous along its fibers,

then

$$\tilde{I}_{\text{flex}}(\psi) = \frac{1}{3}\int_{S'} \tilde{\kappa}_{x',\det(C')}(|n(D'\psi_0)|H - H_0)^2\,dx'.$$

Proof. For all sections F' of $T^*(S';\mathbb{R}^3)$ and $G' \in T^*_{x'}(S';\mathbb{R}^3)$, we have

$$\widetilde{P}_{C'}(F'^T(G' - F'D'n') + (G' - F'D'n')^T F') = P_{C'}(F'^T G' + G'^T F').$$

From Proposition 10 and (31), we deduce that

$$\widetilde{P}_{C'}(F'^T(G' - F'D'n') + (G' - F'D'n')^T F') = \widetilde{P}_{(\det C')^{1/2}\,\mathrm{Id}}\big(\tfrac{1}{2}\alpha(\det C')^{1/2}\,\mathrm{Id}\big),$$

with

$$\alpha = \mathrm{Tr}\big(C'^{-1/2}(F'^T(G' - F'D'n') + (G' - F'D'n')^T F')C'^{-1/2}\big).$$

We thus obtain (41) with

$$\widetilde{K}_{x',\det(C')}(H) = \widetilde{P}_{\det(C')\,\mathrm{Id}}\big(\tfrac{1}{2}H(\det C')^{1/2}\,\mathrm{Id}\big). \tag{42}$$

If the shell is homogeneous along its fibers, then the $\widetilde{P}_{(\det C')\,\mathrm{Id}}$ is homogeneous of degree two, whence the conclusion in this case. □

6. Examples

We are now in position to apply our formal convergence result to derive different models for isometric bending shells, vesicles and RBCs. Note that, in our setting, we do not derive the nonlinear membrane shell model (see [Le Dret and Raoult 1996]) since W_2 cannot be chosen to be equal to zero. Throughout this section, we assume that W^ε and $\widetilde{W}^\varepsilon$ satisfy the assumptions (2), (3) and (33), (34), respectively, that the Dirichlet boundary conditions on Γ^ε are given by (7) and (18) and that the minimizers φ^ε and $\tilde{\varphi}^\varepsilon$ of the total energy admit asymptotic expansions as in (6) and (35), respectively, while their total energies $I_\varepsilon(\varphi^\varepsilon)$ and $\tilde{I}_\varepsilon(\tilde{\varphi}^\varepsilon)$ remain bounded. Moreover, the stored energies are assumed to be frame-indifferent.

6.1. ***Isometric bending shells.*** In this section, we recover the isometric bending shell model by choosing $\widetilde{W}_0 = 0$ and the set $\widetilde{\mathcal{M}}$ of the zeros of $\widetilde{W}_2$ restricted to the midsection to be equal to

$$\widetilde{\mathcal{M}}_{\text{iso}} := S' \times \mathrm{SO}(3). \tag{43}$$

The sequence of minimizers of the energy converges toward the minimizer of an energy whose elastic part depends only on the difference between the second fundamental form of the deformed shell and that of its reference configuration.

Proposition 17. *If $\widetilde{W}^\varepsilon = \varepsilon^{-3}\widetilde{W}_2$ and if $\widetilde{W}_2$ is such that $\widetilde{\mathcal{M}} = \widetilde{\mathcal{M}}_{\text{iso}}$, given by* (43)*, then*

$$\varphi_0 = \arg\min_{\psi_0 \in \Psi_0} \tilde{I}_0(\psi_0),$$

where

$$\tilde{I}_0(\psi_0) = \frac{1}{3}\int_{S'} \widetilde{P}_{x',\mathrm{Id}}\big((D'\psi_0)^T D'n + D'n^T D'\psi_0 - (D'n' + D'n'^T)\big)\,dx' - 2\int_{S'} f_0\cdot\psi_0\,dx', \quad (44)$$

$f_0(x') = f(x',0)$ *for every* $x' \in S'$, $n = n(D'\psi_0)$, $\widetilde{P}_{x',\mathrm{Id}}$ *is a polynomial of degree at most two given by* (39), *and*

$$\Psi_0 = \{\psi_0 \in W^{1,\infty}(S')^3 : (D'\psi_0)^T D'\psi_0 = \mathrm{Id},\, n = n(D'\psi_0) \in W^{1,\infty}(S')^3, \psi_0(x') = \phi_0(x') \text{ and } n(x') = n_\gamma(x') \text{ for every } x' \in \gamma\}.$$

Moreover, if the shell is homogeneous along its fibers, then $\widetilde{P}_{x',\mathrm{Id}}$ *is homogeneous of degree two.*

Proof. It is a straightforward application of Proposition 12 and Proposition 14. □

Example. Let us give a practical example. For instance, one can choose the Saint Venant–Kirchhoff nonlinearly elastic stored energy function

$$\widetilde{W}_2(F) = \mu\,\mathrm{Tr}((F^TF - \mathrm{Id})^2) + \frac{\lambda}{2}\,\mathrm{Tr}(F^TF - \mathrm{Id})^2.$$

A simple computation leads to the energy

$$\tilde{I}_0(\psi_0) = \frac{1}{3}\int_{S'} 2\mu\,\mathrm{Tr}((b - b_{\mathrm{ref}})^2) + \frac{\lambda\mu}{2\mu+\lambda}\,\mathrm{Tr}(b - b_{\mathrm{ref}})^2 - f_0\cdot\psi_0\,dx',$$

where $b = (D'\psi_0)^T D'n + (D'n)^T D'\psi_0$ is the second fundamental form of the deformed shell and $b_{\mathrm{ref}} = (D'n')^T + D'n'$ is the second fundamental form of the undeformed shell.

6.2. ***Vesicles.*** In this section we derive Helfrich functionals, with or without spontaneous curvature, from three-dimensional elasticity. The main difference with the isometric case lies in the fact that we assume the energy to be right invariant under the special linear group $\mathrm{SL}(TS')$. Note that this readily implies that it may not be chosen to be isotropic. The Helfrich functional without spontaneous curvature is derived using the abstract configuration S, while the one with spontaneous curvature is obtained using the geometric configuration $\widetilde{S}$ of the shell.

6.2.1. *Without spontaneous curvature.* In this section we consider the case where the zero set of W_2 restricted to the midsection is given by

$$\mathcal{M}_H := \{(F', F_3) \in T^*(S';\mathbb{R}^3) \times T_0^*((-1,1);\mathbb{R}^3) : \det(F) = 1,\, F_3\cdot v = \det(F', v) \text{ for every } v \in T_0^*((-1,1);\mathbb{R}^3)\}. \quad (45)$$

From Propositions 5 and 10, we obtain that the minimizers of the energy formally converge to the Helfrich functional with no spontaneous curvature.

Proposition 18. *Suppose that* $W^\varepsilon = \varepsilon^{-3} W_2$, W_2 *is right invariant under* $\mathrm{SL}(TS')$, *and that* $\mathcal{M} = \mathcal{M}_H$ *as given by* (45). *If the shell is homogeneous along its fibers in the abstract configuration, then*

$$\varphi_0 = \arg\min_{\psi_0 \in \Psi_0} I_0(\psi_0),$$

where

$$I_0(\psi_0) = \frac{1}{3}\int_{S'} \kappa |H|^2 \, dx' - 2\int_{S'} f_0 \cdot \psi_0 \, dx', \tag{46}$$

$f_0(x') = f(x', 0)$ *for every* $x' \in S'$, H *is the mean curvature of the deformed shell* $\psi_0(S')$, $\kappa(x') = P_{\mathrm{Id}}\big(\frac{1}{2}\,\mathrm{Id}\big)$, *where* P_{Id} *is given by* (25), *and*

$$\Psi_0 = \{\psi_0 \in W^{1,\infty}(S')^3 : n \in W^{1,\infty}(S')^3 \text{ and } \det((D'\psi_0)^T D'\psi_0) = 1,$$
$$\psi_0(x') = \phi_0(x') \text{ and } n(x') = n_\gamma(x') \text{ for } x' \in \gamma\},$$

where n *is the normal to the deformed surface* $\psi_0(S')$.

Example. Proposition 18 can be applied with

$$W_2(F) = W_S(C) = \alpha(\det(C) - 1)^2 + \beta|Ce_3 - e_3|^2, \tag{47}$$

where $C = F^T F$ and α and β are positive real constants. A simple computation leads to

$$D^2 W_S\left[0, \begin{pmatrix} \frac{1}{2}\,\mathrm{Id} & w' \\ w' & w_3 \end{pmatrix}_{\mathrm{Id}}\right]^2 = 2\alpha(1 + w_3)^2 + 2\beta|(w', w_3)|^2.$$

Then, from the expression (25) of P_{Id}, we get

$$\kappa = P_{\mathrm{Id}}\big(\tfrac{1}{2}\,\mathrm{Id}\big) = \inf_w 2\alpha(1 + w_3)^2 + 2\beta|(w', w_3)|^2 = 2(\alpha^{-1} + \beta^{-1})^{-1}.$$

Hence, the limit energy in this case is

$$I_0(\psi_0) = \frac{2}{3}\int_{S'} (\alpha^{-1} + \beta^{-1})^{-1} |H|^2 \, dx' - 2\int_{S'} f_0 \cdot \psi_0 \, dx'.$$

6.2.2. *With spontaneous curvature.* Here we derive from three-dimensional elasticity a model of shells whose limit energy is the Helfrich functional with nonzero spontaneous curvature. Basically, such a model is obtained by using the same assumptions as in the previous case but cast in the geometric configuration, with a set of zeros restricted to the midsection for $\widetilde{W}_2$ given by

$$\widetilde{\mathcal{M}}_H := \{(x', F) \in S' \times \mathbb{R}^{3\times 3} : \det(F) = 1 \text{ and } (\mathrm{Cof}\, F - F)n' = 0\}. \tag{48}$$

The following proposition is a direct application of Propositions 12 and 16.

Proposition 19. *Suppose that* $\widetilde{W}^\varepsilon = \varepsilon^{-3}\widetilde{W}_2$, $\widetilde{W}_2$ *is right invariant under* $\mathrm{SL}(TS')$, *and such that* $\widetilde{\mathcal{M}} = \widetilde{\mathcal{M}}_H$ *as given by* (48). *If the shell is homogeneous along its fibers in the geometric configuration, then*

$$\varphi_0 = \arg\min_{\psi_0 \in \Psi_0} \tilde{I}_0(\psi_0),$$

where

$$\tilde{I}_0(\psi_0) = \frac{1}{3}\int_{S'} \tilde{\kappa}|H - H_0|^2\, dx' - 2\int_{S'} f_0 \cdot \psi_0\, dx', \tag{49}$$

$f_0(x') = \widetilde{f}(x', 0)$ *for every* $x' \in S'$, *where* H *is the mean curvature of the deformed shell* $\psi_0(S')$ *and* H_0 *is the mean curvature of* S', *with*

$$\tilde{\kappa}(x') = \widetilde{P}_{\mathrm{Id}}\big(\tfrac{1}{2}\,\mathrm{Id}\big),$$

where $\widetilde{P}_{\mathrm{Id}}$ *is given by* (40) *and*

$$\Psi_0 = \{\psi_0 \in W^{1,\infty}(S')^3 : n \in W^{1,\infty}(S')^3 \text{ and } \det(D'\psi_0{}^T D'\psi_0) = 1,\\ \psi_0(x') = \phi_0(x') \text{ and } n(x') = n_\gamma(x') \text{ for every } x' \in \gamma\}, \tag{50}$$

where n *is the normal to the deformed surface* $\psi_0(S')$.

Example. The stored energy function

$$\widetilde{W}_2(x', F) = \widetilde{W}_S x' \widetilde{C} = \alpha(\det(F^T F) - 1)^2 + \beta|F^T F n' - n'|^2$$

satisfies the assumptions of Proposition 19, and we have

$$\tilde{\kappa} = \inf_{\substack{w' \in T_{x'}S' \\ w_3 \in \mathbb{R}}} \frac{\partial \widetilde{W}_S}{\partial \widetilde{C}^2}(x', \mathrm{Id})\big[\tfrac{1}{2}\pi'^T\pi' + n'w'^T\pi' + \pi'^T w' n'^T + n' w_3 n'^T\big]^2.$$

Furthermore, we have

$$\frac{\partial \widetilde{W}_S}{\partial \widetilde{C}^2}(x', \mathrm{Id})[\delta C]^2 = 2(\alpha\,\mathrm{Tr}(\delta C)^2 + \beta|\delta C n'|^2),$$

so that

$$\tilde{\kappa} = \inf_{\substack{w' \in T_{x'}S' \\ w_3 \in \mathbb{R}}} 2\big(\alpha(1 + w_3)^2 + \beta(|w'|^2 + w_3^2)\big) = 2(\alpha^{-1} + \beta^{-1})^{-1}.$$

For such a choice of $\widetilde{W}_2$, and under the assumptions made in Proposition 19, the sequence of minimizers $\tilde{\varphi}^\varepsilon$ formally converges toward a minimizer of

$$\tilde{I}_0(\psi_0) = \frac{2}{3}\int_{S'} (\alpha^{-1} + \beta^{-1})^{-1}|H - H_0|^2\, dx' - 2\int_{S'} \widetilde{f_0} \cdot \psi_0\, dx'$$

over Ψ_0 given by (50). Note that this is the set of deformations that preserve the local area of the shell supplemented with boundary conditions on a subset of the boundary.

6.3. ***Red blood cells.*** The mechanical behavior of a red blood cell (RBC) is driven by the nature of its membrane, which is mainly made of a lipid bilayer. Note that in addition to the lipid bilayer, RBCs are also composed of a protein skeleton. This skeleton ensures a small resistance of the RBCs to shear stress. Such a model may be obtained as the limit of the three-dimensional elasticity. In this section, we derive a model of the mechanical behavior of RBCs as the limit of genuine three-dimensional elasticity. To this end, we consider a stored energy W^ε whose asymptotic assumption reads as

$$W^\varepsilon = \varepsilon^{-1}(\varepsilon^{-2}W_2 + W_0),$$

where W_2 satisfies the same assumption as in the study of vesicles without spontaneous curvature (see Section 6.2.1), namely, its zero set restricted to the midsection is given by (45). We get that the sequence φ^ε of minimizers formally converges to

$$\varphi_0 = \operatorname*{arg\,min}_{\psi_0 \in \Psi_0} I_0(\psi_0),$$

where

$$I_0(\psi_0) = \frac{1}{3}\int_{S'} k|H|^2\,dx' + 2\int_{S'} W_0(D'\psi_0, n)\,dx' - 2\int_{S'} f_0 \cdot \psi_0\,dx',$$

$f_0(x') = f(x', 0)$, n is the normal to the deformed shell $\psi_0(S')$ and Ψ_0 is the set of deformations that preserve the local area of the shell and satisfy the boundary conditions

$$\psi_0(x') = \phi_0(x') \quad \text{and} \quad n(x') = n_\gamma(x') \quad \text{for every } x' \in \gamma.$$

Example. As an example, we can choose the nonlinearly elastic Saint Venant–Kirchhoff stored energy function

$$W_0(F) = \mu\,\mathrm{Tr}((C - \mathrm{Id})^2) + \frac{\lambda}{2}\,\mathrm{Tr}(C - \mathrm{Id})^2 \quad \text{with } C = F^T F,$$

and $W_2(F)$ as in (47). This leads to a limit energy

$$I_0(\psi_0) = \frac{2}{3}\int_{S'} (\alpha^{-1} + \beta^{-1})^{-1}|H|^2\,dx' + 2\int_{S'} \Big(\mu\,\mathrm{Tr}((D'\psi_0{}^T D'\psi_0 - \mathrm{Id}')^2) + \frac{\lambda}{2}\,\mathrm{Tr}(D'\psi_0{}^T D'\psi_0 - \mathrm{Id}')^2\Big)\,dx' - 2\int_{S'} f_0 \cdot \psi_0\,dx'.$$

7. Conclusion

In this article we prove, using a formal approach, that new nonlinearly elastic shell models may be derived assuming the shell to be highly anisotropic. Notably, it enables us to derive some models used in the study of vesicles and RBCs. Part of the results presented in this article have since been proved by a Γ-convergence approach in an Eulerian setting for the justification of the modeling of vesicles by Merlet [2013a; 2013b]. Finally, let us recall and emphasize the fact that the computation of the limit energy should include a relaxation step that is not taken into account in our formal framework. The only interesting case being the one where the flexural term $Q^0_{F'}(G, s)$ is not fully degenerate, that is not independent of G. In such a case, a relaxation of the membrane term of the limit energy is expected to take place. The correct limit energy in Proposition 5 should read

$$I_0{}'(\psi_0) = \frac{1}{3}\int_{S'} Q^0_{D'\psi_0}(D'n, 1)\,dx' + 2\int_{S'} \mathcal{Q}'W_0(D'\psi_0, n)\,dx' - 2\int_{S'} f_0 \cdot \psi_0\,dx',$$

where $\mathcal{Q}'W_0$ is the in-plane quasiconvexification of W_0, defined for every element F' of $T^*(S'; \mathbb{R}^3)$ by

$$\mathcal{Q}'W_0(F', n) = \inf_{\varphi \in C^\infty_0(\omega; T_{x'}S')} |\omega|^{-1} \int_\omega W_0(F'(\mathrm{Id}' + D'\varphi), n)\,dy',$$

where $x' = \pi_{S'}(F')$ and ω is a bounded regular open set of $T_{x'}S'$.

Notation

- $\mathbb{R}$, set of reals
- $\mathbb{R}'$, dual set of reals
- $\mathbb{N}$, set of nonnegative integers
- S', midsurface of the shell in the reference configuration
- ε, thickness of the shell
- $S^\varepsilon := S' \times (-\varepsilon, \varepsilon)$, abstract reference configuration of the shell
- $S := S^1$, rescaled abstract reference configuration of the shell
- $\widetilde{S}^\varepsilon$, geometric reference configuration
- $\widetilde{S}$, geometric reference configuration of maximum thickness
- x^ε, element of S^ε
- $\tilde{x}$, element of $\widetilde{S}$
- x', element of S'
- TM, tangent space to M

- T_xM, tangent fiber to M at x
- T^*M, cotangent space to M
- T^*_xM, cotangent space to M at x
- $T(M;\mathbb{R}^3) := TM \oplus TM \oplus TM$, Whitney triple sum of TM
- $T_x(M;\mathbb{R}^3) := T_xM \oplus T_xM \oplus T_xM$, Whitney triple sum of T_xM
- $T^*(M;\mathbb{R}^3) := T^*M \oplus T^*M \oplus T^*M$, Whitney triple sum of T^*M
- $T^*_x(M;\mathbb{R}^3) := T^*_xM \oplus T^*_xM \oplus T^*_xM$, Whitney triple sum of T^*_xM
- $\pi_B : \mathcal{F} \to B$, projection of a fiber bundle $\mathcal{F}$ onto its base B
- $\pi'_{x'} : \mathbb{R}^3 \to T_{x'}S'$, projection onto $T_{x'}S'$
- ψ^ε, deformation of S^ε
- $\psi(\varepsilon) : S \to \mathbb{R}^3$, rescaled map of the deformation ψ^ε
- $\boldsymbol{\psi} = (\psi_k)$, expansion of the deformation $\psi(\varepsilon) = \sum_k \varepsilon^k \psi_k$ with respect to the thickness
- $\boldsymbol{\varphi} = (\varphi_k)$, expansion of the minimizers $\varphi(\varepsilon) = \sum_k \varepsilon^k \varphi_k$ with respect to the thickness
- $D\psi^\varepsilon$, differential of $\psi^\varepsilon : S^\varepsilon \to \mathbb{R}^3$
- $D\psi^\varepsilon(x^\varepsilon)$, differential of $\psi^\varepsilon : S^\varepsilon \to \mathbb{R}^3$ at x^ε
- $(D'\psi^\varepsilon, D_3\psi^\varepsilon)$, decomposition of $D\psi^\varepsilon \in T^*(S^\varepsilon;\mathbb{R}^3)$ with respect to the product $T^*(S';\mathbb{R}^3) \times T^*((-\varepsilon,\varepsilon);\mathbb{R}^3)$, where $\psi^\varepsilon : S^\varepsilon \to \mathbb{R}^3$
- $(D'\psi^\varepsilon(x), D_3\psi^\varepsilon(x))$, decomposition of $D\psi^\varepsilon(x) \in T^*_x(S^\varepsilon;\mathbb{R}^3)$ with respect to the product $T^*_{x'}(S';\mathbb{R}^3) \times T^*_{x_3}((-\varepsilon,\varepsilon);\mathbb{R}^3)$, where $\psi^\varepsilon : S^\varepsilon \to \mathbb{R}^3$
- ∂_3, partial differentiation along the fibers
- $J_\varepsilon(\psi^\varepsilon)$, elastic energy of a deformation $\psi^\varepsilon : S^\varepsilon \to \mathbb{R}^3$
- $I_\varepsilon(\psi^\varepsilon)$, total energy of a deformation $\psi^\varepsilon : S^\varepsilon \to \mathbb{R}^3$
- $L_\varepsilon(\psi^\varepsilon)$, work of the external loads
- $J(\varepsilon)(\psi(\varepsilon))$, rescaled elastic energy of the deformation ψ^ε
- $I(\varepsilon)(\psi(\varepsilon))$, rescaled total energy of the deformation ψ^ε
- $\boldsymbol{J}(\varepsilon)(\boldsymbol{\psi})$, elastic energy of the deformation of asymptotic expansion $\boldsymbol{\psi}$
- $\boldsymbol{I}(\varepsilon)(\boldsymbol{\psi})$, total energy of the deformation of asymptotic expansion $\boldsymbol{\psi}$
- $I_0(\psi_0)$, limit of the total energy (for standard boundary conditions)
- f_ε, external loads
- $W^\varepsilon : T^*(S^\varepsilon;\mathbb{R}^3) \to \overline{\mathbb{R}}^+$, stored energy
- Ψ^ε, set of admissible deformations of S^ε

- $\mathbf{\Psi}$, set of admissible asymptotic expansions for the deformations of S^ε
- $W_k : T^*(S; \mathbb{R}^3) \to \mathbb{R}^+$, k-th term of the asymptotic expansion of the stored energy
- $\mathcal{M} \subset T^*(S; \mathbb{R}^3)$, the restriction of the zero set of W_2 to the midsection S'
- $\mathcal{M}' \subset T^*(S'; \mathbb{R}^3)$, projection of $\mathcal{M}$ on $T^*(S'; \mathbb{R}^3)$
- $n_0 : \mathcal{M}' \to T_0^*((-1, 1); \mathbb{R}^3)$, orientation of the normal fiber in the deformed configuration
- $n : \mathcal{M}' \to \mathbb{R}^3$, orientation of the normal fiber in the deformed configuration (only the vectorial part)
- n', normal to the midsection S' of the reference configuration
- $D^2 W_2$, second derivative of W_2 on $\mathcal{M}$
- $a[\,\cdot\,]^2 = a(\,\cdot\,,\,\cdot\,)$, where a is a bilinear form
- $Q_F : T^*(S'; \mathbb{R}^3) \times \mathbb{R}^3 \to \mathbb{R}$, quadratic form associated to the flexural limit energy, where F is a section of $\mathcal{M}'$
- $Q_F^0 := \inf_v Q_F(\,\cdot\,, v) : T^*(S'; \mathbb{R}^3) \to \mathbb{R}$
- $I_{\text{flex}}(\psi)$, flexural limit energy of a deformation ψ
- dx', 2-dimensional Hausdorff measure restricted to S'
- $\mathrm{Tr}(A)$, the trace of the matrix A
- $\tilde{\psi}^\varepsilon$, $\widetilde{W}^\varepsilon$, $\widetilde{W}_k$, $\widetilde{\mathcal{M}}$, $\widetilde{L}_\varepsilon$, $\tilde{f}_\varepsilon$, ..., variables with tildes are defined on the geometric configuration
- $\mathrm{SO}(TM)$, rotations of TM, that is, the fiber bundle made of all rotations of $T_x M$ (with $x \in M$), where M is a Riemannian manifold
- $\mathrm{SO}(n)$, rotations of $\mathbb{R}^n$
- $\mathrm{SL}(TM)$, special group of TM, that is, the fiber bundle made of all linear diffeomorphisms of $T_x M$ with determinant one (with $x \in M$), where M is a Riemannian manifold
- $\mathscr{E}_M$, set of symmetric bilinear forms on the tangent space of M
- $\mathbb{C}$, fiber bundle over S' whose fibers are the traceless maps from $T_{x'} S'$ into itself
- $\mathscr{P}$, set of polynomials of degree at most two on $\mathscr{E}_{S'}$

References

[Agrawal and Steigmann 2009] A. Agrawal and D. J. Steigmann, "Boundary-value problems in the theory of lipid membranes", *Contin. Mech. Thermodyn.* **21**:1 (2009), 57–82.

[Bonito et al. 2010] A. Bonito, R. H. Nochetto, and M. S. Pauletti, "Parametric FEM for geometric biomembranes", *J. Comput. Phys.* **229**:9 (2010), 3171–3188.

[Bonito et al. 2011] A. Bonito, R. H. Nochetto, and M. S. Pauletti, "Dynamics of biomembranes: effect of the bulk fluid", *Math. Model. Nat. Phenom.* **6**:5 (2011), 25–43.

[Canham 1970] P. Canham, "The minimum energy of bending as a possible explanation of the biconcave shape of the human red blood cell", *Journal of Theoretical Biology* **26**:1 (1970), 61–81.

[Conti and Maggi 2008] S. Conti and F. Maggi, "Confining thin elastic sheets and folding paper", *Arch. Ration. Mech. Anal.* **187**:1 (2008), 1–48.

[Conti et al. 2006] S. Conti, F. Maggi, and S. Müller, "Rigorous derivation of Föppl's theory for clamped elastic membranes leads to relaxation", *SIAM J. Math. Anal.* **38**:2 (2006), 657–680.

[Crowl and Fogelson 2010] L. M. Crowl and A. L. Fogelson, "Computational model of whole blood exhibiting lateral platelet motion induced by red blood cells", *Int. J. Numer. Methods Biomed. Eng.* **26**:3-4 (2010), 471–487.

[Deuling and Helfrich 1976] H. Deuling and W. Helfrich, "The curvature elasticity of fluid membranes : A catalogue of vesicle shapes", *J. Phys. France* **37**:11 (1976), 1335–1345.

[Doyeux et al. 2013] V. Doyeux, Y. Guyot, V. Chabannes, C. Prud'homme, and M. Ismail, "Simulation of two-fluid flows using a finite element/level set method. Application to bubbles and vesicle dynamics", *J. Comput. Appl. Math.* **246** (2013), 251–259.

[Du and Zhang 2008] Q. Du and J. Zhang, "Adaptive finite element method for a phase field bending elasticity model of vesicle membrane deformations", *SIAM J. Sci. Comput.* **30**:3 (2008), 1634–1657.

[Du et al. 2004] Q. Du, C. Liu, and X. Wang, "A phase field approach in the numerical study of the elastic bending energy for vesicle membranes", *J. Comput. Phys.* **198**:2 (2004), 450–468.

[Du et al. 2006] Q. Du, C. Liu, and X. Wang, "Simulating the deformation of vesicle membranes under elastic bending energy in three dimensions", *J. Comput. Phys.* **212**:2 (2006), 757–777.

[Dziuk 2008] G. Dziuk, "Computational parametric Willmore flow", *Numer. Math.* **111**:1 (2008), 55–80.

[Evans 1974] E. Evans, "Bending resistance and chemically induced moments in membrane bilayers", *Biophysical Journal* **14**:12 (1974), 923 – 931.

[Feng and Klug 2006] F. Feng and W. S. Klug, "Finite element modeling of lipid bilayer membranes", *J. Comput. Phys.* **220**:1 (2006), 394–408.

[Fox et al. 1993] D. D. Fox, A. Raoult, and J. C. Simo, "A justification of nonlinear properly invariant plate theories", *Arch. Rational Mech. Anal.* **124**:2 (1993), 157–199.

[Friesecke et al. 2002] G. Friesecke, R. D. James, and S. Müller, "A theorem on geometric rigidity and the derivation of nonlinear plate theory from three-dimensional elasticity", *Comm. Pure Appl. Math.* **55**:11 (2002), 1461–1506.

[Friesecke et al. 2003] G. Friesecke, R. D. James, M. G. Mora, and S. Müller, "Derivation of nonlinear bending theory for shells from three-dimensional nonlinear elasticity by Gamma-convergence", *C. R. Math. Acad. Sci. Paris* **336**:8 (2003), 697–702.

[Friesecke et al. 2006] G. Friesecke, R. D. James, and S. Müller, "A hierarchy of plate models derived from nonlinear elasticity by gamma-convergence", *Arch. Ration. Mech. Anal.* **180**:2 (2006), 183–236.

[Helfrich 1973] W. Helfrich, "Elastic properties of lipid bilayers: theory and possible experiments", *Naturforsch C* **28**:11 (1973), 693–703.

[Jenkins 1977a] J. T. Jenkins, "The equations of mechanical equilibrium of a model membrane", *SIAM J. Appl. Math.* **32**:4 (1977), 755–764.

[Jenkins 1977b] J. T. Jenkins, "Static equilibrium configurations of a model red blood cell", *Journal of Mathematical Biology* **4** (1977), 149–169.

[Kim and Lai 2010] Y. Kim and M.-C. Lai, "Simulating the dynamics of inextensible vesicles by the penalty immersed boundary method", *J. Comput. Phys.* **229**:12 (2010), 4840–4853.

[Krishnaswamy 1996] S. Krishnaswamy, "A cosserat-type model for the red blood cell wall", *International Journal of Engineering Science* **34**:8 (1996), 873–899.

[Le Dret and Raoult 1995] H. Le Dret and A. Raoult, "The nonlinear membrane model as variational limit of nonlinear three-dimensional elasticity", *J. Math. Pures Appl.* (9) **74**:6 (1995), 549–578.

[Le Dret and Raoult 1996] H. Le Dret and A. Raoult, "The membrane shell model in nonlinear elasticity: a variational asymptotic derivation", *J. Nonlinear Sci.* **6**:1 (1996), 59–84.

[Liu and Liu 2006] Y. Liu and W. K. Liu, "Rheology of red blood cell aggregation by computer simulation", *J. Comput. Phys.* **220**:1 (2006), 139–154.

[Liu et al. 2004] Y. Liu, L. Zhang, X. Wang, and W. K. Liu, "Coupling of Navier–Stokes equations with protein molecular dynamics and its application to hemodynamics", *Internat. J. Numer. Methods Fluids* **46**:12 (2004), 1237–1252.

[Luke 1982] J. C. Luke, "A method for the calculation of vesicle shapes", *SIAM J. Appl. Math.* **42**:2 (1982), 333–345.

[Luke and Kaplan 1979] J. C. Luke and J. I. Kaplan, "On the theoretical shapes of bilipid vesicles under conditions of increasing membrane area", *Biophysical Journal* **25**:1 (1979), 107–111.

[Maitre et al. 2009] E. Maitre, T. Milcent, G.-H. Cottet, A. Raoult, and Y. Usson, "Applications of level set methods in computational biophysics", *Math. Comput. Modelling* **49**:11-12 (2009), 2161–2169.

[Merlet 2013a] B. Merlet, "A highly anisotropic nonlinear elasticity model for vesicles I. Eulerian formulation, rigidity estimates and vanishing energy limit", 2013, http://hal.archives-ouvertes.fr/hal-00848547/PDF/Merlet_VesiclesPartI.pdf. 23 pages, preprint.

[Merlet 2013b] B. Merlet, "A highly anisotropic nonlinear elasticity model for vesicles. II. Derivation of the thin bilayer bending theory", 2013, http://hal.archives-ouvertes.fr/hal-00848552/PDF/Merlet_VesiclesPartII.pdf. 57 pages, preprint.

[Miao et al. 1994] L. Miao, U. Seifert, M. Wortis, and H.-G. Döbereiner, "Budding transitions of fluid-bilayer vesicles: The effect of area-difference elasticity", *Phys. Rev. E* **49**:6 (1994), 5389–5407.

[Pantz 2003] O. Pantz, "On the justification of the nonlinear inextensional plate model", *Arch. Ration. Mech. Anal.* **167**:3 (2003), 179–209.

[Peng et al. 2013] Z. Peng, X. Li, I. V. Pivkin, M. Dao, G. E. Karniadakis, and S. Suresh, "Lipid bilayer and cytoskeletal interactions in a red blood cell", *Proc. Nat. Acad. Sci.* (2013).

[Salac and Miksis 2011] D. Salac and M. Miksis, "A level set projection model of lipid vesicles in general flows", *J. Comput. Phys.* **230**:22 (2011), 8192–8215.

[Seifert et al. 1991] U. Seifert, K. Berndl, and R. Lipowsky, "Shape transformations of vesicles: Phase diagram for spontaneous- curvature and bilayer-coupling models", *Phys. Rev. A* **44** (1991), 1182–1202.

[Singer and Nicolson 1972] S. J. Singer and G. L. Nicolson, "The fluid mosaic model of the structure of cell membranes", *Science* **175**:4023 (1972), 720–731.

[Sohn et al. 2010] J. S. Sohn, Y.-H. Tseng, S. Li, A. Voigt, and J. S. Lowengrub, "Dynamics of multicomponent vesicles in a viscous fluid", *J. Comput. Phys.* **229**:1 (2010), 119–144.

[Steigmann 1999] D. J. Steigmann, "Fluid films with curvature elasticity", *Archive for Rational Mechanics and Analysis* **150** (1999), 127–152.

[Svetina and Žekš 1989] S. Svetina and B. Žekš, "Membrane bending energy and shape determination of phospholipid vesicles and red blood cells", *European Biophysics Journal* **17** (1989), 101–111.

[Veerapaneni et al. 2009] S. K. Veerapaneni, D. Gueyffier, D. Zorin, and G. Biros, "A boundary integral method for simulating the dynamics of inextensible vesicles suspended in a viscous fluid in 2D", *J. Comput. Phys.* **228**:7 (2009), 2334–2353.

Received 24 Jul 2012. Revised 6 Jun 2014. Accepted 9 Jun 2014.

OLIVIER PANTZ: olivier.pantz@polytechnique.org
Centre de Mathématiques Appliquées, École Polytechnique, Route de Saclay, 91128 Palaiseau CEDEX, France

KARIM TRABELSI: karim.trabelsi@polytechnique.edu
Direction du la Recherche et de l'Innovation, Institut Polytechnique des Sciences Avancées, 5-9 rue Maurice Grandcoing, 94200 Ivry-sur-Seine, France

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 3, No. 2, 2015
dx.doi.org/10.2140/memocs.2015.3.139

CANONICAL DUALITY THEORY AND TRIALITY FOR SOLVING GENERAL GLOBAL OPTIMIZATION PROBLEMS IN COMPLEX SYSTEMS

DANIEL MORALES-SILVA AND DAVID Y. GAO

General nonconvex optimization problems are studied by using the canonical duality-triality theory. The triality theory is proved for sums of exponentials and quartic polynomials, which solved an open problem left in 2003. This theory can be used to find the global minimum and local extrema, which bridges a gap between global optimization and nonconvex mechanics. Detailed applications are illustrated by several examples.

1. Introduction and motivation

This paper intends to solve the following nonconvex optimization problem ($(\mathcal{P})$ in short):

$$(\mathcal{P}) : \operatorname{ext}\big\{\Pi(\boldsymbol{x}) = W(\boldsymbol{x}) + \tfrac{1}{2}\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} - \boldsymbol{f}^t\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\big\}, \tag{1}$$

where $\operatorname{ext}\{\cdot\}$ denotes finding extremum points of a function given in $\{\cdot\}$, $\boldsymbol{f} \in \mathbb{R}^n$ is a given (input) vector, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a given symmetric matrix, and $W : \mathbb{R}^n \to \mathbb{R}$ is a combination of fourth-order polynomials (double-well functions) and quadratic-exponential functions, namely

$$W(\boldsymbol{x}) := \sum_{i\in I_m} \exp\big(\tfrac{1}{2}\boldsymbol{x}^t\boldsymbol{B}_i\boldsymbol{x} - \alpha_i\big) + \sum_{j\in I_p} \tfrac{1}{2}b_j\big(\tfrac{1}{2}\boldsymbol{x}^t\boldsymbol{C}_j\boldsymbol{x} - \theta_j\big)^2,$$

where $I_m = \{1, \dots, m\}$ and $I_p = \{1, \dots, p\}$ are two integer sets with m and p that are fixed integers; all the coefficients b_j with $j \in I_p$ are positive constants, and $\alpha_i, \theta_j \in \mathbb{R}$ for all $i \in I_m$ and $j \in I_p$ are given parameters; the matrices $\{\boldsymbol{B}_i\}_{i\in I_m}$ and $\{\boldsymbol{C}_j\}_{j\in I_p}$ are assumed to be symmetric, positive semidefinite such that the cone generated by them contains a positive-definite matrix.

The nonconvex optimization problem $(\mathcal{P})$ arises naturally in complex systems with a wide range of applications, including chaotic dynamical systems [Gao 2003a; Gao and Ogden 2008a; Gao and Ruan 2008], computational biology [Zhang et al.

Communicated by Martin Ostoja-Starzewski.
MSC2010: 49N15, 90C26.
Keywords: canonical duality, triality theory, nonlinear analysis, nonconvex optimization, complex systems.

2011], chemical-database analysis [Xie and Schlick 2000], large-deformation computational mechanics [Gao 1996; Santos and Gao 2012], population growing [Ruan and Gao 2014a], location/allocation, network communication [Gao et al. 2012], and transitions of solids [Gao and Ogden 2008a; 2008b; Gao and Yu 2008], etc.

For example, the popular sensor-network location problem is to solve the following system of nonlinear equations (see [Aspnes et al. 2004; Moré and Wu 1997]):

$$\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2^2 = d_{ij}^2 \quad \forall (i,j) \in \mathcal{J}_p, \qquad \boldsymbol{u}_k = \boldsymbol{a}_k \quad \forall k \in \mathcal{J}_b, \tag{2}$$

where the vectors $\boldsymbol{u}_i = \{u_i^\alpha\} \in \mathbb{R}^d$ $(i = 1, \ldots, p)$ represent the locations of the unknown sensors, $\mathcal{J}_p = \{(i,j) : i < j,\ d_{ij} \text{ is specified}\}$ and $\mathcal{J}_b = \{k : \boldsymbol{u}_k = \boldsymbol{a}_k \text{ is specified}\}$ are two given index sets, d_{ij} are given distances for $(i,j) \in \mathcal{J}_p$, and the given vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q \in \mathbb{R}^d$ are the so-called anchors. The notation $\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2$ denotes the Euclidean distance between $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$; i.e.,

$$\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2 = \sqrt{\sum_{\alpha=1}^{d} (u_i^\alpha - u_j^\alpha)^2}.$$

By using the least-squares method, the quadratic equations (2) of the sensor-localization problem can be reformulated as an optimization problem:

$$\min\left\{ P(\boldsymbol{u}) = \sum_{(i,j)\in\mathcal{J}_p} \tfrac{1}{2}\big(\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2^2 - d_{ij}^2\big)^2 : \boldsymbol{u}_i \in \mathcal{U}_a \right\}, \tag{3}$$

where $\mathcal{U}_a = \{\boldsymbol{u} \in \mathbb{R}^{d\times p} : \boldsymbol{u}_k = \boldsymbol{a}_k\ \forall k \in \mathcal{J}_b\}$ is a feasible space. Let $\boldsymbol{x} = \{\{u_1^1, \ldots, u_1^d\}, \ldots, \{u_p^1, \ldots, u_p^d\}\} \in \mathbb{R}^n$ $(n = d \times p)$ denote an extended vector. By using the Lagrange-multiplier method to relax the boundary conditions in $\mathcal{U}_a$, the least-squares method for the sensor-localization problem (3) can be written in the problem (1) for certain properly defined matrices $\{\boldsymbol{C}_j\}$, which are the so-called deformation matrices in structural mechanics. The sensor-network-localization-type problems also appear in computational biology, Euclidean ball packing, molecular confirmation, recently wireless network communication, etc. [Ruan and Gao 2014b; Zhang et al. 2011]. Due to the nonconvexity, the sensor-network localization problem is considered to be NP-hard even for the simplest case $d = 1$ [Moré and Wu 1997; Saxe 1979]. A recent result of Aspnes et al. [2004] shows that the problem of computing a realization of the sensors on the plane is NP-complete in general.

Mathematics and mechanics have been complementary partners since Newton. Many fundamental ideas, concepts, and mathematical methods extensively used in calculus of variations and optimization originated from mechanics. For example, the Lagrange-multiplier method was first proposed by Lagrange from the classical analytic mechanics while the concepts of superpotential and subdifferential in modern convex analysis were introduced in [Moreau 1968; Moreau et al. 1988] from

frictional mechanics. From the point of view of computational large-deformation mechanics, both the fourth-order polynomial-minimization problem ($\mathcal{P}$) and the sensor-localization problem (3) are actually two special cases of discretized finite-deformation problems [Gao 1996]. It is known that, in continuum mechanics and differential geometry, the deformation $\boldsymbol{u}(\boldsymbol{x}) : \Omega \to \mathbb{R}^r$ is a vector field over an open domain $\Omega \subset \mathbb{R}^r$, and the minimal-potential variational problem is defined by

$$\min\left\{ P(\boldsymbol{u}) = \int_\Omega [W(\nabla \boldsymbol{u}) - \boldsymbol{u}^T \boldsymbol{f}]\, \mathrm{d}\Omega : \boldsymbol{u} \in \mathcal{U}_a \right\}, \tag{4}$$

where $W(\boldsymbol{F})$ is the so-called *stored strain energy*, which is usually a nonconvex function of the deformation gradient $\boldsymbol{F} = \nabla \boldsymbol{u}$, the feasible set $\mathcal{U}_a$ in this nonconvex variational problem is called the *kinematically admissible space*, where certain boundary conditions are prescribed. According to the hyperelasticity law (see [Gao 2000b, Chapter 6.1.2] or [Marsden and Hughes 1983]), the stored strain energy should be an *objective function* of the deformation gradient $\boldsymbol{F}$; i.e., there exists an objective strain measure $\boldsymbol{E}(\boldsymbol{F})$ and a convex function $V(\boldsymbol{E})$ such that

$$W(\nabla \boldsymbol{u}) = V(\boldsymbol{E}(\nabla \boldsymbol{u})). \tag{5}$$

One of the most simple objective strain measures is the well known *Green–Saint-Venant strain tensor* $\boldsymbol{E} = \frac{1}{2}[\boldsymbol{F}^T \boldsymbol{F} - I]$. Clearly, this strain measure satisfies the objectivity condition; i.e., $\boldsymbol{E}(\boldsymbol{Q}\boldsymbol{F}) = \boldsymbol{E}(\boldsymbol{F})$ for any given orthonormal (rotation) matrix $\boldsymbol{Q}$. For the most simple *Saint-Venant–Kirchhoff material*, $V(\boldsymbol{E})$ is a quadratic function of $\boldsymbol{E}$; i.e.,

$$V(\boldsymbol{E}) = \tfrac{1}{2}\lambda(\operatorname{tr} \boldsymbol{E})^2 + \mu \operatorname{tr}(\boldsymbol{E})^2, \tag{6}$$

where $\lambda, \mu > 0$ are the classical Lamé constants and $\operatorname{tr} \boldsymbol{E}$ represents the trace of $\boldsymbol{E}$. Therefore, the stored energy $W(\boldsymbol{F})$ is a fourth-order polynomial tensor function of $\boldsymbol{F} = \nabla \boldsymbol{u}$ while for biomaterials the stored energy could be the combination of the polynomial and exponential functions of the Cauchy–Green strain tensor. By using a finite-difference method (FDM), the deformation gradient $\nabla \boldsymbol{u}$ can be directly approximated by the difference $\boldsymbol{D}\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}_i) - \boldsymbol{u}(\boldsymbol{x}_j) = \boldsymbol{u}_i - \boldsymbol{u}_j$ while, in a finite-element method (FEM), the domain $\Omega = \bigcup_e^m \Omega^e$ is discretized by a finite number of elements $\Omega^e \subset \Omega$ and, in each element, the deformation field $\boldsymbol{u}(x) = \sum_i \boldsymbol{N}_i(\boldsymbol{x})\boldsymbol{u}_i$ is numerically represented by the nodal vectors $\boldsymbol{u}_i$ via piecewise interpolation (polynomial) function $\boldsymbol{N}_i(\boldsymbol{x})$ [Gao 1996]. Therefore, by either FDM or FEM, the minimal potential variational problem (4) can be eventually reduced to a very complicated large-scale fourth-order polynomial/exponential minimization problem with the problems ($\mathcal{P}$) as its the most simple case. In the contact mechanics and elastoplastic design of large deformed structures, the nonconvex problems are usually subjected to inequality constraints. In these cases, the global

optimal solution could be local minima (see [Cai et al. 2014]), and to solve such problems is fundamentally difficult by using traditional direct methods.

Canonical duality theory was developed originally by Gao and Strang [1989] for solving general variational problem (4) in finite-deformation theory, where the stored energy $W(\boldsymbol{F})$ is nonconvex and even nonsmooth. By introducing a so-called complementary gap function, they recovered the complementary energy principle in large deformation (geometrically nonlinear) systems. They proved that the nonnegative gap function can be used to identify the global minimizer of the nonconvex potential variational problems. Seven years later, it was discovered that the negative gap function can be used to identify the largest local minimum and maximum. Therefore, a so-called triality theory was first proposed in nonconvex mechanics [Gao 1997] and then generalized to global optimization [Gao 2000a]. This triality theory is composed of a canonical min-max duality and two pairs of double-min, double-max dualities, which reveals an intrinsic duality pattern in complex systems and has been used successfully for solving a wide class of challenging problems in complex systems [Gao 1998; 1999; 2009; Gao and Sherali 2009]. However, it was realized [Gao 2003a; 2003b] that the double-min duality holds under certain additional conditions. Recently, this problem is partly solved for a class of fourth-order polynomial optimization problems [Gao and Wu 2012]. Based on these results, this paper intends to solve the more challenging problem $(\mathcal{P})$. We will show that, by the canonical dual transformation, all critical solutions of $(\mathcal{P})$ can be analytically presented in terms of the canonical dual solutions. The extremality of these solutions can be identified by the triality theory. Several solved examples are listed in the last section.

2. Canonical dual problem and analytical solutions

Following the standard procedure of the canonical dual transformation (e.g., [Gao 2003b]), first we need to choose a geometric operator $\Lambda = (\Lambda_1(\boldsymbol{x}), \Lambda_2(\boldsymbol{x})) : \mathbb{R}^n \to \mathbb{R}^{m+p}$, where

$$\Lambda_1(\boldsymbol{x}) = \left\{ \tfrac{1}{2}\boldsymbol{x}^t \boldsymbol{B}_i \boldsymbol{x} - \alpha_i \right\} : \mathbb{R}^n \to \mathbb{R}^m,$$
$$\Lambda_2(\boldsymbol{x}) = \left\{ \tfrac{1}{2}\boldsymbol{x}^t \boldsymbol{C}_j \boldsymbol{x} - \theta_j \right\} : \mathbb{R}^n \to \mathbb{R}^p.$$

Therefore, the nonconvex function $W(\boldsymbol{x})$ can be written in the canonical form

$$W(\boldsymbol{x}) = V(\Lambda(\boldsymbol{x})) = V_1(\Lambda_1(\boldsymbol{x})) + V_2(\Lambda_2(\boldsymbol{x})) \tag{7}$$

with

$$V_1(\boldsymbol{\epsilon}) = \sum_{i \in I_m} \exp(\epsilon_i) \quad \text{and} \quad V_2(\boldsymbol{\gamma}) = \sum_{j \in I_p} \tfrac{1}{2} b_j \gamma_j^2. \tag{8}$$

Clearly, the canonical function $V(\boldsymbol{\varepsilon})$ is convex on

$$\mathcal{V}_a = \{\boldsymbol{\varepsilon} = (\boldsymbol{\epsilon}, \boldsymbol{\gamma}) \in \mathbb{R}^{m+p} : \epsilon_i \in [-\alpha_i, +\infty),\ \gamma_j \in [-\theta_j, +\infty)\ \forall i \in I_m,\ j \in I_p\} \quad (9)$$

such that the canonical dual variable $\boldsymbol{\varsigma} = (\boldsymbol{\tau}, \boldsymbol{\sigma})$ of $\boldsymbol{\varepsilon} = (\boldsymbol{\epsilon}, \boldsymbol{\gamma})$ can be uniquely defined by

$$\boldsymbol{\varsigma} = \nabla V(\boldsymbol{\varepsilon}) \implies \boldsymbol{\tau} = \nabla V_1(\boldsymbol{\epsilon}) = \{\exp(\epsilon_i)\},\ \boldsymbol{\sigma} = \nabla V_2(\boldsymbol{\gamma}) = \{b_j \gamma_j\}, \quad (10)$$

and on the canonical dual space

$$\mathcal{V}_a^* = \{\boldsymbol{\varsigma} = (\boldsymbol{\tau}, \boldsymbol{\sigma}) \in \mathbb{R}^{m+p} : \tau_i \in [\exp(-\alpha_i), \infty),\ \sigma_j \in [-b_j\theta_j, \infty),\ \forall i \in I_m,\ j \in I_p\},$$

the Legendre conjugate of $V(\boldsymbol{\varepsilon})$ can be defined by

$$V^c(\boldsymbol{\varsigma}) = \text{sta}\{\boldsymbol{\varepsilon}^t\boldsymbol{\varsigma} - V(\boldsymbol{\varepsilon}) : \boldsymbol{\varepsilon} \in \mathcal{V}_a\} = V_1^c(\boldsymbol{\tau}) + V_2^c(\boldsymbol{\sigma}), \quad (11)$$

where sta$\{*\}$ denotes finding stationary points of the function given in $\{*\}$ and

$$V_1^c(\boldsymbol{\tau}) = \sum_{i \in I_m} (\tau_i \ln \tau_i - \tau_i) \quad \text{and} \quad V_2^c(\boldsymbol{\sigma}) = \sum_{j \in I_p} \frac{1}{2b_j}\sigma_j^2. \quad (12)$$

By using the canonical dual transformation $W(\boldsymbol{x}) = V(\Lambda(\boldsymbol{x})) = \Lambda(\boldsymbol{x})^T\boldsymbol{\varsigma} - V^c(\boldsymbol{\varsigma})$, the Gao–Strang *total complementary function* $\Xi : \mathbb{R}^n \times \mathcal{V}_a^* \to \mathbb{R}$ associated with the problem $(\mathcal{P})$ can be given by

$$\begin{aligned}\Xi(\boldsymbol{x}, \boldsymbol{\varsigma}) &= \langle \Lambda(\boldsymbol{x}), \boldsymbol{\varsigma}\rangle - V^c(\boldsymbol{\varsigma}) + \tfrac{1}{2}\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} - \boldsymbol{f}^t\boldsymbol{x} \\ &= \tfrac{1}{2}\boldsymbol{x}^t\boldsymbol{G}(\boldsymbol{\varsigma})\boldsymbol{x} - \boldsymbol{\alpha}^t\boldsymbol{\tau} - \boldsymbol{\theta}^t\boldsymbol{\sigma} - V_1^c(\boldsymbol{\tau}) - V_2^c(\boldsymbol{\sigma}) - \boldsymbol{f}^t\boldsymbol{x}, \end{aligned} \quad (13)$$

where

$$\boldsymbol{G}(\boldsymbol{\varsigma}) = \boldsymbol{A} + \sum_{i \in I_m} \tau_i \boldsymbol{B}_i + \sum_{j \in I_p} \sigma_j \boldsymbol{C}_j. \quad (14)$$

Via this $\Xi(\boldsymbol{x}, \boldsymbol{\varsigma})$, the canonical dual function $\Pi^d : \mathcal{V}_a^* \to \mathbb{R}$ can be defined by

$$\Pi^d(\boldsymbol{\varsigma}) := \text{sta}\{\Xi(\boldsymbol{x}, \boldsymbol{\varsigma}) : \boldsymbol{x} \in \mathbb{R}^n\} = \{\Xi(\boldsymbol{x}(\boldsymbol{\varsigma}), \boldsymbol{\varsigma}) : \nabla_{\boldsymbol{x}}\Xi(\boldsymbol{x}(\boldsymbol{\varsigma}), \boldsymbol{\varsigma}) = 0\}.$$

Notice that $\nabla_{\boldsymbol{x}}\Xi(\boldsymbol{x}, \boldsymbol{\varsigma}) = \boldsymbol{G}(\boldsymbol{\varsigma})\boldsymbol{x} - \boldsymbol{f} = 0$ if and only if

$$\boldsymbol{G}(\boldsymbol{\varsigma})\boldsymbol{x} = \boldsymbol{f}. \quad (15)$$

Let $\mathcal{C}_{ol}(\boldsymbol{G}(\boldsymbol{\varsigma}))$ be the space generated by the columns of the matrix $\boldsymbol{G}(\boldsymbol{\varsigma})$. Then on the dual feasible space

$$\mathcal{S}_a = \{\boldsymbol{\varsigma} \in \mathcal{V}_a^* : \boldsymbol{f} \in \mathcal{C}_{ol}(\boldsymbol{G}(\boldsymbol{\varsigma}))\},$$

the primal solution $\boldsymbol{x} = (\boldsymbol{G}(\boldsymbol{\varsigma}))^{-1}\boldsymbol{f}$ is well defined (if $\boldsymbol{G}(\boldsymbol{\varsigma})$ is singular, $(\boldsymbol{G}(\boldsymbol{\varsigma}))^{-1}$ denotes its pseudoinverse; see [Desoer and Whalen 1963; Peters and Wilkinson

1970] and references therein), and we have $\Pi^d : \mathcal{S}_a \to \mathbb{R}$, where

$$\Pi^d(\boldsymbol{\varsigma}) = -\tfrac{1}{2}\boldsymbol{f}^t(\boldsymbol{G}(\boldsymbol{\varsigma}))^{-1}\boldsymbol{f} - V_1^c(\boldsymbol{\tau}) - V_2^c(\boldsymbol{\sigma}) - \boldsymbol{\alpha}^t\boldsymbol{\tau} - \boldsymbol{\theta}^t\boldsymbol{\sigma}. \tag{16}$$

Therefore, the canonical dual problem is proposed in the form

$$(\mathcal{P}^d) : \operatorname{ext}\{\Pi^d(\boldsymbol{\varsigma}) : \boldsymbol{\varsigma} \in \mathcal{S}_a\}. \tag{17}$$

By the canonical duality theory, it is not difficult to show that

$$\Pi(\boldsymbol{x}) = \operatorname{sta}\{\Xi(\boldsymbol{x}, \boldsymbol{\varsigma}) : \boldsymbol{\varsigma} \in \mathcal{S}_a\} = \Xi(\boldsymbol{x}, \boldsymbol{\varsigma}(\boldsymbol{x})), \tag{18}$$

where $\boldsymbol{\varsigma}(\boldsymbol{x}) = (\boldsymbol{\tau}(\boldsymbol{x}), \boldsymbol{\sigma}(\boldsymbol{x}))$ and

$$\begin{aligned} (\boldsymbol{\tau}(\boldsymbol{x}))_i &= \exp((\Lambda_1(\boldsymbol{x}))_i), \quad i \in I_m, \\ (\boldsymbol{\sigma}(\boldsymbol{x}))_j &= b_j(\Lambda_2(\boldsymbol{x}))_j, \quad j \in I_p. \end{aligned}$$

According to the general theory presented in [Gao 2003b], we have:

Theorem 1 (analytical solutions). *Suppose that for a given $\boldsymbol{f} \in \mathbb{R}^n$ the canonical dual space $\mathcal{S}_a$ is not empty. If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a$ is a stationary point of Π^d, then*

$$\bar{\boldsymbol{x}} = (\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{f} \tag{19}$$

is a stationary point of Π and

$$\Pi(\bar{\boldsymbol{x}}) = \Pi^d(\bar{\boldsymbol{\varsigma}}). \tag{20}$$

Proof. Let us calculate $\nabla\Pi^d(\boldsymbol{\varsigma})$ and $\nabla^2\Pi^d(\boldsymbol{\varsigma})$. We know that

$$\nabla\Pi^d(\boldsymbol{\varsigma}) = \begin{bmatrix} \nabla_{\boldsymbol{\tau}}\Pi^d(\boldsymbol{\varsigma}) \\ \nabla_{\boldsymbol{\sigma}}\Pi^d(\boldsymbol{\varsigma}) \end{bmatrix} \in \mathbb{R}^{m+p};$$

then

$$(\nabla_{\boldsymbol{\tau}}\Pi^d(\boldsymbol{\varsigma}))_i = \tfrac{1}{2}\boldsymbol{f}^t(\boldsymbol{G}(\boldsymbol{\varsigma}))^{-1}\boldsymbol{B}_i(\boldsymbol{G}(\boldsymbol{\varsigma}))^{-1}\boldsymbol{f} - \ln\tau_i - \alpha_i, \quad i \in I_m, \tag{21}$$

$$(\nabla_{\boldsymbol{\sigma}}\Pi^d(\boldsymbol{\varsigma}))_j = \tfrac{1}{2}\boldsymbol{f}^t(\boldsymbol{G}(\boldsymbol{\varsigma}))^{-1}\boldsymbol{C}_j(\boldsymbol{G}(\boldsymbol{\varsigma}))^{-1}\boldsymbol{f} - \frac{\sigma_j}{b_j} - \theta_j, \quad j \in I_p. \tag{22}$$

On the other hand,

$$\nabla^2\Pi^d(\boldsymbol{\varsigma}) = \begin{bmatrix} \nabla^2_{\boldsymbol{\tau}\boldsymbol{\tau}}\Pi^d(\boldsymbol{\varsigma}) & \nabla^2_{\boldsymbol{\tau}\boldsymbol{\sigma}}\Pi^d(\boldsymbol{\varsigma}) \\ \nabla^2_{\boldsymbol{\sigma}\boldsymbol{\tau}}\Pi^d(\boldsymbol{\varsigma}) & \nabla^2_{\boldsymbol{\sigma}\boldsymbol{\sigma}}\Pi^d(\boldsymbol{\varsigma}) \end{bmatrix} \in \mathbb{R}^{m+p} \times \mathbb{R}^{m+p},$$

where $\nabla^2_{\tau\sigma}\Pi^d(\varsigma) := (\nabla_\tau(\nabla_\sigma\Pi^d(\varsigma))^t)$. Let δ_{ij} be the Kronecker delta. Then

$$(\nabla^2_{\tau\tau}\Pi^d(\varsigma))_{ij} = -f^t(G(\varsigma))^{-1}B_i(G(\varsigma))^{-1}B_j(G(\varsigma))^{-1}f - \frac{\delta_{ij}}{\tau_j}, \quad i,j\in I_m,$$
$$(\nabla^2_{\tau\sigma}\Pi^d(\varsigma))_{ij} = -f^t(G(\varsigma))^{-1}B_i(G(\varsigma))^{-1}C_j(G(\varsigma))^{-1}f, \quad i\in I_m,\ j\in I_p,$$
$$(\nabla^2_{\sigma\tau}\Pi^d(\varsigma))_{ij} = -f^t(G(\varsigma))^{-1}C_i(G(\varsigma))^{-1}B_j(G(\varsigma))^{-1}f, \quad i\in I_m,\ j\in I_p,$$
$$(\nabla^2_{\sigma\sigma}\Pi^d(\varsigma))_{ij} = -f^t(G(\varsigma))^{-1}C_i(G(\varsigma))^{-1}C_j(G(\varsigma))^{-1}f - \frac{\delta_{ij}}{b_j}, \quad i,j\in I_p.$$

By making $x = (G(\varsigma))^{-1}f$ and $F(x)\in\mathbb{R}^{n\times(m+p)}$ be $F(x) = [B_1x,\dots,B_mx, C_1x,\dots,C_px]$, we have

$$\nabla^2\Pi^d(\varsigma) = -F(x)^t(G(\varsigma))^{-1}F(x) - \mathrm{Diag}\Big(\frac{1}{\tau_1},\dots,\frac{1}{\tau_m},\frac{1}{b_1},\dots,\frac{1}{b_p}\Big). \tag{23}$$

Let $D = \mathrm{Diag}(\tau_1,\dots,\tau_m,b_1,\dots,b_p)$; then $\nabla^2\Pi^d(\varsigma)$ can be written as

$$\nabla^2\Pi^d(\varsigma) = -F(x)^t(G(\varsigma))^{-1}F(x) - D^{-1}. \tag{24}$$

Calculating $\nabla\Pi(x)$ and $\nabla^2\Pi(x)$, we have respectively

$$\nabla\Pi(x) = \sum_{i\in I_m}\exp\big(\tfrac12x^tB_ix-\alpha_i\big)B_ix + \sum_{j\in I_p}b_j\big(\tfrac12x^tC_jx-\theta_j\big)C_jx + Ax - f, \tag{25}$$

$$\nabla^2\Pi(x) = A + \sum_{i\in I_m}\exp\big(\tfrac12x^tB_ix-\alpha_i\big)(B_ix(B_ix)^t + B_i) + \sum_{j\in I_p}b_j\big(C_jx(C_jx)^t + \big(\tfrac12x^tC_jx-\theta_j\big)C_j\big). \tag{26}$$

Since $\bar\varsigma = (\bar\tau,\bar\sigma)$ is a stationary point of Π^d, then by (21) and (22),

$$(\Lambda_1(\bar x))_i = \ln\bar\tau_i, \quad i\in I_m, \tag{27}$$
$$(\Lambda_2(\bar x))_j = \frac{\bar\sigma_j}{b_j}, \quad j\in I_p. \tag{28}$$

Using (27) and (28) in (25), we obtain

$$\nabla\Pi(\bar x) = G(\bar\varsigma)\bar x - f = G(\bar\varsigma)(G(\bar\varsigma))^{-1}f - f = 0.$$

Notice that (27) and (28) together with (16) and (18) imply that

$$\Pi(\bar x) = \Xi(\bar x,\bar\varsigma) = \Xi((G(\bar\varsigma))^{-1}f,\bar\varsigma) = \Pi^d(\bar\varsigma). \tag{29}$$

And this finishes the proof. □

Remark 1. This theorem shows that the problem $(\mathcal{P}^d)$ is canonical dual to the nonconvex primal problem $(\mathcal{P})$ in the sense that $\Pi(\bar x) = \Pi^d(\bar\varsigma)$ at each critical point

of $\Xi(\boldsymbol{x}, \boldsymbol{\varsigma})$. By the criticality condition (15), we know that, if $\boldsymbol{G}(\boldsymbol{\varsigma})$ is singular at $\bar{\boldsymbol{\varsigma}}$, the canonical equilibrium equation (15) may have an infinite number of solutions: $\bar{\boldsymbol{x}} = \boldsymbol{G}(\bar{\boldsymbol{\varsigma}})^{\dagger}\boldsymbol{f} + \boldsymbol{N}\boldsymbol{x}^o$, where $\boldsymbol{G}^{\dagger}$ represents the Moore–Penrose generalized inverse, $\boldsymbol{N}$ is a basis matrix of the null space of $\boldsymbol{G}(\bar{\boldsymbol{\varsigma}})$, and $\boldsymbol{x}^o$ is a free vector. In this case, Theorem 1 still holds, but the canonical dual function Π^d will have the additional parametric vector $\boldsymbol{x}^o$. In order to avoid this case, a quadratic perturbation method is introduced in [Ruan and Gao 2014b]; i.e., in the case that $\boldsymbol{G}(\bar{\boldsymbol{\varsigma}})$ is singular, replace it by the perturbed form

$$\boldsymbol{G}_{\alpha}(\bar{\boldsymbol{\varsigma}}) = \boldsymbol{G}(\bar{\boldsymbol{\varsigma}}) + \alpha\boldsymbol{D}, \tag{30}$$

where $\alpha > 0$ is a perturbation parameter and $\boldsymbol{D}$ is a given positive-definite matrix. Very often, $\boldsymbol{D} = \boldsymbol{I}$. A detailed study on this quadratic perturbation method is given in [Ruan and Gao 2014b].

In the next section, we will show that the extremality of some of these solutions can be identified by a refined triality theory.

3. Triality theory

Before presenting the refined triality theory, we need the sets

$$\mathcal{S}_a^+ := \{\boldsymbol{\varsigma} \in \mathcal{S}_a : G(\boldsymbol{\varsigma}) \succeq 0\} \quad \text{and} \quad \mathcal{S}_a^- := \{\boldsymbol{\varsigma} \in \mathcal{S}_a : G(\boldsymbol{\varsigma}) \prec 0\}.$$

Lemma 1. *Suppose that $m + p < n$, $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a^-$ is a stationary point and a local minimizer of Π^d, and $\bar{\boldsymbol{x}} = (\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{f}$. Then there exists a matrix $\boldsymbol{L} \in \mathbb{R}^{n\times(m+p)}$ with* $\mathrm{Rank}(\boldsymbol{L}) = m + p$ *such that*

$$\boldsymbol{L}^t\nabla^2\Pi(\bar{\boldsymbol{x}})\boldsymbol{L} \succeq 0. \tag{31}$$

Proof. Since $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a^-$ is a local minimizer of Π^d, we have that $\nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}}) \succeq 0$. It follows from (24) that

$$-\boldsymbol{F}(\bar{\boldsymbol{x}})^t(\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{F}(\bar{\boldsymbol{x}}) \succeq \boldsymbol{D}^{-1} \succ 0.$$

Thus, $\mathrm{Rank}(\boldsymbol{F}(\bar{\boldsymbol{x}})) = m + p$. Since $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a^-$ and $\boldsymbol{F}(\bar{\boldsymbol{x}})\boldsymbol{D}\boldsymbol{F}(\bar{\boldsymbol{x}})^t \succeq 0$, there exists a nonsingular matrix $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ such that

$$\boldsymbol{T}^t\boldsymbol{G}(\bar{\boldsymbol{\varsigma}})\boldsymbol{T} = \mathrm{Diag}(-\lambda_1, \ldots, -\lambda_n), \tag{32}$$

$$\boldsymbol{T}^t\boldsymbol{F}(\bar{\boldsymbol{x}})\boldsymbol{D}\boldsymbol{F}(\bar{\boldsymbol{x}})^t\boldsymbol{T} = \mathrm{Diag}(a_1, \ldots, a_{m_1+m_2}, 0, \ldots, 0), \tag{33}$$

where $\lambda_i > 0$ for every $i = 1, \ldots, n$ and $a_j > 0$ for every $j = 1, \ldots, m + p$ (see [Feng et al. 2012; Horn and Johnson 1985] and references therein). According to Lemma A1 in the Appendix, we know that there exist orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{E} \in \mathbb{R}^{(m+p)\times(m+p)}$ such that

$$\boldsymbol{T}^t\boldsymbol{F}(\bar{\boldsymbol{x}})\boldsymbol{D}^{1/2} = \boldsymbol{U}\boldsymbol{R}\boldsymbol{E}, \tag{34}$$

where $\boldsymbol{R} \in \mathbb{R}^{n\times(m+p)}$ and

$$\boldsymbol{R}_{ij} = \begin{cases} \sqrt{a_i} & \text{if } i = j \text{ and } i = 1, \dots, m+p, \\ 0 & \text{otherwise.} \end{cases}$$

According to the singular value decomposition theory, we know that $\boldsymbol{U}$ is the identity matrix. Then

$$\begin{aligned} \nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}}) &= -\boldsymbol{F}(\bar{\boldsymbol{x}})^t(\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{F}(\bar{\boldsymbol{x}}) - \boldsymbol{D}^{-1} \\ &= -(\boldsymbol{F}(\bar{\boldsymbol{x}})^t\boldsymbol{T})[\boldsymbol{T}^t\boldsymbol{G}(\bar{\boldsymbol{\varsigma}})\boldsymbol{T}]^{-1}(\boldsymbol{T}^t\boldsymbol{F}(\bar{\boldsymbol{x}})) - \boldsymbol{D}^{-1} \\ &= -\boldsymbol{D}^{-1/2}\boldsymbol{E}^t\boldsymbol{R}^t \operatorname{Diag}\Big(-\frac{1}{\lambda_1}, \dots, -\frac{1}{\lambda_n}\Big)\boldsymbol{R}\boldsymbol{E}\boldsymbol{D}^{-1/2} - \boldsymbol{D}^{-1} \succeq 0. \end{aligned}$$

Multiplying by $\boldsymbol{D}^{1/2}$ from the left and the right,

$$\boldsymbol{D}^{1/2}\nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}})\boldsymbol{D}^{1/2} = -\boldsymbol{E}^t\boldsymbol{R}^t \operatorname{Diag}\Big(-\frac{1}{\lambda_1}, \dots, -\frac{1}{\lambda_n}\Big)\boldsymbol{R}\boldsymbol{E} - \boldsymbol{I}_{(m+p)\times(m+p)} \succeq 0.$$

If we multiply the right side of the last equation by $\boldsymbol{E}$ from the left and $\boldsymbol{E}^t$ from the right, we have

$$\begin{aligned} 0 &\preceq -\boldsymbol{R}^t \operatorname{Diag}\Big(-\frac{1}{\lambda_1}, \dots, -\frac{1}{\lambda_n}\Big)\boldsymbol{R} - \boldsymbol{I}_{(m+p)\times(m+p)} \\ &\preceq \operatorname{Diag}\Big(\frac{a_1}{\lambda_1} - 1, \dots, \frac{a_{m+p}}{\lambda_{m+p}} - 1\Big); \end{aligned} \tag{35}$$

thus, $a_i \geq \lambda_i$ for every $i = 1, \dots, m+p$. On the other hand,

$$\begin{aligned} \boldsymbol{T}^t\nabla^2\Pi(\bar{\boldsymbol{x}})\boldsymbol{T} &= \boldsymbol{T}^t\boldsymbol{G}(\bar{\boldsymbol{\varsigma}})\boldsymbol{T} + \boldsymbol{T}^t\boldsymbol{F}(\bar{\boldsymbol{x}})\boldsymbol{D}\boldsymbol{F}(\bar{\boldsymbol{x}})^t\boldsymbol{T} \\ &= \operatorname{Diag}(-\lambda_1, \dots, -\lambda_n) + \operatorname{Diag}(a_1, \dots, a_{m+p}, 0, \dots, 0) \\ &= \operatorname{Diag}(a_1 - \lambda_1, \dots, a_{m+p} - \lambda_{m+p}, -\lambda_{m+p+1}, \dots, -\lambda_n). \end{aligned}$$

Let $\boldsymbol{J} \in \mathbb{R}^{n\times n}$ be defined by

$$J_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and } i = 1, \dots, m+p, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have

$$\boldsymbol{J}^t\boldsymbol{T}^t\nabla^2\Pi(\bar{\boldsymbol{x}})\boldsymbol{T}\boldsymbol{J} = \operatorname{Diag}(a_1 - \lambda_1, \dots, a_{m+p} - \lambda_{m+p}) \succeq 0. \tag{36}$$

Let $\boldsymbol{L} = \boldsymbol{T}\boldsymbol{J}$; clearly $\operatorname{Rank}(\boldsymbol{L}) = m+p$ and $\boldsymbol{L}^t\nabla^2\Pi(\bar{\boldsymbol{x}})\boldsymbol{L} \succeq 0$. □

In a similar way, we can prove the following lemma:

Lemma 2. *Suppose that $m + p > n$, $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_a^-$ is a stationary point Π^d, and $\bar{\boldsymbol{x}} = (\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{f}$ is a local minimizer of Π. Then there exists a matrix $\boldsymbol{Q} \in \mathbb{R}^{(m+p)\times n}$ with $\operatorname{Rank}(\boldsymbol{Q}) = n$ such that*

$$\boldsymbol{Q}^t\nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}})\boldsymbol{Q} \succeq 0. \tag{37}$$

Let the $m+p$ column vectors of $\boldsymbol{L}$ be $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{m+p}$, respectively, and the n column vectors of $\boldsymbol{Q}$ be $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$, respectively. Clearly, $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{m+p}$ are $m+p$ independent vectors and $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ are n independent vectors. Now the subspaces $\mathscr{X}_b$ and $\mathscr{S}_b$ are defined as

$$\mathscr{X}_b = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} = \bar{\boldsymbol{x}} + \sum_{i=1}^{m+p} \upsilon_i \boldsymbol{l}_i,\ \{\upsilon_i\}_{i=1}^{m+p} \subset \mathbb{R} \right\}, \tag{38}$$

$$\mathscr{S}_b = \left\{ \boldsymbol{\varsigma} \in \mathbb{R}^{m+p} : \boldsymbol{\varsigma} = \bar{\boldsymbol{\varsigma}} + \sum_{j=1}^{n} \vartheta_j \boldsymbol{q}_j,\ \{\vartheta_j\}_{j=1}^{n} \subset \mathbb{R} \right\}. \tag{39}$$

Now we are ready to present the refined triality theory.

Theorem 2 (triality theory). *Let $\bar{\boldsymbol{\varsigma}}$ be a stationary point of Π^d and $\bar{\boldsymbol{x}} = (\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{f}$. Assume that* $\det(\nabla^2 \Pi(\bar{\boldsymbol{x}})) \neq 0$.

(i) *If $\bar{\boldsymbol{\varsigma}} \in \mathscr{S}_a^+$, then $\bar{\boldsymbol{\varsigma}}$ is the only global maximizer of Π^d in $\mathscr{S}_a^+$ and $\bar{\boldsymbol{x}}$ is the only global minimizer of Π.*

(ii) *If $\bar{\boldsymbol{\varsigma}} \in \mathscr{S}_a^-$, then $\bar{\boldsymbol{\varsigma}}$ is a local maximizer of Π^d in $\mathscr{S}_a^-$ if and only if $\bar{\boldsymbol{x}}$ is a local maximizer of Π.*

(iii) *If $\bar{\boldsymbol{\varsigma}} \in \mathscr{S}_a^-$ and*:

(a) *If $n = m + p$, then $\bar{\boldsymbol{\varsigma}}$ is a local minimizer of Π^d if and only if $\bar{\boldsymbol{x}}$ is a local minimizer of Π; i.e., there exist neighborhoods $\mathscr{X}, \mathscr{S} \subset \mathbb{R}^n$ of $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{\varsigma}}$, respectively, such that*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x} \in \mathscr{X}} \Pi(\boldsymbol{x}) = \min_{\boldsymbol{\varsigma} \in \mathscr{S}} \Pi^d(\boldsymbol{\varsigma}) = \Pi^d(\bar{\boldsymbol{\varsigma}}). \tag{40}$$

(b) *If $m + p < n$ and $\bar{\boldsymbol{\varsigma}}$ is a local minimizer of Π^d, then $\bar{\boldsymbol{x}}$ is a saddle point of Π and there exist neighborhoods $\mathscr{X}, \mathscr{S} \subset \mathbb{R}^n$ of $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{\varsigma}}$, respectively, such that*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x} \in \mathscr{X} \cap \mathscr{X}_b} \Pi(\boldsymbol{x}) = \min_{\boldsymbol{\varsigma} \in \mathscr{S}} \Pi^d(\boldsymbol{\varsigma}) = \Pi^d(\bar{\boldsymbol{\varsigma}}). \tag{41}$$

(c) *If $n < m + p$ and $\bar{\boldsymbol{x}}$ is a local minimizer of Π, then $\bar{\boldsymbol{\varsigma}}$ is a saddle point of Π^d and there exist neighborhoods $\mathscr{X}, \mathscr{S} \subset \mathbb{R}^n$ of $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{\varsigma}}$, respectively, such that*

$$\Pi(\bar{\boldsymbol{x}}) = \min_{\boldsymbol{x} \in \mathscr{X}} \Pi(\boldsymbol{x}) = \min_{\boldsymbol{\varsigma} \in \mathscr{S} \cap \mathscr{S}_b} \Pi^d(\boldsymbol{\varsigma}) = \Pi^d(\bar{\boldsymbol{\varsigma}}). \tag{42}$$

Proof. (i) Since $\bar{\boldsymbol{\varsigma}} \in \mathscr{S}_a^+$, from (24), it is not difficult to show that Π^d is strictly concave in $\mathscr{S}_a^+$ and $\Xi(\cdot, \bar{\boldsymbol{\varsigma}})$ is strictly convex in $\mathbb{R}^n$ and therefore $\bar{\boldsymbol{\varsigma}}$ must be the only global maximizer of Π^d in $\mathscr{S}_a^+$ and $\bar{\boldsymbol{x}}$ is the only global minimizer of $\Xi(\cdot, \bar{\boldsymbol{\varsigma}})$.

By the definition of Ξ given in (13) and the convexity of V, the Fenchel inequality leads to

$$\Xi(\boldsymbol{x}, \boldsymbol{\varsigma}) \leq \Pi(\boldsymbol{x}) \quad \forall (\boldsymbol{x}, \boldsymbol{\varsigma}) \in \mathbb{R}^n \times \mathcal{S}_a.$$

Let us assume now that there exists a vector $\boldsymbol{x}' \in \mathbb{R}^n \setminus \{\bar{\boldsymbol{x}}\}$ such that $\Pi(\boldsymbol{x}') \leq \Pi(\bar{\boldsymbol{x}})$; then

$$\Pi(\bar{\boldsymbol{x}}) \geq \Pi(\boldsymbol{x}') \geq \Xi(\boldsymbol{x}', \bar{\boldsymbol{\varsigma}}) > \Xi(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\varsigma}}) = \Pi(\bar{\boldsymbol{x}}),$$

where the last equality comes from (29). This contradiction proves that $\bar{\boldsymbol{x}}$ must be the only global minimizer of Π.

(ii) Notice first that using (27) and (28) in (26) we have

$$\nabla^2 \Pi(\bar{\boldsymbol{x}}) = \boldsymbol{G}(\bar{\boldsymbol{\varsigma}}) + \boldsymbol{F}(\bar{\boldsymbol{x}}) \boldsymbol{D} \boldsymbol{F}(\bar{\boldsymbol{x}})^t, \tag{43}$$

where $F(\boldsymbol{x})$ and D are defined in (24). If $\bar{\boldsymbol{\varsigma}}$ is a local maximizer of Π^d in $\mathcal{S}_a^-$, we must have that $\nabla^2 \Pi^d(\bar{\boldsymbol{\varsigma}}) \preceq 0$ from (24), which is equivalent to

$$\boldsymbol{D}^{-1} + \boldsymbol{F}(\bar{\boldsymbol{x}})^t (\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1} \boldsymbol{F}(\bar{\boldsymbol{x}}) \succeq 0. \tag{44}$$

• If $m + p = n$ and $\boldsymbol{F}$ is invertible, multiplying (44) by $(\boldsymbol{F}(\bar{\boldsymbol{x}})^t)^{-1}$ from the left and $(\boldsymbol{F}(\bar{\boldsymbol{x}}))^{-1}$ from the right, we have

$$(\boldsymbol{F}(\bar{\boldsymbol{x}})^t)^{-1} \boldsymbol{D}^{-1} (\boldsymbol{F}(\bar{\boldsymbol{x}}))^{-1} + (\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1} \succeq 0. \tag{45}$$

This is equivalent to

$$(\boldsymbol{F}(\bar{\boldsymbol{x}})^t)^{-1} \boldsymbol{D}^{-1} (\boldsymbol{F}(\bar{\boldsymbol{x}}))^{-1} \succeq -(\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1} \succ 0,$$

which in turn is equivalent to (Lemma A2 in the Appendix)

$$-\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}) \succeq \boldsymbol{F}(\bar{\boldsymbol{x}}) \boldsymbol{D} \boldsymbol{F}(\bar{\boldsymbol{x}})^t \iff \nabla^2 \Pi(\bar{\boldsymbol{x}}) \preceq 0.$$

By assumption, $\det(\nabla^2 \Pi(\bar{\boldsymbol{x}})) \neq 0$; then $\bar{\boldsymbol{x}}$ is a local maximum of Π.

• If $m + p \neq n$ or $\boldsymbol{F}$ is not invertible, then by Lemma A1, there exist orthogonal matrices $\boldsymbol{E} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{K} \in \mathbb{R}^{(m+p) \times (m+p)}$ and a matrix $\boldsymbol{R} \in \mathbb{R}^{n \times (m+p)}$ such that

$$R_{ij} = \begin{cases} s_i & \text{if } i = j \text{ and } i = 1, \ldots, r, \\ 0 & \text{otherwise,} \end{cases}$$

where $s_i > 0$ for every i, $r = \text{Rank}(\boldsymbol{F}(\bar{\boldsymbol{x}}))$, and

$$\boldsymbol{F}(\bar{\boldsymbol{x}}) \boldsymbol{D}^{1/2} = \boldsymbol{E} \boldsymbol{R} \boldsymbol{K}. \tag{46}$$

Using (46), (44) can be rewritten as

$$\boldsymbol{D}^{-1} + \boldsymbol{D}^{-1/2} \boldsymbol{K}^t \boldsymbol{R}^t \boldsymbol{E}^t (\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1} \boldsymbol{E} \boldsymbol{R} \boldsymbol{K} \boldsymbol{D}^{-1/2} \succeq 0.$$

After multiplying this equation by $\boldsymbol{K}\boldsymbol{D}^{1/2}$ from the left and $\boldsymbol{D}^{1/2}\boldsymbol{K}^t$ from the right, we have

$$\boldsymbol{I}_{(m+p)\times(m+p)} + \boldsymbol{R}^t(\boldsymbol{E}^t\boldsymbol{G}(\bar{\boldsymbol{\varsigma}})\boldsymbol{E})^{-1}\boldsymbol{R} \succeq 0.$$

This equation is equivalent to

$$-\boldsymbol{I}_{(m+p)\times(m+p)} - \boldsymbol{R}^t(\boldsymbol{E}^t\boldsymbol{G}(\bar{\boldsymbol{\varsigma}})\boldsymbol{E})^{-1}\boldsymbol{R} \preceq 0.$$

By Lemma A3 in the Appendix, the last equation is equivalent to

$$0 \succeq \boldsymbol{E}^t\boldsymbol{G}(\bar{\boldsymbol{\varsigma}})\boldsymbol{E} + \boldsymbol{R}\boldsymbol{R}^t = \boldsymbol{E}^t\boldsymbol{G}(\bar{\boldsymbol{\varsigma}})\boldsymbol{E} + \boldsymbol{R}(\boldsymbol{K}\boldsymbol{D}^{-1/2}\boldsymbol{D}\boldsymbol{D}^{-1/2}\boldsymbol{K}^t)\boldsymbol{R}^t.$$

Multiplying by $\boldsymbol{E}$ from the left and $\boldsymbol{E}^t$ from the right, we can obtain that

$$0 \succeq \boldsymbol{G}(\bar{\boldsymbol{\varsigma}}) + (\boldsymbol{E}\boldsymbol{R}\boldsymbol{K}\boldsymbol{D}^{-1/2})\boldsymbol{D}(\boldsymbol{D}^{-1/2}\boldsymbol{K}^t\boldsymbol{R}^t\boldsymbol{E}^t) = \boldsymbol{G}(\bar{\boldsymbol{\varsigma}}) + \boldsymbol{F}(\bar{\boldsymbol{x}})\boldsymbol{D}\boldsymbol{F}(\bar{\boldsymbol{x}})^t = \nabla^2\Pi(\bar{\boldsymbol{x}}).$$

By the assumption $\det(\nabla^2\Pi(\bar{\boldsymbol{x}})) \neq 0$, $\bar{\boldsymbol{x}}$ is a local maximum of Π.

Notice that every step of the proof is equivalent, so if $\bar{\boldsymbol{x}}$ is a local maximum of Π, then $\bar{\boldsymbol{\varsigma}}$ must be a local maximum of Π^d.

(iii) Let us consider the three cases:

(a) Assume $n = m + p$. If $\bar{\boldsymbol{\varsigma}}$ is a local minimizer of Π^d, then

$$\nabla^2\Pi^d(\bar{\boldsymbol{\varsigma}}) = -\boldsymbol{F}(\bar{\boldsymbol{x}})^t(\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{F}(\bar{\boldsymbol{x}}) - \boldsymbol{D}^{-1} \succeq 0 \iff -\boldsymbol{F}(\bar{\boldsymbol{x}})^t(\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{F}(\bar{\boldsymbol{x}}) \succeq \boldsymbol{D}^{-1}.$$

This implies that $\mathrm{Rank}(\boldsymbol{F}(\bar{\boldsymbol{x}})) = n$. By multiplying the last inequality by $(\boldsymbol{F}(\bar{\boldsymbol{x}})^t)^{-1}$ from the left and by $(\boldsymbol{F}(\bar{\boldsymbol{x}}))^{-1}$ from the right, we have

$$-(\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1} \succeq (\boldsymbol{F}(\bar{\boldsymbol{x}})^t)^{-1}\boldsymbol{D}^{-1}(\boldsymbol{F}(\bar{\boldsymbol{x}}))^{-1}.$$

By Lemma A2, this is equivalent to

$$-\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}) \preceq \boldsymbol{F}(\bar{\boldsymbol{x}})\boldsymbol{D}\boldsymbol{F}(\bar{\boldsymbol{x}})^t \iff \nabla^2\Pi(\bar{\boldsymbol{x}}) \succeq 0.$$

And since $\det(\nabla^2\Pi(\bar{\boldsymbol{x}})) \neq 0$, $\bar{\boldsymbol{x}}$ is a local minimizer of Π. In a similar way, we can prove the converse.

(b) From (24), we know that

$$-\boldsymbol{F}(\bar{\boldsymbol{x}})^t(\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{F}(\bar{\boldsymbol{x}}) \succeq \boldsymbol{D}^{-1}.$$

Then $-\boldsymbol{F}(\bar{\boldsymbol{x}})^t(\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}))^{-1}\boldsymbol{F}(\bar{\boldsymbol{x}})$ is a nonsingular matrix, and $\mathrm{Rank}(\boldsymbol{F}(\bar{\boldsymbol{x}})) = m + p < n$. We claim now that $\bar{\boldsymbol{x}}$ is not a local minimizer of Π. This is because, if $\bar{\boldsymbol{x}}$ were also a local minimizer, we would have

$$\nabla^2\Pi(\bar{\boldsymbol{x}}) = \boldsymbol{G}(\bar{\boldsymbol{\varsigma}}) + \boldsymbol{F}(\bar{\boldsymbol{x}})\boldsymbol{D}\boldsymbol{F}(\bar{\boldsymbol{x}})^t \succeq 0.$$

Thus,

$$\boldsymbol{F}(\bar{\boldsymbol{x}})\boldsymbol{D}\boldsymbol{F}(\bar{\boldsymbol{x}})^t \succeq -\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}).$$

This implies that

$$n = \text{Rank}(-\boldsymbol{G}(\bar{\varsigma})) = \text{Rank}(\boldsymbol{F}(\bar{\boldsymbol{x}})\boldsymbol{D}\boldsymbol{F}(\bar{\boldsymbol{x}})^t) = m + p,$$

which is a contradiction. Therefore, $\bar{\boldsymbol{x}}$ is a saddle point of Π.

To prove (41), we let $\boldsymbol{L}$ be the matrix as given in Lemma 1 and $\{\boldsymbol{l}_i\}_{i=1}^{m+p}$ be the column vectors of $\boldsymbol{L}$. Define

$$\varphi(t_1, \ldots, t_{m+p}) := \Pi(\bar{\boldsymbol{x}} + t_1\boldsymbol{l}_1 + \cdots + t_{m+p}\boldsymbol{l}_{m+p}).$$

We need to show that $(0, \ldots, 0) \in \mathbb{R}^{m+p}$ is a local minimizer of the function φ. Notice that

$$\nabla\varphi(0, \ldots, 0) = \boldsymbol{L}^t\nabla\Pi(\bar{\boldsymbol{x}}) = 0$$

and

$$\nabla^2\varphi(0, \ldots, 0) = \boldsymbol{L}^t\nabla^2\Pi(\bar{\boldsymbol{x}})\boldsymbol{L} \succeq 0,$$

which is a consequence of Lemma 1. Furthermore, from (36) we have that

$$\nabla^2\varphi(0, \ldots, 0) = \text{Diag}(a_1 - \lambda_1, \ldots, a_{m+p} - \lambda_{m+p}),$$

and since $\det(\nabla^2\Pi(\bar{\boldsymbol{x}})) \neq 0$, it can be proven that $a_i > \lambda_i$ for every i.

(c) The proof is similar to that of part (b). □

Remark 2. Theorem 2 shows that, in order to solve the problem $(\mathcal{P})$ by means of the canonical duality theory, a necessary condition is that the problem $(\mathcal{P})$ should have a unique solution. It was indicated in [Ruan and Gao 2014b] that, if the nonconvex minimization problem has more than one global minimizer, it could be NP-hard. In order to solve this type of problems, the perturbation methods should be used.

Remark 3. The triality theory states precisely that, if $\boldsymbol{\varsigma}$ is a global maximizer of Π^d on a certain set, then $\boldsymbol{x}$ is a global minimizer for Π. This is known from the general result by Gao and Strang [1989]. If $\boldsymbol{\varsigma}$ is a local maximizer for Π^d, then $\boldsymbol{x}$ is also a local maximizer for Π. This is the so-called double-max duality statement. If $\boldsymbol{\varsigma}$ is a local minimizer for Π^d, then $\boldsymbol{x}$ is also a local minimizer for Π in certain directions. This is the so-called double-min duality in the standard triality form proposed in [Gao 2000b]. The triality theory was first discovered in nonconvex mechanics [Gao 1997]. Gao [2003a; 2003b] realized that the double-min duality holds under certain additional condition, which was left as an open problem. Recently, this open problem was solved for the quartic polynomial optimization problem [Gao and Wu 2012]. This result is now generalized to the general nonconvex problem $(\mathcal{P})$. Part (iii) of Theorem 2 shows that, if $m + p = n$, then $\boldsymbol{\varsigma}$ is a local minimizer if and only if $\boldsymbol{x}$ is also a local minimizer. In other cases, either $\boldsymbol{x}$ is a saddle point of Π or $\boldsymbol{\varsigma}$ is a saddle point of Π^d.

Remark 4. The canonical duality-triality theory has been challenged recently by C. Zălinescu and his coworkers R. Strugariu and M. D. Voisei in several papers (see [Strugariu et al. 2011]). By listing some simple "counterexamples", they claimed that this theory is false. Unfortunately, most of these counterexamples are not new and were first discovered by Gao [2003a; 2003b], who was never cited. Some of their "counterexamples" are fundamentally wrong; i.e. they incorrectly choose linear functions as the stored energy and nonlinear functions as external energy (see [Voisei and Zălinescu 2011]). These conceptual mistakes show a big gap between mathematics and mechanics.

4. Numerical examples

In the following examples, $m = p = 1$ and $b_1 = 1$. The graphs provided and the numerical results were obtained using MAXIMA [2010].

4.1. ***One stationary point in*** $\mathcal{S}_a^+$**.** First, we consider the case that the primal function has a unique solution. We let $\alpha_1 = \theta_1 = 1$ and

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \boldsymbol{B}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \boldsymbol{C}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{f} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Clearly, the function $\Pi : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$\Pi(x, y) = \exp\left(\tfrac{1}{2}(x^2 + 2y^2) - 1\right) + \tfrac{1}{2}\left(\tfrac{1}{2}(x^2 + y^2) - 1\right)^2 + \tfrac{1}{2}(x^2 - y^2) - x - y,$$

and the dual function has the form of

$$\Pi^d(\tau, \sigma) = -\frac{1}{2}\left(\frac{1}{1+\tau+\sigma} + \frac{1}{2\tau+\sigma-1}\right) - \tau \cdot \ln(\tau) - \tfrac{1}{2}\sigma^2 - \sigma.$$

It can be shown that Π^d has only one critical point in $\mathcal{S}_a^+$ and it is given (approximately) by

$$\bar{\boldsymbol{\varsigma}} = (1.171057661103504, -0.34599084656216).$$

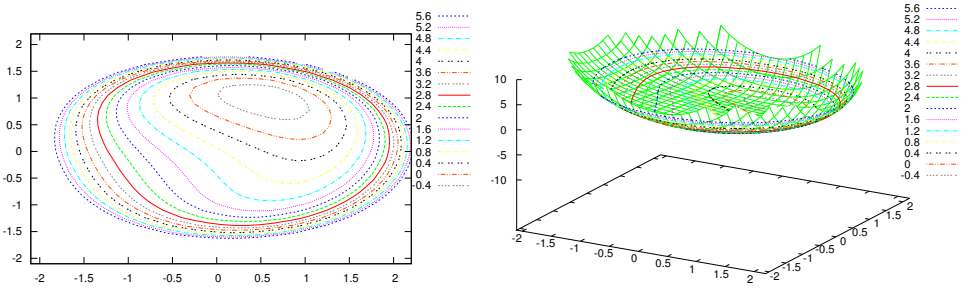


Figure 1. Contours and graph of the primal function Π of Section 4.1.

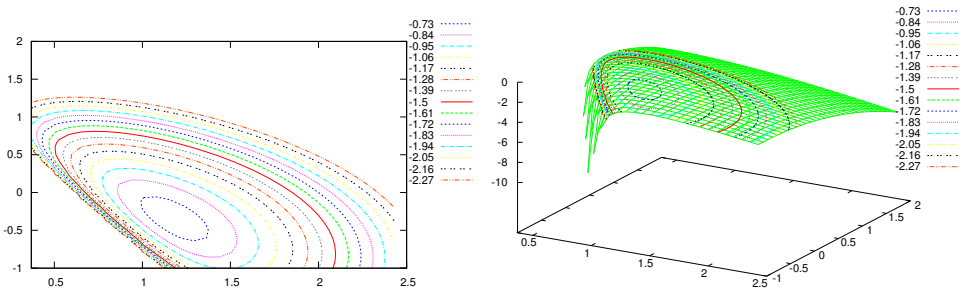


Figure 2. Contours and graph of the dual function Π^d of Section 4.1.

By the triality theory, the vector

$$\bar{\boldsymbol{x}} = \boldsymbol{G}(\bar{\varsigma})^{-1}\boldsymbol{f} = (0.54792514555217, 1.003890602479819)$$

is the only global minimizer of the primal problem.

4.2. ***One stationary point in*** $\mathcal{S}_a^+$ ***and one in*** $\mathcal{S}_a^-$**.** Let $\alpha_1 = 1$, $\theta_1 = 50$, and

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & -16 \end{bmatrix}, \quad \boldsymbol{B}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{C}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{f} = \begin{bmatrix} -25 \\ 9 \end{bmatrix}.$$

The primal function $\Pi : \mathbb{R}^2 \to \mathbb{R}$ is then given by

$$\Pi(x, y) = \exp\left(\tfrac{1}{2}(x^2+y^2)-1\right) + \tfrac{1}{2}\left(\tfrac{1}{2}(x^2+2y^2)-50\right)^2 + \tfrac{1}{2}(x^2-16y^2) + 25x - 9y,$$

and its canonical dual is

$$\Pi^d(\tau, \sigma) = -\frac{1}{2}\left(\frac{81}{-16+\tau+2\sigma} + \frac{625}{1+\tau+\sigma}\right) - \tau \cdot \ln(\tau) - \tfrac{1}{2}\sigma^2 - 50\sigma,$$

which has two critical points:

$$\overline{\varsigma_1} = (96.61711963278241, -38.94928057661689) \in \mathcal{S}_a^+,$$
$$\overline{\varsigma_2} = (\ 0.42157060067968, -49.86072154366873) \in \mathcal{S}_a^-.$$

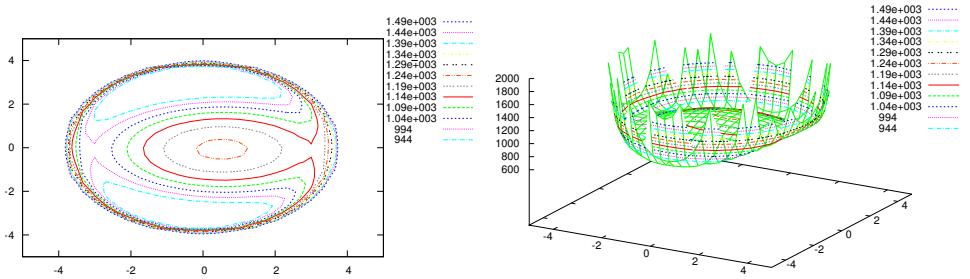


Figure 3. Contours and graph of the primal function Π of Section 4.2.

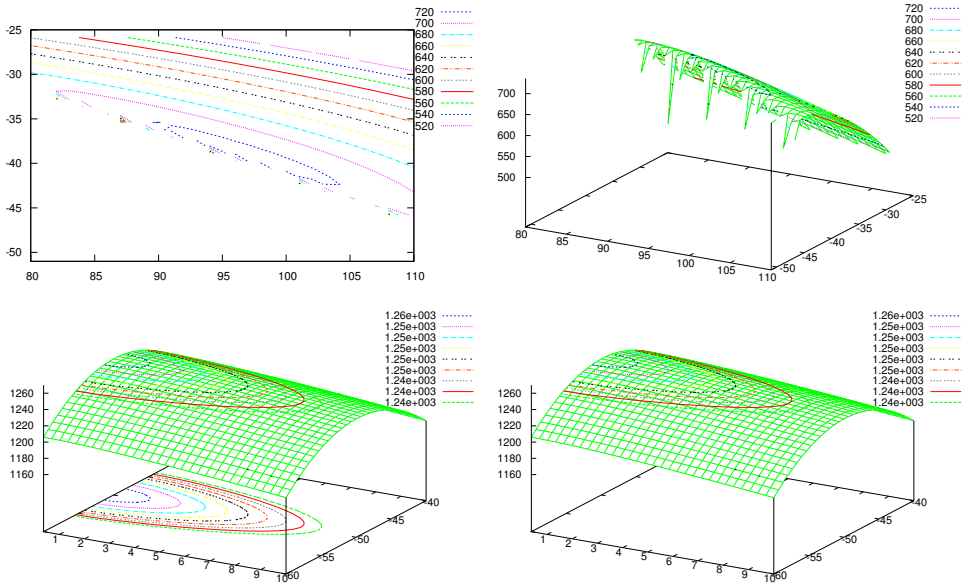


Figure 4. Contours and graph of the dual function Π^d of Section 4.2 in $\mathcal{S}_a^+$ (top) and in $\mathcal{S}_a^-$ (bottom).

Therefore, by the triality theory, the associated vector

$$\overline{\boldsymbol{x}_1} = \boldsymbol{G}(\overline{\boldsymbol{\varsigma}_1})^{-1}\boldsymbol{f} = (-0.42612784793499,\, 3.310578038951848)$$

is the only global minimizer of $\Pi(\boldsymbol{x})$, and

$$\overline{\boldsymbol{x}_2} = (0.51611144112381,\, -0.078057328303129)$$

is a local maximizer (see Figure 3) since $\overline{\boldsymbol{\varsigma}_2}$ is a local maximum of Π^d in $\mathcal{S}_a^-$ (see Figure 4, bottom).

4.3. ***One stationary point in $\mathcal{S}_a^+$ and two in $\mathcal{S}_a^-$.*** In order to illustrate the triality theory, we let $\alpha_1 = \theta_1 = 2$ and

$$\boldsymbol{A} = \begin{bmatrix} -16 & 0 \\ 0 & -4 \end{bmatrix}, \quad \boldsymbol{B}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{C}_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{f} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

Accordingly, we have

$$\Pi(x, y) = \exp\left(\tfrac{1}{2}x^2 - 2\right) + \tfrac{1}{2}\left(\tfrac{1}{2}y^2 - 2\right)^2 + \tfrac{1}{2}(-16x^2 - 4y^2) - 2x - 2y,$$

$$\Pi^d(\tau, \sigma) = -\frac{1}{2}\left(\frac{4}{\sigma - 4} + \frac{4}{\tau - 16}\right) - \tau \cdot \ln(\tau) - \tau - \tfrac{1}{2}\sigma^2 - 2\sigma.$$

In this case, Π^d has in total six critical points but only one in $\mathcal{S}_a^+$,

$$\overline{\boldsymbol{\varsigma}_1} = (16.64468576727409,\, 4.552474610531074) \in \mathcal{S}_a^+$$

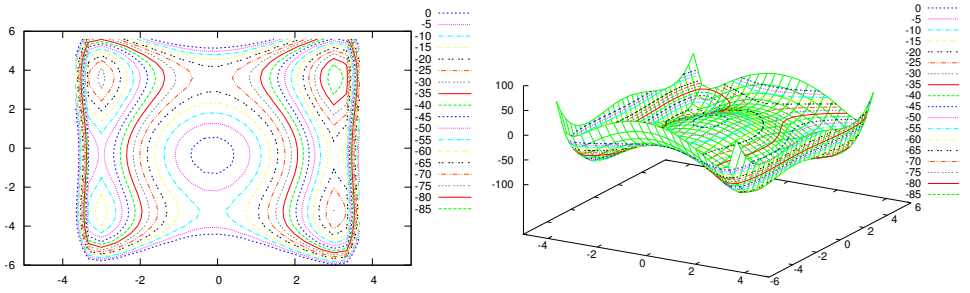


Figure 5. Contours and graph of the primal function Π of Section 4.3.

(see Figure 6, top), and two in $\mathcal{S}_a^-$:

$$\overline{\varsigma_2} = (0.13641513779858, -1.943380912562619) \in \mathcal{S}_a^-,$$
$$\overline{\varsigma_3} = (15.34981976568548, 3.390906302031545) \in \mathcal{S}_a^-.$$

From Figure 6, bottom, we can see that $\overline{\varsigma_2}$ is a local maximizer and $\overline{\varsigma_3}$ is a local minimizer of Π^d. Therefore, by the triality theory, we know that

$$\overline{\boldsymbol{x}_1} = \boldsymbol{G}(\overline{\boldsymbol{\varsigma}_1})^{-1}\boldsymbol{f} = (3.102286573591542, 3.620075858467906)$$

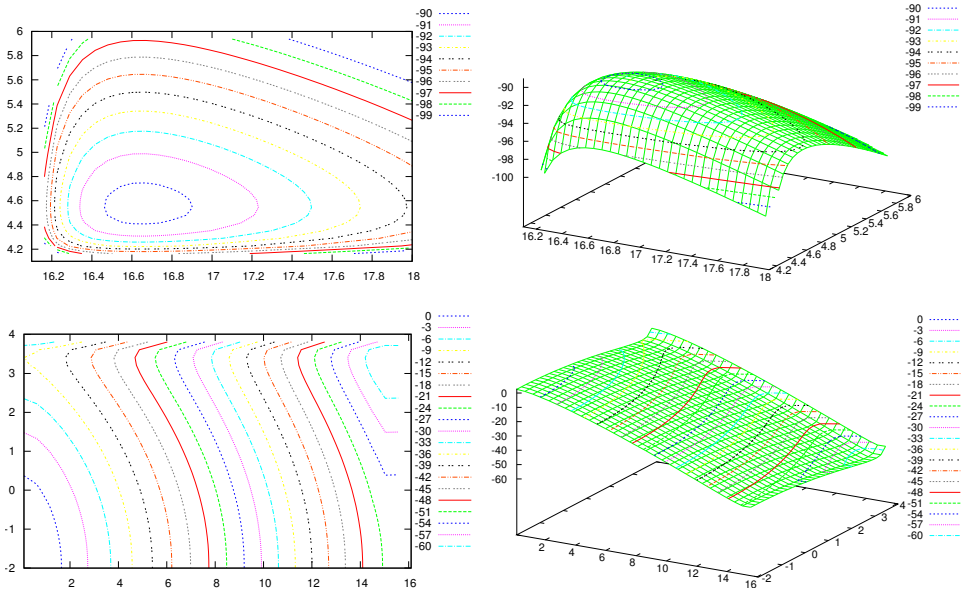


Figure 6. Contours and graph of the dual function Π^d of Section 4.3 in $\mathcal{S}_a^+$ (top) and in $\mathcal{S}_a^-$ (bottom).

is the only global minimizer,

$$\overline{\boldsymbol{x}_2} = (-0.12607490787063, -0.33650880356205)$$

is a local maximizer, and

$$\overline{\boldsymbol{x}_3} = (-3.076070133243102, -3.283567054905852)$$

is a local minimizer of $\Pi(\boldsymbol{x})$ (see Figure 5).

4.4. ***Nonunique global minima.*** In the case that no stationary point can be found in $\mathcal{S}_a^+$, the primal problem could have more than one global minima. To see this, we let $\boldsymbol{f} \equiv 0$, $\alpha_1 = \theta_1 = 2$, and

$$\boldsymbol{A} \equiv 0, \quad \boldsymbol{B}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{C}_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

In this case, the primal function

$$\Pi(x, y) = \exp\left(\tfrac{1}{2}x^2 - 2\right) + \tfrac{1}{2}\left(\tfrac{1}{2}y^2 - 2\right)^2$$

has two global minimums at $(0, -2)$ and $(0, 2)$ and a local maximum at $(0, 0)$. While the dual function

$$\Pi^d(\tau, \sigma) = -\tau \ln \tau - \tau - \tfrac{1}{2}\sigma^2 - 2\sigma$$

does not have a stationary point in $\mathcal{S}_a^+$. There is however a critical point in the boundary of $\mathcal{S}_a^+$, namely $\bar{\boldsymbol{\varsigma}} = (\exp(-2), 0)$. By defining $\bar{\boldsymbol{x}} = \boldsymbol{G}(\bar{\boldsymbol{\varsigma}})^{-1}\boldsymbol{f}$, we have that $\bar{\boldsymbol{x}} = (0, 0)$.

In order to find a global minimum of Π, we need to introduce the perturbations

$$\boldsymbol{A}_n = \begin{bmatrix} -\frac{16}{n} & 0 \\ 0 & -\frac{4}{n} \end{bmatrix} \quad \text{and} \quad \boldsymbol{f}_n = \begin{bmatrix} \frac{2}{n} \\ \frac{2}{n} \end{bmatrix} \quad \text{for every } n \in \mathbb{N}.$$

Then the associated primal and dual functions are

$$\Pi_n(x, y) = \exp\left(\tfrac{1}{2}x^2 - 2\right) + \tfrac{1}{2}\left(\tfrac{1}{2}y^2 - 2\right)^2 + \tfrac{1}{2}\left(-\tfrac{16}{n}x^2 - \tfrac{4}{n}y^2\right) - \tfrac{2}{n}x - \tfrac{2}{n}y,$$

$$\Pi_n^d(\tau, \sigma) = -\frac{1}{2}\left(\frac{4}{n^2\left(\tau - \frac{16}{n}\right)} + \frac{4}{n^2\left(\sigma - \frac{4}{n}\right)}\right) - \tau \ln \tau + \tau - \tfrac{1}{2}\sigma^2 - 2\tau - 2\sigma.$$

Notice that if $n = 1$ we are in the case presented in Section 4.3. Let us show that for sufficiently large values of n we can find a stationary point for Π_n^d in $\mathcal{S}_a^+$, namely $\overline{\boldsymbol{\varsigma}_n}$. Furthermore, by defining $\bar{\boldsymbol{x}}_n = \boldsymbol{G}(\bar{\boldsymbol{\varsigma}}_n)^{-1}\boldsymbol{f}_n$, we will have a convergent sequence.

Let us calculate the gradient of Π_n^d:

$$\nabla \Pi_n^d(\tau, \sigma) = \begin{bmatrix} -2 - \ln \tau + \frac{2}{(n\tau - 16)^2} \\ -\sigma - 2 + \frac{2}{(n\sigma - 4)^2} \end{bmatrix}.$$

Let $h(\tau) = -2 - \ln\tau + 2/(n\tau - 16)^2$ and $g(\sigma) = -\sigma - 2 + 2/(n\sigma - 4)^2$. It is not difficult to show that there exists a sufficiently large $N \in \mathbb{N}$ such that, if $n > N$,

- $n \cdot \exp\big(-2 + \frac{1}{n}\big) - 16$ and $n \cdot \exp(-2) - 16$ are positive numbers,
- $h\big(\exp\big(-2 + \frac{1}{n}\big)\big) = \dfrac{2}{\big(n \cdot \exp\big(-2 + \frac{1}{n}\big) - 16\big)^2} - \frac{1}{n} < 0 < h(\exp(-2)) = \dfrac{2}{(n \cdot \exp(-2) - 16)^2}$,
- $g\big(\frac{5.1}{n}\big) \approx -\frac{5.1}{n} - 0.34710743801 < 0 < g\big(\frac{4.9}{n}\big) \approx 0.46913580247 - \frac{4.9}{n}$.

Based on these results, we know that, for every $n > N$, $\nabla\Pi_n^d$ has a stationary point $\bar{\boldsymbol{\varsigma}}_n = (\bar\tau_n, \bar\sigma_n) \in \big[\exp(-2), \exp\big(-2 + \frac{1}{n}\big)\big] \times \big[\frac{4.9}{n}, \frac{5.1}{n}\big]$. Moreover, since $g(\bar\sigma_n) = 0$, it is easy to obtain $\lim_{n\to+\infty} n \cdot \bar\sigma_n = 5$.

Notice also that

$$\boldsymbol{G}(\bar{\boldsymbol{\varsigma}}_n) = \begin{bmatrix} \bar\tau_n - \frac{16}{n} & 0 \\ 0 & \bar\sigma_n - \frac{4}{n} \end{bmatrix}$$

is positive definite. Therefore, the perturbed solution can be obtained as

$$\bar{\boldsymbol{x}}_n = \boldsymbol{G}(\bar{\boldsymbol{\varsigma}}_n)^{-1} \boldsymbol{f}_n = \begin{bmatrix} 2/(n \cdot \bar\tau_n - 16) \\ 2/(n \cdot \bar\sigma_n - 4) \end{bmatrix}.$$

Since $\bar\tau_n \in \big[\exp(-2), \exp\big(-2 + \frac{1}{n}\big)\big]$, we have $\lim_{n\to+\infty} \tau_n = \exp(-2)$. From the fact that $\lim_{n\to+\infty} n \cdot \bar\sigma_n = 5$, we get

$$\lim_{n\to+\infty} \bar{\boldsymbol{x}}_n = \begin{bmatrix} 0 \\ 2 \end{bmatrix},$$

which is a solution of Π.

Canonical perturbation method was originally introduced in [Ruan et al. 2010] for solving nonconvex polynomial minimization problems. This method has been used successfully in integer programming and network communication (see [Gao et al. 2012; Wang et al. 2012]).

5. Future research

Some open questions that will be studied in the future are the following:

- As stated in Remark 1, in order to use the canonical dual transformation, a necessary condition is that $(\mathcal{P})$ has a unique solution. Is this also a sufficient condition? In other words, given $(\mathcal{P})$ such that it has a unique solution, can we find a stationary point of Π^d in $\mathcal{S}_a^+$?
- Section 4.4 shows an interesting perturbation method that allows us to solve a problem when the necessary condition of Remark 1 is not satisfied. Can we generalize this method and develop an algorithm?

Appendix A: Some lemmas in matrix analysis

The following results are needed in the proofs of Section 2:

Lemma A1 (singular-value decomposition [Horn and Johnson 1985]). *For any given matrix* $\boldsymbol{M} \subset \mathbb{R}^{m\times n}$ *with* $\operatorname{Rank}(\boldsymbol{M}) = r$, *there exist* $\boldsymbol{U} \subset \mathbb{R}^{m\times m}$, $\boldsymbol{R} \subset \mathbb{R}^{m\times n}$, *and* $\boldsymbol{E} \subset \mathbb{R}^{n\times n}$ *such that*

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{R}\boldsymbol{E},$$

where $\boldsymbol{U}$ *and* $\boldsymbol{E}$ *are orthogonal matrices*, *and*

$$R_{ij} = \begin{cases} s_i & \text{if } i = j \text{ and } i = 1, \dots, r, \\ 0 & \text{if } i \neq j, \end{cases}$$

where $s_i > 0$ *for every* $i = 1, \dots, r$.

Lemma A2 [Horn and Johnson 1985]. *If* $\boldsymbol{G}$ *and* $\boldsymbol{U}$ *are positive-definite matrices in* $\mathbb{R}^{n\times n}$, *then* $\boldsymbol{G} \succeq \boldsymbol{U}$ *if and only if* $\boldsymbol{U}^{-1} \succeq \boldsymbol{G}^{-1}$.

Lemma A3 [Gao and Wu 2012]. *Suppose* $\boldsymbol{P}$, $\boldsymbol{U}$, *and* $\boldsymbol{D}$ *are three matrices in* $\mathbb{R}^{n\times n}$ *such that*

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & 0_{m\times(n-m)} \\ 0_{(n-m)\times n} & 0_{(n-m)\times(n-m)} \end{bmatrix},$$

where $\boldsymbol{D}_{11} \in \mathbb{R}^{m\times m}$ *is nonsingular and*

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22} \end{bmatrix} \prec 0 \quad \text{and} \quad \boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_{11} & 0_{m\times(n-m)} \\ 0_{(n-m)\times m} & \boldsymbol{U}_{22} \end{bmatrix} \succ 0,$$

where $\boldsymbol{P}_{ij}$ *and* $\boldsymbol{U}_{ii}$ *are appropriate-dimensional matrices for* $i, j = 1, 2$. *Then*

$$\boldsymbol{P} + \boldsymbol{D}\boldsymbol{U}\boldsymbol{D}^t \preceq 0 \iff -\boldsymbol{D}^t\boldsymbol{P}^{-1}\boldsymbol{D} - \boldsymbol{U}^{-1} \preceq 0.$$

Acknowledgements

This research is supported by the US Air Force Office of Scientific Research under the grant AFOSR FA9550-10-1-0487. Comments and suggestions from the editor and reviewers are sincerely acknowledged.

References

[Aspnes et al. 2004] J. Aspnes, D. Goldberg, and Y. R. Yang, "On the computational complexity of sensor network localization", pp. 32–44 in *Algorithmic aspects of wireless sensor networks* (Turku, 2004), edited by S. Nikoletseas and J. D. P. Rolim, Lecture Notes in Computer Science **3121**, Springer, Berlin, 2004.

[Cai et al. 2014] K. Cai, D. Y. Gao, and Q. H. Qin, "Post-buckling solutions of hyper-elastic beam by canonical dual finite element method", *Math. Mech. Solids* **19**:6 (2014), 659–671. arXiv 1302.4136

[Desoer and Whalen 1963] C. A. Desoer and B. H. Whalen, "A note on pseudoinverses", *J. Soc. Indust. Appl. Math.* **11** (1963), 442–447.

[Feng et al. 2012] J.-M. Feng, G.-X. Lin, R.-L. Sheu, and Y. Xia, "Duality and solutions for quadratic programming over single non-homogeneous quadratic constraint", *J. Glob. Optim.* **54**:2 (2012), 275–293.

[Gao 1996] D. Y. Gao, "Complementary finite-element method for finite deformation nonsmooth mechanics", *J. Eng. Math.* **30**:3 (1996), 339–353.

[Gao 1997] D. Y. Gao, "Dual extremum principles in finite deformation theory with applications to post-buckling analysis of extended nonlinear beam theory", *Appl. Mech. Rev.* (*ASME*) **50**:11S (1997), S64–S71.

[Gao 1998] D. Y. Gao, "Duality, triality and complementary extremum principles in non-convex parametric variational problems with applications", *IMA J. Appl. Math.* **61**:3 (1998), 199–235.

[Gao 1999] D. Y. Gao, "General analytic solutions and complementary variational principles for large deformation nonsmooth mechanics", *Meccanica* (*Milano*) **34**:3 (1999), 169–198.

[Gao 2000a] D. Y. Gao, "Canonical dual transformation method and generalized triality theory in nonsmooth global optimizaton", *J. Glob. Optim.* **17**:1-4 (2000), 127–160.

[Gao 2000b] D. Y. Gao, *Duality principles in nonconvex systems: theory, methods and applications*, Nonconvex Optimization and its Applications **39**, Kluwer, Dordrecht, 2000.

[Gao 2003a] D. Y. Gao, "Nonconvex semi-linear problems and canonical duality solutions", pp. 261–312 in *Advances in mechanics and mathematics*, vol. 2, edited by D. Y. Gao and R. W. Ogden, Adv. Mech. Math. **4**, Kluwer, Boston, 2003.

[Gao 2003b] D. Y. Gao, "Perfect duality theory and complete solutions to a class of global optimization problems", *Optimization* **52**:4-5 (2003), 467–493.

[Gao 2009] D. Y. Gao, "Canonical duality theory: theory, method, and applications in global optimization", *Comput. Chem. Eng.* **33**:12 (2009), 1964–1972.

[Gao and Ogden 2008a] D. Y. Gao and R. W. Ogden, "Multiple solutions to non-convex variational problems with implications for phase transitions and numerical computation", *Quart. J. Mech. Appl. Math.* **61**:4 (2008), 496–522.

[Gao and Ogden 2008b] D. Y. Gao and R. W. Ogden, "Closed-form solutions, extremality and nonsmoothness criteria in a large deformation elasticity problem", *Z. Angew. Math. Phys.* **59**:3 (2008), 498–517.

[Gao and Ruan 2008] D. Y. Gao and N. Ruan, "Solutions and optimality criteria for nonconvex quadratic-exponential minimization problem", *Math. Methods Oper. Res.* **67**:3 (2008), 479–491.

[Gao and Sherali 2009] D. Y. Gao and H. D. Sherali, "Canonical duality theory: connections between nonconvex mechanics and global optimization", pp. 257–326 in *Advances in applied mathematics and global optimization*, edited by D. Y. Gao and H. D. Sherali, Adv. Mech. Math. **17**, Springer, New York, 2009.

[Gao and Strang 1989] D. Y. Gao and G. Strang, "Geometric nonlinearity: potential energy, complementary energy, and the gap function", *Quart. Appl. Math.* **47**:3 (1989), 487–504.

[Gao and Wu 2012] D. Y. Gao and C. Wu, "On the triality theory for a quartic polynomial optimization problem", *J. Ind. Manag. Optim.* **8**:1 (2012), 229–242.

[Gao and Yu 2008] D. Y. Gao and H. Yu, "Multi-scale modelling and canonical dual finite element method in phase transitions of solids", *Int. J. Solids Struct.* **45**:13 (2008), 3660–3673.

[Gao et al. 2012] D. Y. Gao, N. Ruan, and P. M. Pardalos, "Canonical dual solutions to sum of fourth-order polynomials minimization problems with applications to sensor network localization",

pp. 37–54 in *Sensors: theory, algorithms, and applications*, edited by V. L. Boginski et al., Springer Optim. Appl. **61**, Springer, New York, 2012.

[Horn and Johnson 1985] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, 1985.

[Marsden and Hughes 1983] J. E. Marsden and T. J. R. Hughes, *Mathematical foundations of elasticity*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[MAXIMA 2010] MAXIMA, "Maxima, a computer algebra system", 2010, Available at http://maxima.sourceforge.net. Version 5.22.1.

[Moré and Wu 1997] J. J. Moré and Z. Wu, "Global continuation for distance geometry problems", *SIAM J. Optim.* **7**:3 (1997), 814–836.

[Moreau 1968] J.-J. Moreau, "La notion de sur-potentiel et les liaisons unilatérales en élastostatique", *C. R. Acad. Sci. Paris Sér. A-B* **267A** (1968), A954–A957.

[Moreau et al. 1988] J.-J. Moreau, P. D. Panagiotopoulos, and G. Strang (editors), *Topics in nonsmooth mechanics*, Birkhäuser, Basel, 1988.

[Peters and Wilkinson 1970] G. Peters and J. H. Wilkinson, "The least squares problem and pseudo-inverses", *Comput. J.* **13**:3 (1970), 309–316.

[Ruan and Gao 2014a] N. Ruan and D. Y. Gao, "Canonical duality approach for non-linear dynamical systems", *IMA J. Appl. Math.* **79**:2 (2014), 313–325.

[Ruan and Gao 2014b] N. Ruan and D. Y. Gao, "Global optimal solutions to a general sensor network localization problem", *Perform. Eval.* **75–76** (2014), 1–16.

[Ruan et al. 2010] N. Ruan, D. Y. Gao, and Y. Jiao, "Canonical dual least square method for solving general nonlinear systems of quadratic equations", *Comput. Optim. Appl.* **47**:2 (2010), 335–347.

[Santos and Gao 2012] H. A. F. A. Santos and D. Y. Gao, "Canonical dual finite element method for solving post-buckling problems of a large deformation elastic beam", *Int. J. Nonlinear Mech.* **47**:2 (2012), 240–247.

[Saxe 1979] J. B. Saxe, "Embeddability of weighted graphs in k-space is strongly NP-hard", pp. 480–489 in *Proceedings of the 17th Allerton Conference on Communications, Control, and Computing* (Monticello, IL, 1979), University of Illinois at Urbana-Champaign, Urbana, IL, 1979. Also in *Two papers on graph embedding problems*, preprint, Carnegie Mellon University, Pittsburgh, PA, 1980.

[Strugariu et al. 2011] R. Strugariu, M. D. Voisei, and C. Zălinescu, "Counter-examples in bi-duality, triality and tri-duality", *Discrete Contin. Dyn. Syst.* **31**:4 (2011), 1453–1468.

[Voisei and Zălinescu 2011] M. D. Voisei and C. Zălinescu, "Some remarks concerning Gao–Strang's complementary gap function", *Appl. Anal.* **90**:6 (2011), 1111–1121.

[Wang et al. 2012] Z. Wang, S.-C. Fang, D. Y. Gao, and W. Xing, "Canonical dual approach to solving the maximum cut problem", *J. Glob. Optim.* **54**:2 (2012), 341–351.

[Xie and Schlick 2000] D. Xie and T. Schlick, "Visualization of chemical databases using the singular value decomposition and truncated-Newton minimization", pp. 267–286 in *Optimization in computational chemistry and molecular biology: local and global approaches*, edited by C. A. Floudas and P. M. Pardalos, Optimization in Computational Chemistry and Molecular Biology **40**, Springer, Dordrecht, 2000.

[Zhang et al. 2011] J. Zhang, D. Y. Gao, and J. Yearwood, "A novel canonical dual computational approach for prion AGAAAAGA amyloid fibril molecular modeling", *J. Theoret. Biol.* **284** (2011), 149–157.

Received 22 Apr 2013. Revised 13 May 2014. Accepted 29 Jul 2014.

DANIEL MORALES-SILVA: d.moralessilva@federation.edu.au
School of Science, Information Technology and Engineering, Federation University, Mt. Helen VIC 3353, Australia

DAVID Y. GAO: d.gao@federation.edu.au
david.gao@anu.edu.au
School of Science, Information Technology and Engineering, Federation University, Mt. Helen, VIC 3353, Australia

and

Research School of Engineering, Australian National University, Canberra, ACT 0200, Australia

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 3, No. 2, 2015
dx.doi.org/10.2140/memocs.2015.3.163

NEUTRALITY OF ECCENTRICALLY COATED ELASTIC INCLUSIONS

XU WANG AND PETER SCHIAVONE

In the analysis of neutral coated circular holes in an isotropic medium, it is well-known that neutrality to a general class of applied uniform fields can be realized only by the *concentrically coated circle* construction. It is of interest to examine to what degree *eccentric* circular coatings can be used to achieve effective or near-neutrality in the presence of a wider and more general class of applied fields. To this end, we consider the neutrality of a circular elastic inclusion bonded to its surrounding matrix through $N-2$ *eccentric* circular coatings ($N \geq 3$) when the matrix is subjected to remote nonuniform antiplane shear stresses characterized by arbitrary polynomials of order $M \leq N-2$. In our design, the first $N-M-1$ generalized polarization tensors associated with the N-phase structure vanish. Our results demonstrate conclusively that for arbitrary applied nonuniform fields, the stress disturbance in the matrix becomes negligible as N becomes sufficiently large, indicating that the inclusion can be made "near-neutral" for a given N and completely neutral *as N approaches infinity.*

1. Introduction

The idea of a "neutral hole" was initiated by Mansfield [1953] who found that certain reinforced holes in a uniformly stressed plate do not alter the original stress field in the uncut body. In other words, the hole shape and corresponding reinforcing layer could be designed to make the hole "invisible" to the surrounding stress field. This idea was later extended to the concept of a "neutral inclusion" in which the insertion of certain shapes of inclusion into an elastic body causes no disturbance in the body's original stress field. In this case, "neutrality" is achieved by adding one or more specifically designed coatings between the inclusion and the surrounding body (see [Milton and Serkov 2001; Chen et al. 2002; Schiavone 2003; Mahboob and Schiavone 2005; Vasudevan and Schiavone 2006; Bertoldi et al. 2007; Benveniste and Miloh 2007; Jarczyk and Mityushev 2012; Wang and

Communicated by Francesco dell'Isola.
MSC2010: primary 74B05; secondary 74M25.
Keywords: neutral inclusion, remote nonuniform loading, multiple coating, Apollonius circles, generalized polarization tensor.

Schiavone 2012a; 2012b] for a comprehensive account of fundamental investigations in this area). This concept of "neutrality" finds significant application in the design of advanced composite materials and structures (for example, in the design of implants in biomechanics) but is also topical in that it is often taken to be equivalent to the modern ideas of "cloaking", "invisibility" or "stealth" [Milton et al. 2006; Liu 2010; Ammari et al. 2013a; 2013b] in that the inclusion becomes "invisible" to the original stress distribution.

Milton and Serkov [2001] showed that, for an isotropic medium, neutrality to multiple applied uniform fields can be realized only by the *concentrically coated circle* construction. This fact has also been observed by Ru [1998]. In the present paper we intend to show that a circular elastic inclusion with $N-2$ eccentric coatings can be made "almost neutral" to multiple applied nonuniform fields. In fact, we show that when the matrix surrounding the inclusion is subjected to remote nonuniform stresses characterized by arbitrary polynomials of order $M \le N-2$ in the complex variable z, the generalized polarization tensors (GPTs) [Ammari et al. 2013a] of up to order $N-M-1$ vanish on the introduction of the $N-2$ eccentric coatings. For a sufficiently large value of N and a relatively low value of M, the disturbance in the matrix is minimal since only GPTs of orders higher than $N-M-1$ exist. As N approaches infinity, there will be no stress disturbance in the matrix as a result of the cancellation of all GPTs.

2. Design of neutral circular inclusions with multiple eccentric coatings

Let (x_1, x_2, x_3) describe a Cartesian coordinate system in $\mathbb{R}^3$. In the theory of antiplane shear deformations, the out-of-plane displacement $w(x_1, x_2)$, the stress function $\phi(x_1, x_2)$, and the stress components $\sigma_{32}(x_1, x_2)$ and $\sigma_{31}(x_1, x_2)$, can be expressed more conveniently in terms of an analytic function $f(z)$ of the complex variable $z = x_1 + ix_2 = r\exp(i\theta)$ where $r = \sqrt{x_1^2 + x_2^2}$ and $\tan\theta = x_2/x_1$, as

$$\mu^{-1}\phi + iw = f(z), \quad \sigma_{32} + i\sigma_{31} = \mu f'(z), \tag{1}$$

where μ is the shear modulus of the material. The two stress components can be expressed in terms of the stress function ϕ as [Ting 1996]

$$\sigma_{32} = \phi_{,1}, \quad \sigma_{31} = -\phi_{,2}, \tag{2}$$

where the notation $(\cdot)_{,s}$ denotes differentiation with respect to x_s, $s = 1, 2$.

We consider a circular elastic inclusion bonded to the surrounding matrix through $N-2$ eccentric circular coatings (Figure 1). Let S_1 denote the unbounded matrix, $S_2, \ldots, S_{N-1}$ the $N-2$ intermediate coatings, and S_N the inner circular inclusion. We assume perfect bonding across the $N-1$ eccentric circles $L_1, \ldots, L_{N-1}$. Clearly, the interface L_k is formed by the outer S_k and the inner S_{k+1}. The center

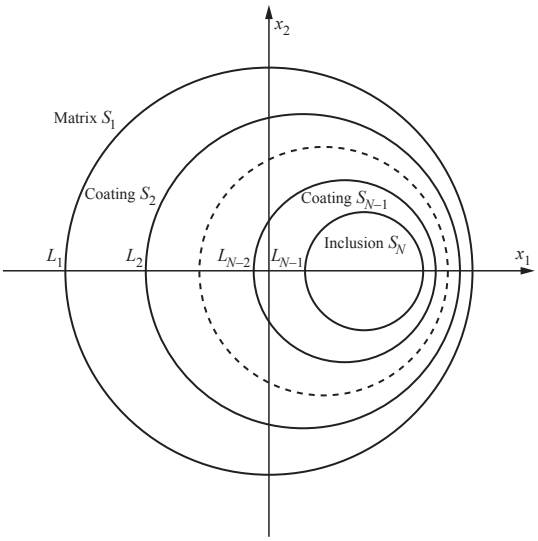


Figure 1. A circular elastic inclusion with $(N-2)$ eccentric circular coatings.

of the unit circle L_1 is at the origin, so that $L_1, \ldots, L_{N-1}$ are Apollonius circles in the sense that, if we introduce the conformal map

$$z = \omega(\xi) = \frac{\xi - a}{a\xi - 1}, \quad \xi = \omega^{-1}(z) = \frac{z - a}{az - 1} \quad (a > 1), \tag{3}$$

then the eccentric circles $L_1, \ldots, L_{N-1}$ in the z-plane are mapped onto the $N-1$ concentric circles $|\xi| = R_1, \ldots, |\xi| = R_{N-1}$ in the ξ-plane, respectively, where $R_{N-1} > R_{N-2} > \cdots > R_2 > R_1 = 1$, as shown in Figure 2. It follows from (3) that (i) the Apollonius circles L_k can be described by

$$\frac{|z - a|}{|az - 1|} = R_k,$$

which implies that the centers of these eccentric circles are all on the real axis; (ii) $z = a$ in S_1 is mapped to $\xi = 0$, and $z = 1/a$ in S_N is mapped to $\xi = \infty$. In addition, we first assume that the matrix is subjected to remote uniform antiplane stresses σ_{32}^{∞} and σ_{31}^{∞} (it will be seen in the following analysis that the remote applied stresses can be nonuniform). Throughout the remainder of this paper, the subscript j or the superscript (j) will denote the corresponding quantities associated with S_j. For convenience and without loss of generality, we adopt the notation $f(z) = f(\omega(\xi)) = f(\xi)$.

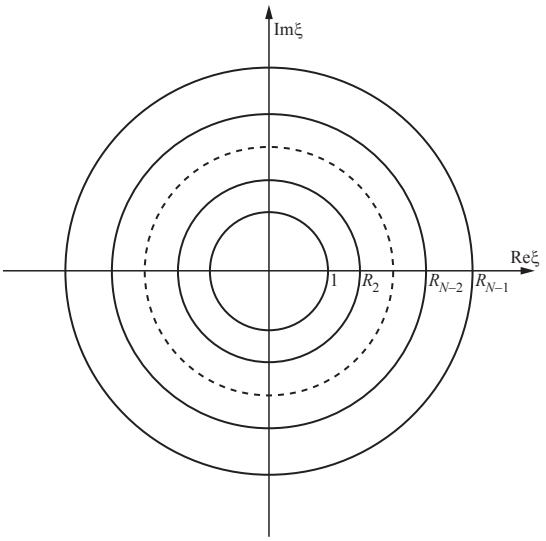


Figure 2. The mapped ξ-plane.

The analytic function $f_j(\xi)$ defined in phase j can be expanded into the convergent Laurent series

$$f_j(\xi) = \sum_{n=1}^{\infty} [A_n^{(j)} \xi^{-n} + B_n^{(j)} \xi^n], \tag{4}$$

where $A_n^{(j)}$ and $B_n^{(j)}$ are complex constants to be determined (note that we do not include the constant term ($n = 0$) in the Laurent expansion since this term corresponds to a rigid body translation and does not affect the corresponding stress field). It should be pointed out that this expansion of $f_j(\xi)$ for $j = 1$ is convergent only for $1/a < |\xi| < 1$, and the convergent expression for $f_1(\xi)$ in $|\xi| < 1$ is

$$f_1(\xi) = \frac{C}{a\xi - 1} + \sum_{n=1}^{\infty} B_n^{(1)} \xi^n, \quad |\xi| < 1, \tag{5}$$

where the complex constant C is determined from the remote uniform stresses as

$$C = \frac{(a^{-1} - a)(\sigma_{32}^{\infty} + i\sigma_{31}^{\infty})}{\mu_1}. \tag{6}$$

Remark 1. Our reasoning in obtaining (5) is as follows. In the physical z-plane,

$$f_1(z) = \frac{\sigma_{32}^{\infty} + i\sigma_{31}^{\infty}}{\mu_1} z + f_0(z), \quad |z| > 1,$$

where $f_0(z)$ is analytic everywhere in the matrix, including the point at infinity. Thus, in the mapped ξ-plane,

$$f_1(\xi) = \frac{C}{a\xi - 1} + f_0(\xi), \quad |\xi| < 1,$$

where $f_0(\xi)$ is analytic in $|\xi| < 1$ and can be expanded in a Taylor series. Consequently, (5) can be obtained with the constant term disregarded.

By enforcing the continuity condition of displacement and traction across the (perfect) interface $|\xi| = R_j$ (i.e., $\phi_j = \phi_{j+1}$, $w_j = w_{j+1}$ on $|\xi| = R_j$), we arrive at the recurrence relation

$$\begin{bmatrix} A_n^{(j+1)} \\ \bar{B}_n^{(j+1)} \end{bmatrix} = \boldsymbol{P}_n^{(j)} \begin{bmatrix} A_n^{(j)} \\ \bar{B}_n^{(j)} \end{bmatrix}, \quad n = 1, 2, \ldots, \tag{7}$$

where the transfer matrix $\boldsymbol{P}_n^{(j)}$ is given by

$$\boldsymbol{P}_n^{(j)} = \frac{1}{1 - \lambda_j} \begin{bmatrix} 1 & R_j^{2n}\lambda_j \\ R_j^{-2n}\lambda_j & 1 \end{bmatrix}, \tag{8}$$

with λ_j being the mismatch parameter defined by

$$\lambda_j = \frac{\mu_j - \mu_{j+1}}{\mu_j + \mu_{j+1}} \quad (|\lambda_j| < 1). \tag{9}$$

It follows from (7) that

$$\begin{bmatrix} A_n^{(N)} \\ \bar{B}_n^{(N)} \end{bmatrix} = \boldsymbol{S}_n \begin{bmatrix} A_n^{(1)} \\ \bar{B}_n^{(1)} \end{bmatrix}, \quad n = 1, 2, \ldots, \tag{10}$$

where

$$\boldsymbol{S}_n = \begin{bmatrix} S_n^{11} & S_n^{12} \\ S_n^{21} & S_n^{22} \end{bmatrix} = \boldsymbol{P}_n^{(N-1)} \boldsymbol{P}_n^{(N-2)} \cdots \boldsymbol{P}_n^{(2)} \boldsymbol{P}_n^{(1)}. \tag{11}$$

In order to ensure that $f_N(\xi)$ is analytic in the region $R_{N-1} < |\xi| < \infty$, including the point at infinity, we must have $B_n^{(N)} = 0$ $(n = 1, 2, \ldots)$. In addition, it can be easily deduced from (4) and (5) that

$$A_n^{(1)} = Ca^{-n}, \quad n = 1, 2, \ldots. \tag{12}$$

By imposing the above additional conditions on (10), we arrive at

$$B_n^{(1)} = -\frac{\bar{C}}{a^n} \frac{S_n^{21}}{S_n^{22}}, \quad n = 1, 2, \ldots. \tag{13}$$

Thus $f_1(\xi)$ defined in the matrix can be uniquely determined as

$$f_1(\xi) = \frac{C}{a\xi - 1} - \bar{C} \sum_{n=1}^{\infty} \left(a^{-n} \xi^n \frac{S_n^{21}}{S_n^{22}} \right), \quad (|\xi| < 1). \tag{14}$$

In order to arrive at a GPT-vanishing structure of order $N-2$, the following $N-2$ conditions should be satisfied:

$$g^{(k)}(a^{-1}) = 0, \quad k = 1, 2, \ldots, N-2, \tag{15}$$

where the superscript (k) denotes the k-th order derivative, and

$$g(\xi) = f_1(\xi) - \frac{C}{a\xi - 1}. \tag{16}$$

Remark 2. The conditions given by (15) result in the following asymptotic behavior of $f_1(z)$ at infinity:

$$f_1(z) \cong \frac{\sigma_{32}^{\infty} + i\sigma_{31}^{\infty}}{\mu_1} z + O(1/z^{N-1}) \quad \text{as } |z| \to \infty,$$

which indicates that the GPTs up to order $N-2$ all vanish.

In view of (14), (15) can be written explicitly as

$$\begin{aligned} \sum_{n=1}^{\infty} \left[n a^{-2n} \frac{S_n^{21}}{S_n^{22}} \right] &= 0, \\ \sum_{n=1}^{\infty} \left[n(n-1) a^{-2n} \frac{S_n^{21}}{S_n^{22}} \right] &= 0, \\ &\vdots \\ \sum_{n=1}^{\infty} \left[n(n-1)\cdots(n+3-N) a^{-2n} \frac{S_n^{21}}{S_n^{22}} \right] &= 0, \end{aligned} \tag{17}$$

which are independent of the remote uniform stresses characterized by the complex constant C. In addition, we have the following more general result.

Theorem 1. *Equation* (17) *is also the condition leading to a GPT-vanishing structure of order* $N-M-1$ *with* $M \le N-2$ *when the matrix is subjected to remote nonuniform stresses characterized by*

$$f_1(z) \cong \sum_{n=1}^{M} D_n z^n + O(1/z^{N-M}), \quad M \le N-2, \quad \text{as } |z| \to \infty, \tag{18}$$

where D_n $(n = 1, 2, \ldots, M)$ *are complex constants.*

Remark 3. In writing (18), it has been implied that the GPTs up to the order $N-M-1$ vanish. As N approaches infinity, keeping M finite, the inclusion will become ideally neutral to arbitrary remote nonuniform stresses.

Proof of Theorem 1. In the region $|\xi|<1$, $f_1(\xi)$ can be written in the convergent form

$$f_1(\xi)=\sum_{n=1}^{M}\frac{C_n}{(a\xi-1)^n}+\sum_{n=1}^{\infty}B_n^{(1)}\xi^n,\quad |\xi|<1, \tag{19}$$

where the complex constants C_n can be determined from the nonuniform remote loading characterized by (18).

Through satisfaction of the continuity conditions of traction and displacement across all the existing interfaces, $f_1(\xi)$ can be finally determined as

$$\begin{aligned} f_1(\xi)=&\sum_{n=1}^{M}\frac{C_n}{(a\xi-1)^n}\\ &-\sum_{n=1}^{\infty}\left[\bar{C}_1+\sum_{m=2}^{M}\bar{C}_m\frac{(n-1)(n-2)\cdots(n-m+1)}{(m-1)!}\right]\frac{S_n^{21}}{S_n^{22}}\frac{\xi^n}{a^n},\quad |\xi|<1. \end{aligned}\tag{20}$$

If we define the function

$$h(\xi)=f_1(\xi)-\sum_{n=1}^{M}\frac{C_n}{(a\xi-1)^n}, \tag{21}$$

the $N-2$ conditions in (17) will lead to $h^{(k)}(a^{-1})=0$, $(k=1,2,\ldots,N-M-1)$. This fact implies that the GPTs up to order $N-M-1$ vanish. This completes the proof. □

If the $N-1$ geometric parameters a and $R_2, R_3, \ldots, R_{N-1}$ are given, (17) can be considered as a set of $N-2$ nonlinear equations for the $N-1$ mismatch parameters $\lambda_1, \lambda_2, \ldots, \lambda_{N-1}$, which can be solved through iteration. In addition, it can be shown that if $(\lambda_1, \lambda_2, \ldots, \lambda_{N-1})$ is a solution to (17), then $(-\lambda_1, -\lambda_2, \ldots, -\lambda_{N-1})$ is also a solution.

For example, when $N=4$, (17) becomes

$$\begin{aligned} \sum_{n=1}^{\infty}\left[na^{-2n}\frac{\lambda_1+R_2^{-2n}\lambda_2+R_3^{-2n}\lambda_3+R_2^{2n}R_3^{-2n}\lambda_1\lambda_2\lambda_3}{1+R_2^{-2n}\lambda_1\lambda_2+R_3^{-2n}\lambda_1\lambda_3+R_2^{2n}R_3^{-2n}\lambda_2\lambda_3}\right]=0,\\ \sum_{n=1}^{\infty}\left[n(n-1)a^{-2n}\frac{\lambda_1+R_2^{-2n}\lambda_2+R_3^{-2n}\lambda_3+R_2^{2n}R_3^{-2n}\lambda_1\lambda_2\lambda_3}{1+R_2^{-2n}\lambda_1\lambda_2+R_3^{-2n}\lambda_1\lambda_3+R_2^{2n}R_3^{-2n}\lambda_2\lambda_3}\right]=0. \end{aligned}\tag{22}$$

3. Results and discussions

In this section, we will present numerical results for the cases $N = 3$, $N = 4$ and $N \geq 5$. It is of interest to note that very simple approximate closed-form solutions for $N = 3$ and $N = 4$ can be obtained which, in turn, can be used to quickly (albeit roughly) determine the values of the mismatch parameters. For simplicity it is assumed that the matrix is subjected to only remote uniform stresses ($M = 1$).

3.1. ***GPT-vanishing structures of order* 1 ($N = 3$).** In the case of $N = 3$, the following single nonlinear equation should be solved:

$$\sum_{n=1}^{\infty}\left[na^{-2n}\frac{\lambda_1 + R_2^{-2n}\lambda_2}{1 + R_2^{-2n}\lambda_1\lambda_2}\right] = 0. \tag{23}$$

In view of the fact that $R_2 > 1$ and $|\lambda_1|, |\lambda_2| < 1$, the denominator on the left-hand side of (23) can be taken as approximately equal to one (i.e., $1 + R_2^{-2n}\lambda_1\lambda_2 \approx 1$). Consequently, the following approximate closed-form solution is obtained:

$$\frac{\lambda_1}{\lambda_2} \approx -\frac{R_2^2(a^2-1)^2}{(a^2R_2^2-1)^2} = -R_0^2, \tag{24}$$

where R_0 is the radius of the inner circular inclusion. We recall that this is just the condition for the existence of a neutral three-phase inclusion with two concentric circular interfaces with radii R_0 and 1 ($R_0 < 1$) [Ammari et al. 2013a; Ru 1999]. This implies that if a concentrically single-coated inclusion is neutral to a remote uniform stress field, the GPT of order 1 of the shifted structure *nearly* vanishes. We illustrate in Figure 3 the values of (λ_1, λ_2) found for four different values of a, namely $a = 1.2, 1.5, 2, 10$, with $R_2 = 1.5$. The solid lines are obtained by iteratively solving (23), whilst the dashed lines are obtained by using the approximate solution (24). It is observed that the approximate results are very close to the exact ones. As $a \to \infty$ (i.e., the eccentricity becomes minimal), (24) simply recovers the exact solution.

3.2. ***GPT-vanishing structures of order* 2 ($N = 4$).** In the case of $N = 4$, the equation (22) should be solved iteratively. In addition, the following approximate closed-form solution can be derived:

If $R_3 \neq R_2^2$, we have the approximate solution

$$\begin{aligned}
\lambda_2 &\approx \frac{-c_1 + \sqrt{c_1^2 - 4c_0c_2}}{2c_2},\\
\lambda_1 &\approx -\frac{(a^2-1)^2(a^2R_3^2-R_2^2)^2\big[R_2^2(a^2R_3^2-1)^2\lambda_2 + R_3^2(a^2R_2^2-1)^2\lambda_3\big]}{(a^2R_2^2-1)^2(a^2R_3^2-1)^2\big[(a^2R_3^2-R_2^2)^2 + R_2^2R_3^2(a^2-1)^2\lambda_2\lambda_3\big]},
\end{aligned} \tag{25}$$

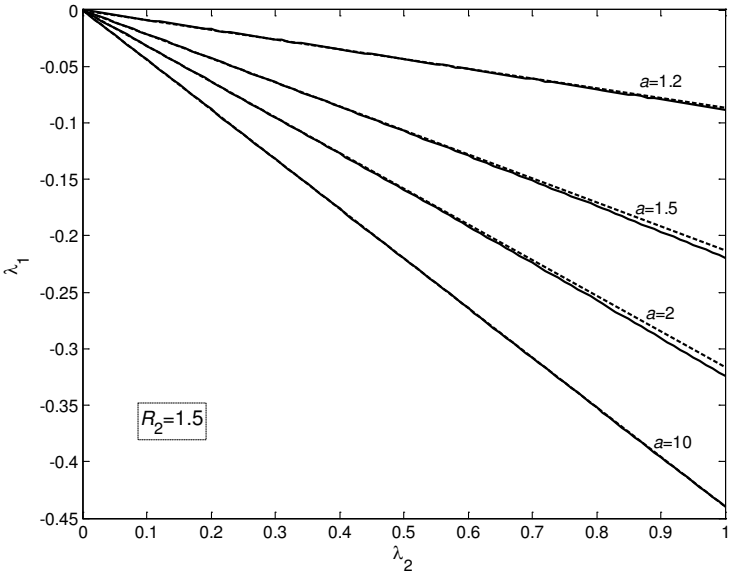


Figure 3. Obtained values of (λ_1, λ_2) for $a = 1.2, 1.5, 2, 10$, with $R_2 = 1.5$.

where

$$\begin{aligned}
c_2 &= R_2^4 R_3^2 (a^2-1)^3 (a^2R_3^2-1)^3 (R_2^4 - R_3^2)\lambda_3,\\
c_1 &= R_2^2 (R_2^2-1)\big[R_3^6 (a^2-1)^3 (a^2R_2^2-1)^3 \lambda_3^2 + (a^2R_3^2-1)^3 (a^2R_3^2 - R_2^2)^3\big],\\
c_0 &= R_3^2 (R_3^2-1)(a^2R_3^2 - R_2^2)^3 (a^2R_2^2-1)^3 \lambda_3.
\end{aligned} \tag{26}$$

On the other hand, if $R_3 = R_2^2$, we have the approximate solution

$$\begin{aligned}
\lambda_2 &\approx -\frac{\lambda_3 R_3 (R_3+1)(a^2R_3-1)^3}{R_3^3(a^2-1)^3\lambda_3^2 + (a^2R_3^2-1)^3},\\
\lambda_1 &\approx \frac{\lambda_3 R_3^3 (a^2-1)^3\big[R_3(a^2-1)^3\lambda_3^2 + (a^2R_3^2-1)^3\big]}{(a^2R_3^2-1)^3\big[R_3^3(a^2-1)^3\lambda_3^2 + (a^2R_3^2-1)^3\big]}\\
&\qquad \times \frac{R_3^2(a^2-1)^2\lambda_3^2 - (a^2R_3^2-1)^2}{(a^2-1)^2\lambda_3^2 - (a^2R_3^2-1)^2}.
\end{aligned} \tag{27}$$

We illustrate in Figure 4 the variations of λ_1 and λ_2 as functions of λ_3 for three values of a, namely $a = 1.01, 1.5, 3$, with $R_2 = 1.5$, $R_3 = 2$. The solid lines are the exact results obtained by iteratively solving (22), whereas the dashed lines are the approximate results found from (25). It is observed from Figure 4 that the approximate results are quite satisfactory, especially when $\lambda_3 < 0.7$.

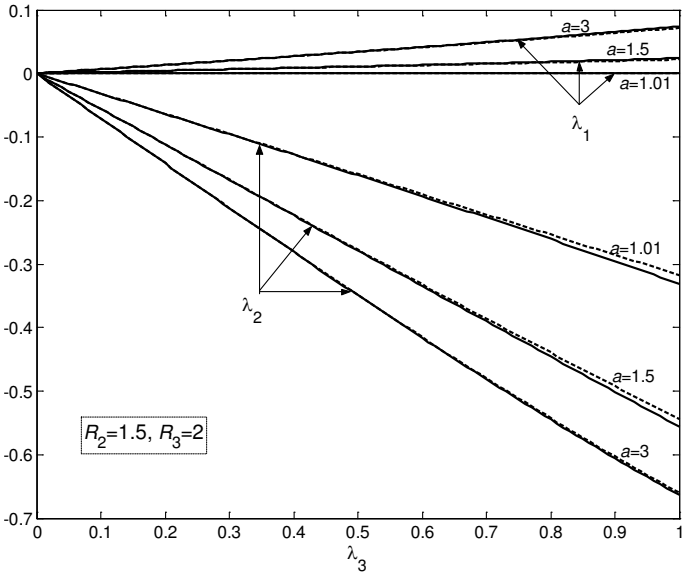


Figure 4. Variations of λ_1 and λ_2 as functions of λ_3 for $a = 1.01, 1.5, 3$, with $R_2 = 1.5$, $R_3 = 2$.

3.3. ***GPT-vanishing structures of order* 3 *or higher* ($N \geq 5$).** When $N \geq 5$, the solutions can be obtained only by solving (17) iteratively. Listed in Table 1 are typical results. In performing the calculations, we set the $N-1$ geometric parameters to $a = 1.2$ and $R_{k+1} = 1 + k/(N-2)$, $(k = 1, 2, \ldots, N-2)$. It is observed that λ_k and λ_{k+1} always have opposite signs.

3.4. ***Stress disturbance in the matrix.*** The concept of neutral holes and inclusions was originally proposed to completely eliminate stress concentrations in the matrix [Mansfield 1953; Milton and Serkov 2001; Ru 1998]. In our design, however, the stress disturbance in the matrix cannot be completely avoided due to the existence of GPTs of orders higher than $N-2$ when the remote loading is uniform. However,

	$N=5$	$N=6$	$N=7$	$N=8$	$N=9$	$N=10$
λ_1	5.464×10^{-4}	4.203×10^{-5}	4.42×10^{-6}	4.289×10^{-7}	3.613×10^{-8}	2.221×10^{-9}
$-\lambda_2$	0.1075	0.0147	0.0024	3.197×10^{-4}	3.573×10^{-5}	2.800×10^{-6}
λ_3	0.6415	0.2041	0.0592	0.0131	0.0022	2.471×10^{-4}
$-\lambda_4$	0.8	0.5751	0.3239	0.1225	0.0323	0.0053
λ_5		0.5	0.5868	0.3964	0.1720	0.0430
$-\lambda_6$			0.4	0.5437	0.3979	0.1594
λ_7				0.3	0.4445	0.2939
$-\lambda_8$					0.2	0.2728
λ_9						0.1

Table 1. Obtained mismatch parameters for $N = 5, 6, 7, 8, 9, 10$.

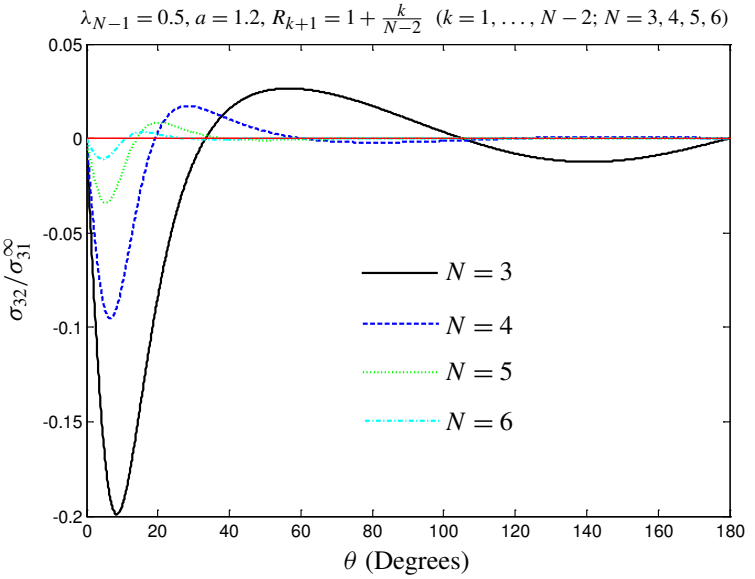


Figure 5. The stress disturbance $\sigma_{32}/\sigma_{31}^{\infty}$ along the coating/matrix interface $|z| = 1$ on the matrix side for $N = 3, 4, 5, 6$ when the matrix is subjected to $\sigma_{31} \neq 0$, $\sigma_{32}^{\infty} = 0$.

the stress disturbance is expected to be insignificant as N becomes sufficiently large. In fact, we illustrate in Figures 5 and 6 the stress disturbance along the coating/matrix interface $|z| = 1$ on the matrix side for $N = 3, 4, 5, 6$ when the matrix is subjected to the loading given by $\sigma_{31}^{\infty} \neq 0$, $\sigma_{32}^{\infty} = 0$. It is observed from the two figures that, as N increases, the most significant stress disturbance occurs in a more localized region of θ: $\theta < 60°, 40°$ and $20°$ for $N = 4, 5$ and 6, respectively. Clearly, when $N = 6$, the stress disturbance along the whole interface $|z| = 1$ is minimal.

4. Conclusions

By adopting the GPT cancellation method proposed in [Ammari et al. 2013a], we design "near-neutral" circular elastic inclusions with multiple eccentric circular coatings. When the matrix is subjected to remote nonuniform stress characterized by (18), the GPTs up to the $(N-M-1)$-th order are canceled by appropriately adding $N-2$ eccentric coatings between the inclusion and the matrix. Condition (17) is, in fact, independent of the remote applied nonuniform loading given by (18). Consequently, our design of an N-phase circular inclusion is "almost neutral" to the remote nonuniform stresses characterized by any polynomials in z of order M no greater than $N-2$. In order to make the analysis tractable, we

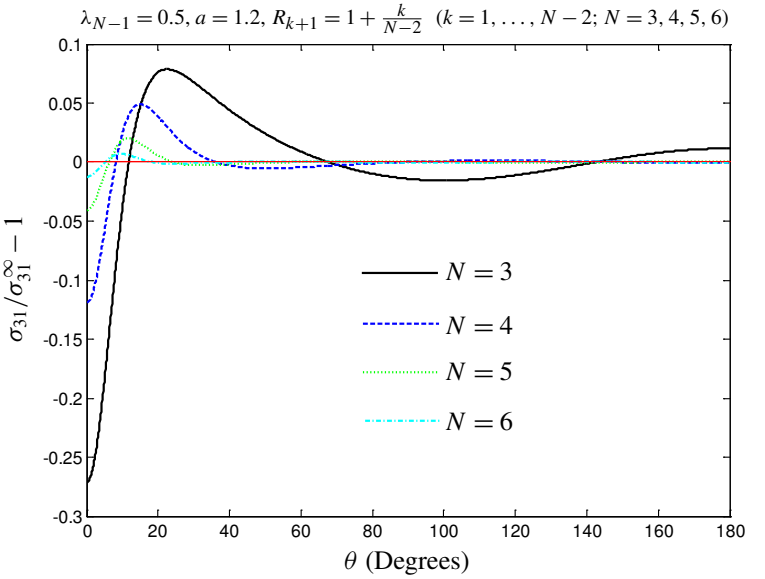


Figure 6. The stress disturbance $\sigma_{31}/\sigma_{31}^{\infty} - 1$ along the coating/matrix interface $|z| = 1$ on the matrix side for $N = 3, 4, 5, 6$ when the matrix is subjected to $\sigma_{31}^{\infty} \neq 0$, $\sigma_{32}^{\infty} = 0$.

assume that all the eccentric circular interfaces L_k $(k = 1, 2, \ldots, N-1)$ are Apollonius circles. Approximate closed-form solutions (24) for $N = 3$ and (25)–(27) for $N = 4$ are obtained. One consequence and potential application of the results here arises from the finding that multiple closely spaced and eccentrically coated inclusions can be inserted into a nonuniformly stressed finite matrix with minimal stress disturbance in the matrix.

Acknowledgements

The authors are indebted to two very meticulous reviewers, whose comments greatly improved the paper. This work is supported by the National Natural Science Foundation of China (grant number 11272121) and through a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

References

[Ammari et al. 2013a] H. Ammari, H. Kang, H. Lee, and M. Lim, "Enhancement of near cloaking using generalized polarization tensors vanishing structures, I: The conductivity problem", *Comm. Math. Phys.* **317**:1 (2013), 253–266. arXiv 1104.3936v1

[Ammari et al. 2013b] H. Ammari, H. Kang, H. Lee, and M. Lim, "Enhancement of near-cloaking, II: The Helmholtz equation", *Comm. Math. Phys.* **317**:2 (2013), 485–502. arXiv 1110.1922v1

[Benveniste and Miloh 2007] Y. Benveniste and T. Miloh, "Soft neutral elastic inhomogeneities with membrane-type interface conditions", *J. Elasticity* **88**:2 (2007), 87–111.

[Bertoldi et al. 2007] K. Bertoldi, D. Bigoni, and W. J. Drugan, "Structural interfaces in linear elasticity, II: Effective properties and neutrality", *J. Mech. Phys. Solids* **55**:1 (2007), 35–63.

[Chen et al. 2002] T. Chen, Y. Benveniste, and P. C. Chuang, "Exact solutions in torsion of composite bars: thickly coated neutral inhomogeneities and composite cylinder assemblages", *Proc. R. Soc. Lond. A* **458**:2023 (2002), 1719–1759.

[Jarczyk and Mityushev 2012] P. Jarczyk and V. Mityushev, "Neutral coated inclusions of finite conductivity", *Proc. R. Soc. Lond. A* **468**:2140 (2012), 954–970.

[Liu 2010] L. P. Liu, "Neutral shells and their applications in the design of electromagnetic shields", *Proc. R. Soc. Lond. A* **466**:2124 (2010), 3659–3677.

[Mahboob and Schiavone 2005] M. Mahboob and P. Schiavone, "Designing a neutral elliptic inhomogeneity in the case of a general non-uniform loading", *Appl. Math. Lett.* **18**:11 (2005), 1312–1318.

[Mansfield 1953] E. H. Mansfield, "Neutral holes in plane sheet: reinforced holes which are elastically equivalent to the uncut sheet", *Quart. J. Mech. Appl. Math.* **6** (1953), 370–378.

[Milton and Serkov 2001] G. W. Milton and S. K. Serkov, "Neutral coated inclusions in conductivity and anti-plane elasticity", *Proc. R. Soc. Lond. A* **457**:2012 (2001), 1973–1997.

[Milton et al. 2006] G. W. Milton, M. Briane, and J. R. Willis, "On cloaking for elasticity and physical equations with a transformation invariant form", *New J. Phys.* **8**:10 (2006), Article ID #248.

[Ru 1998] C.-Q. Ru, "Interface design of neutral elastic inclusions", *Int. J. Solids Struct.* **35**:7–8 (1998), 559–572.

[Ru 1999] C.-Q. Ru, "A new method for an inhomogeneity with stepwise graded interphase under thermomechanical loadings", *J. Elasticity* **56**:2 (1999), 107–127.

[Schiavone 2003] P. Schiavone, "Neutrality of the elliptic inhomogeneity in the case of non-uniform loading", *Int. J. Eng. Sci.* **41**:18 (2003), 2081–2090.

[Ting 1996] T. C. T. Ting, *Anisotropic elasticity: theory and applications*, Oxford Engineering Science Series **45**, Oxford University Press, New York, 1996.

[Vasudevan and Schiavone 2006] M. Vasudevan and P. Schiavone, "New results concerning the identification of neutral inhomogeneities in plane elasticity", *Arch. Mech. Stos.* **58**:1 (2006), 45–58.

[Wang and Schiavone 2012a] X. Wang and P. Schiavone, "Neutral coated circular inclusions in finite plane elasticity of harmonic materials", *Eur. J. Mech. A Solids* **33** (2012), 75–81.

[Wang and Schiavone 2012b] X. Wang and P. Schiavone, "Neutrality in the case of N-phase elliptical inclusions with internal uniform hydrostatic stresses", *Int. J. Solids Struct.* **49**:5 (2012), 800–807.

Received 1 Nov 2013. Revised 13 Feb 2014. Accepted 21 Apr 2014.

XU WANG: xuwang@ecust.edu.cn
School of Mechanical and Power Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai, 200237, China

PETER SCHIAVONE: p.schiavone@ualberta.ca
Department of Mechanical Engineering, University of Alberta, 4-9 Mechanical Engineering Building, Edmonton, Alberta, Canada T6G 2G8


MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 3, No. 2, 2015
dx.doi.org/10.2140/memocs.2015.3.177

RELATIVE CAUCHY EVOLUTION FOR THE VECTOR POTENTIAL ON GLOBALLY HYPERBOLIC SPACETIMES

MARCO BENINI

The dynamics of the electromagnetic vector potential is analyzed in full detail in view of the principle of general local covariance of Brunetti, Fredenhagen and Verch. Exploiting this result, the relative Cauchy evolution for the vector potential is introduced and its relation with the energy-momentum tensor is established, extending the well known results for Klein–Gordon and Dirac fields.

1. Introduction

The *principle of general local covariance* of Brunetti, Fredenhagen and Verch [Brunetti et al. 2003] provides a very satisfactory framework to deal with quantum field theory on curved spacetimes. The success of the axiomatic approach of general local covariance relies on its capability to establish how a quantum field theory is expected to behave on different spacetimes and in particular what kind of relation one should expect between the observables defined on two spacetimes when one of them is isometrically embedded into the other. An effective way being available to relate via embeddings quantum field theories on different curved spacetimes, the way is paved to tackle the question of the sensitivity of the model under small fluctuations of the background geometry. This is what the relative Cauchy evolution is meant for, namely to provide information about the modification induced on any observable by a small change in the metric of the background spacetime where the dynamics of the quantum field takes place.

The core idea of the relative Cauchy evolution can be traced back to the fact that a normally hyperbolic equation which rules the dynamics of a field on a globally hyperbolic spacetime admits a well-posed initial value problem; see for example [Bär et al. 2007, Section 3.2]. This means that all information about the field is determined by suitable initial data specified on a Cauchy surface, thus enabling us to sketch the behavior of the relative Cauchy evolution in terms of initial data only. One can consider a perturbation of the spacetime metric supported in a compact

Communicated by Mauro Carfora.
MSC2010: 81T20, 81T05, 81T13.
Keywords: quantum field theory on curved spacetimes, Maxwell equation, general local covariance, relative Cauchy evolution.

region, say K, and a set of initial data on a Cauchy surface lying in the past of K. Propagating such initial data to a Cauchy surface in the future of K both in the presence of the metric perturbation and without it, one is able to compare the two different outcomes of the dynamical evolution (one when the background geometry is perturbed and the other when this is not the case). This procedure allows one to quantify the effect induced on the dynamics of the field by suitable modifications of the background geometry.

The aim of the present paper is to study the relative Cauchy evolution for the gauge field theory of the electromagnetic vector potential. Given a globally hyperbolic spacetime, the dynamics of the field $A \in \Omega^1(M)$ is ruled by the *non*hyperbolic equation $\delta\, \mathrm{d}\, A = 0$, d and δ being respectively the differential and the codifferential defined for forms over M. The gauge symmetry of the vector potential is specified by the equivalence relation

$$A \sim A' \quad \Longleftrightarrow \quad \exists \chi \in \mathrm{C}^\infty(M) :\ A' = A + \mathrm{d}\chi\,.$$

Although the sketch of the relative Cauchy evolution presented above cannot be directly applied to the case of interest, the dynamics being nonhyperbolic, we can, following [Dimock 1992; Fewster and Pfenning 2003; Pfenning 2009; Dappiaggi 2011; Sanders et al. 2014], exploit the gauge symmetry to recover hyperbolicity in the gauge-fixed dynamics. This eventually leads us to the fulfillment of (almost all) requirements of the generally covariant locality principle. In particular, we are allowed to introduce the relative Cauchy evolution for the vector potential A and analyze its properties, the main result consisting of the extension of a fact which is known to hold for Klein–Gordon and Dirac fields [Brunetti et al. 2003; Sanders 2010], namely the relation between the relative Cauchy evolution of a field and its quantized energy-momentum tensor.

Such a relation between the relative Cauchy evolution and the quantized energy-momentum tensor is relevant when one is dealing with the semiclassical Einstein equation (see [Wald 1994, Section 4.6] for an introduction to this topic) in the presence of a quantized electromagnetic vector potential. As a matter of fact, in this case one is supposed to equal the Einstein tensor with the expectation value of the quantized energy-momentum tensor of electromagnetism in order to account for the back reaction effect on the spacetime metric induced by the presence of a quantized electromagnetic field, whose dynamics is in turn affected by the spacetime geometry. Fortunately, one can access the behavior of the quantized energy-momentum tensor in relation to suitable changes of the background metric by means of the relative Cauchy evolution. Therefore an important step towards a consistent approach to the solution of the semiclassical Einstein equation in the presence of a quantized electromagnetic field consists of a detailed analysis of the relative Cauchy evolution for the electromagnetic field as well as of its relation

with the quantized energy-momentum tensor. This fact motivates our interest in analyzing the relative Cauchy evolution of the vector potential.

The paper is organized in the following way: Section 2 is intended to provide the background information and notation which are needed in the rest of the paper. Some aspects of Lorentzian geometry are briefly discussed in Section 2A, focusing on global hyperbolicity in particular. A short collection of the most relevant properties of Green-hyperbolic linear differential operators follows in Section 2B. We recall the notion of a locally covariant quantum field theory and of its relative Cauchy evolution in Section 2C. Section 2 is completed with the description of a quantization procedure based on the Borchers–Uhlmann construction and recalling the notion of an algebraic state. In Section 3 we analyze the dynamics of the vector potential providing a convenient characterization of the space of solutions for the equation $\delta \mathrm{d} A = 0$. This leads in Section 4 to the assignment of a suitable space of observables for the vector potential and its quantization. This section ends recalling the definition of a Hadamard state for the vector potential, together with references to the literature where positive results about the existence of such a state can be found. The core of the paper is Section 5, where the relative Cauchy evolution for the vector potential is computed and its relation with the quantized energy-momentum tensor is established, thus extending a result which was already known to hold for Klein–Gordon [Brunetti et al. 2003] and Dirac fields [Sanders 2010].

2. Preliminaries

In this section we collect the background material and, at the same time, we introduce some notation needed later. First, we will briefly recall few notions about Lorentzian geometry focusing the attention on *globally hyperbolic spacetimes*, whose physical relevance is related to initial value problems for hyperbolic linear partial differential equations. As a matter of fact, globally hyperbolic spacetimes provide a sufficiently general setting for proving existence and uniqueness theorems for solutions of partial differential equations of hyperbolic type once proper initial data are given; see [Bär et al. 2007, Chapter 3]. This leads us to the second part of the present section, which is devoted to differential operators. We will focus the attention on the class of *Green-hyperbolic* operators, being characterized by the existence of *retarded and advanced Green functions*. We will take the chance to recollect from the literature few fundamental results, which will turn out to be useful throughout the rest of the paper. In the third part of this section, we first provide the framework for the *relative Cauchy evolution*, namely we introduce *general local covariance* following [Brunetti et al. 2003] and in particular the *time slice axiom*, and then we define it using a simple geometrical construction. Since in the end we are interested in the quantization of our model, we recall an

algebraic procedure to assign *canonical commutation relations*, we define states from an algebraic perspective and we make contact with the usual Hilbert space representation of quantum field theory via the Gelfand–Naimark–Segal theorem.

2A. ***Lorentzian geometry.*** We recall here few basic notions of Lorentzian geometry, global hyperbolicity in particular, and we take the chance to introduce some notation. For a detailed analysis of these topics, the reader should refer to the literature; see, e.g.,[Beem et al. 1996; Bär et al. 2007; Waldmann 2012].

In the following all manifolds and functions between manifolds are considered to be smooth, unless otherwise stated. Sometimes we will also restrict ourselves to *manifolds of finite type*, namely manifolds admitting a finite good cover. This will enable us to fully recover full Poincaré duality; see [Bott and Tu 1982, Chapter 1, Section 5].

A *Lorentzian manifold* $(M, g, \mathfrak{o})$ is a d-dimensional, orientable, connected, second-countable, Hausdorff manifold endowed with a Lorentzian metric g and a choice of orientation $\mathfrak{o}$. We adopt the convention $-+\cdots+$ for the signature of g. This structure already enables us to distinguish among *timelike*, *lightlike* (all together *causal*) and *spacelike* tangent vectors $0 \neq v \in T_x M$ at a point $x \in M$ according to the negative, null or positive value of $g(v, v)$. Moreover, the choice of an orientation $\mathfrak{o}$, together with the metric g, uniquely identifies a volume form vol on M, which is used to integrate functions defined M.

In order to account for the dynamical evolution of a physical system, proper notions of future and past are required. This is achieved taking a time-orientable Lorentzian manifold and fixing a time-orientation specified by a timelike vector field $\mathfrak{t}$, which is used as a reference to distinguish between future- and past-directed causal tangent vectors v according to the sign of $g(\mathfrak{t}, v)$ (future for negative values). The quadruple $(M, g, \mathfrak{o}, \mathfrak{t})$ defines a *spacetime*, where the notion of *causal future/past* of $O \subseteq M$, $J_M^{\pm}(O)$, is available. $J_M^{\pm}(O)$ is defined as the set of points in M that can be reached via a future-/past-directed causal curve in M emanating from O, namely a curve whose tangent vector field is everywhere causal and future-/past-directed. If we take into account only timelike curves in the last definition, we obtain the chronological future/past of O, $I_M^{\pm}(O)$. With the notion of causal future and past at hand, we can characterize subregions of M which are compatible with the causal structure of $(M, g, \mathfrak{o}, \mathfrak{t})$, as well as maps between spacetimes preserving causal structures. Specifically, a region $S \subseteq M$ is called *causally compatible* provided that $J_S^{\pm}(x) = J_M^{\pm}(x) \cap S$ regardless of the choice of $x \in S$.[1] Furthermore, a *causal embedding* f between the spacetimes $(M_1, g_1, \mathfrak{o}_1, \mathfrak{t}_1)$ and $(M_2, g_2, \mathfrak{o}_2, \mathfrak{t}_2)$ is defined as an embedding $f : M_1 \to M_2$ such that $f^* g_2 = g_1$, which preserves

[1] Note that in the definition of $J_S^{\pm}(x)$ only curves which never leave S are taken into account.

both the orientations and the time-orientations, whose image $f(M_1)$ is open and causally compatible as a subset of M_2.

In order to provide appropriate initial data for Cauchy problems a *Cauchy surface* is needed; see, for example, [Bär et al. 2007, Section 3.2]. This is a subset Σ of a spacetime $(M, g, \mathfrak{o}, \mathfrak{t})$ which meets exactly once each inextensible future-directed timelike curve. Cauchy surfaces provide a definition of globally hyperbolic spacetimes. As a matter of fact, a spacetime $(M, g, \mathfrak{o}, \mathfrak{t})$ is called *globally hyperbolic* when it admits a Cauchy surface. In the following we will denote globally hyperbolic spacetimes with M, the metric g, the orientation $\mathfrak{o}$, and the time-orientation $\mathfrak{t}$ being understood.

For later purposes, we introduce here some nomenclature for spacetime subregions. Following [Bär 2013; Sanders 2013], we call a subset S of a globally hyperbolic spacetime M:

- *spacelike-compact* (sc) if it is closed and there exists $K \subseteq M$ compact such that $S \subseteq J_M(K) = J_M^+(K) \cup J_M^-(K)$;
- *past-compact/future-compact* (pc/fc) if $S \cap J_M^{\mp}(K)$ is compact for each compact subset K of M.

If a region $S \subseteq M$ is both pc and fc, we call it *timelike-compact* (tc). If it is both pc/fc and sc, we say that it is *strictly past-compact/strictly future-compact* (spc/sfc).

2B. ***Green-hyperbolic differential operators.*** Following the definitions in [Bär and Ginoux 2012a; 2012b; Bär 2013], this is a class of linear differential operators admitting retarded and advanced Green functions on globally hyperbolic spacetimes. This class includes, of course, all wave operators, such as the d'Alembert operator $\Box_\nabla = g^{\mu\nu}\nabla_\mu\nabla_\nu$ defined out of any connection ∇ on a vector bundle, but from a physical perspective it has the advantage of encompassing other relevant cases, such as the Dirac and Proca equations.

Here we briefly review the definitions of retarded and advanced Green functions. At the same time we recall few fundamental results for the so-called causal propagator. For a detailed discussion, as well as proofs of the forthcoming statements, the reader may refer to the papers just cited, as well as [Bär et al. 2007]. A review, with some physically relevant examples, is available in [Benini et al. 2013].

Definition 2.1. Let V and W be vector bundles over a globally hyperbolic spacetime M and consider a linear differential operator $P : \Gamma(V) \to \Gamma(W)$ defined between the corresponding spaces of sections. We call *retarded/advanced Green operator* for P a linear map $G_\pm : \Gamma_c(W) \to \Gamma(V)$ such that the following conditions hold for each $\sigma \in \Gamma_c(W)$ and $\tau \in \Gamma_c(V)$:

$$PG_\pm\sigma = \sigma\,, \quad G_\pm P\tau = \tau\,, \quad \operatorname{supp}(G_\pm\sigma) \subseteq J_M^\pm(\operatorname{supp}(\sigma))\,. \tag{2-1}$$

If we endow V with a nondegenerate inner product on the fibers, denoted by $\langle \cdot, \cdot \rangle_V : V \times V \to M \times \mathbb{R}$, and taking into account the volume form vol of M, we can introduce an inner product $(\cdot, \cdot)_V$ on the corresponding space of sections. This is defined according to

$$(\sigma, \tau)_V = \int_M \langle \sigma, \tau \rangle_V \, \mathrm{vol}, \tag{2-2}$$

for each $\sigma, \tau \in \Gamma(V)$ with compact overlapping supports, namely such that $\mathrm{supp}(\sigma) \cap \mathrm{supp}(\tau)$ is compact.

Given a linear differential operator $P : \Gamma(V) \to \Gamma(W)$ as in Definition 2.1 and assuming that both V and W are endowed with nondegenerate inner products, we can introduce the formal adjoint $P^* : \Gamma(W) \to \Gamma(V)$ of P by setting

$$(P^*\sigma, \tau)_V = (\sigma, P\tau)_W \,, \tag{2-3}$$

for each $\sigma \in \Gamma(W)$ and $\tau \in \Gamma(V)$ with compact overlapping support. We are now ready to define linear differential operators of Green-hyperbolic type.

Definition 2.2. Let V, W be vector bundles over a globally hyperbolic spacetime M endowed with nondegenerate inner products. A linear differential operator $P : \Gamma(V) \to \Gamma(W)$ is of *Green-hyperbolic* type if it admits retarded and advanced Green operators, together with its formal adjoint $P^* : \Gamma(W) \to \Gamma(V)$.

The fact that P^* is the formal adjoint of P entails a relation between the corresponding Green functions:

$$(G^*_\pm \sigma, \tau)_W = (\sigma, G_\mp \tau)_V \,, \tag{2-4}$$

for each $\sigma \in \Gamma_{\mathrm{c}}(V)$ and $\tau \in \Gamma_{\mathrm{c}}(W)$. As a consequence, retarded and advanced Green operators for both P and P^* are unique.

Moreover, Green functions for Green-hyperbolic operators admit unique continuous extensions to larger spaces of sections; see [Bär 2013, Section 3] and [Sanders 2013, Section 5]. With a slight abuse of notation, we denote with $G_\pm$ also the extended Green operators for P:

$$G_+ : \Gamma_{\mathrm{pc}}(W) \to \Gamma(V), \quad G_- : \Gamma_{\mathrm{fc}}(W) \to \Gamma(V) \,, \tag{2-5}$$

the subscripts "pc" and "fc" referring to the supports of sections, which are past-compact in the first case and future-compact in the second; see Section 2A. Extended Green operators share the same properties of the original ones, but in a broader sense: For each $\sigma \in \Gamma_{\mathrm{pc}}(W)$ and $\tau \in \Gamma_{\mathrm{pc}}(V)$, we have

$$PG_+\sigma = \sigma \,, \quad G_+P\tau = \tau \,, \quad \mathrm{supp}(G_+\sigma) \subseteq J^+_M(\mathrm{supp}(\sigma)) \,. \tag{2-6}$$

Replacing pc and $+$ with fc and $-$, we get the properties of the extended Green function G_-. Similarly, the Green operators $G^*_\pm$ for the formal adjoint P^* admit unique extensions.

Introducing the *causal propagator* $G = G_+ - G_- : \Gamma_c(W) \to \Gamma(V)$ for a Green-hyperbolic operator $P : \Gamma(V) \to \Gamma(W)$ as the difference between the retarded and the advanced Green operators and taking into account the support properties of Green operators (see Definition 2.1), we realize that G maps to $\Gamma_{sc}(V)$, the space of sections with spacelike-compact support. We get the following exact sequence of vector spaces:

$$0 \longrightarrow \Gamma_c(V) \xrightarrow{P} \Gamma_c(W) \xrightarrow{G} \Gamma_{sc}(V) \xrightarrow{P} \Gamma_{sc}(W) \longrightarrow 0. \tag{2-7}$$

The proof of this fact easily follows from (2-6) and can be found, e.g., in [Bär et al. 2007, Section 3.4], except for surjectivity of $P : \Gamma_{sc}(V) \to \Gamma_{sc}(W)$, which is shown by the following argument; see also [Khavkine 2014b, Proposition 2.1]. Given $\tau \in \Gamma_{sc}(W)$ and taking a partition of unity $\{\chi_+, \chi_-\}$ on M such that $\chi_+ = 1$ in a past-compact region, while $\chi_- = 1$ in a future-compact one, we deduce that $\operatorname{supp}(\chi_+\tau)$ is strictly past-compact, while $\operatorname{supp}(\chi_-\tau)$ is strictly future-compact. Exploiting the extended Green operators, we are able to introduce a section

$$\sigma = G_+(\chi_+\tau) + G_-(\chi_-\tau) \in \Gamma_{sc}(V)$$

such that $P\sigma = \tau$. This is a direct consequence of (2-6).

An exact sequence similar to (2-7) holds true for the causal propagator G^* of the formal adjoint P^* as well.

Extended Green operators provide also an extension of the causal propagator $G : \Gamma_{tc}(W) \to \Gamma(V)$. Minor modifications to the proof of (2-7) give the following exact sequence:

$$0 \longrightarrow \Gamma_{tc}(V) \xrightarrow{P} \Gamma_{tc}(W) \xrightarrow{G} \Gamma(V) \xrightarrow{P} \Gamma(W) \longrightarrow 0. \tag{2-8}$$

This sequence is particularly useful to characterize the space of solutions to the equation $P\sigma = 0$ for $\sigma \in \Gamma(V)$, that is to say $\ker(P)$, the kernel of P. As a matter of fact, (2-8) entails that G induces an isomorphism from $\Gamma_{tc}(W)/P(\Gamma_{tc}(V))$ to $\ker(P)$.

Following [Bär et al. 2007, Section 4.3], one can also relate Green operators over different globally hyperbolic spacetimes provided that the corresponding differential operators are related by vector bundle maps covering a causal embedding between the bases. Specifically, suppose we are given vector bundle maps $C : V_1 \to V_2$ and $D : W_1 \to W_2$ preserving the inner products of the relevant vector bundles and which cover a causal embedding $f : M_1 \to M_2$ between globally hyperbolic spacetimes. Exploiting invertibility of vector bundle maps when restricted to a

fiber, we can define maps between spaces of sections:

$$C^\Gamma : \Gamma(V_2) \to \Gamma(V_1)\,, \qquad \sigma_2 \mapsto C^{-1} \circ \sigma_2 \circ f\,, \tag{2-9}$$

$$C_{\Gamma_c} : \Gamma_c(V_1) \to \Gamma_c(V_2)\,, \qquad \tau_1 \mapsto C \circ \tau_1 \circ f^{-1}, \tag{2-10}$$

and similarly for D. Furthermore, consider Green-hyperbolic differential operators $P_i : \Gamma(V_i) \to \Gamma(W_i)$, $i \in \{1, 2\}$, such that $P_1 C^\Gamma = D^\Gamma P_2$. This simply means that C and D are compatible with the differential operators P_1 over M_1 and P_2 over M_2. Denoting the retarded/advanced Green operators for P_i with $G_{i\pm} : \Gamma_c(W_i) \to \Gamma(V_i)$, we can introduce $H_{1\pm} = C^\Gamma G_{2\pm} D_{\Gamma_c} : \Gamma_c(W_1) \to \Gamma(V_1)$ and compare it with $G_{1\pm}$. Exploiting the fact that f is a causal embedding, it is easy to check that $H_{1\pm}$ fulfills the requirements in Definition 2.1, hence it is a retarded/advanced Green operator for P_1. By uniqueness, we conclude that $H_{1\pm} = G_{1\pm}$. Therefore, we have established a relation between $G_{1\pm}$ and $G_{2\pm}$, namely

$$C^\Gamma G_{2\pm} D_{\Gamma_c} = G_{1\pm}. \tag{2-11}$$

2C. ***General local covariance.*** We recall here the definition of a *locally covariant quantum field theory* according to [Brunetti et al. 2003] and briefly provide some motivation for this axiomatic approach to quantum field theory on curved spacetimes. This requires some basic notions coming from category theory, which can be found, e.g., in [MacLane 1971, Chapter 1].

We first introduce the relevant categories. As a source, we take a category GHyp having globally hyperbolic spacetimes M as objects and causal embeddings $f : M \to N$ as morphisms; see Section 2A. This category provides the physical background where it is possible to sensibly discuss field theory, essentially because objects in this category possess a structure which is rich enough to make sense of initial value problems for hyperbolic partial differential equations, while morphisms are sufficiently well-behaved to allow us to relate Cauchy problems defined on different objects; see the end of Section 2B. The target category Alg is an algebraic one. Objects are unital $*$-algebras and morphisms are unit-preserving $*$-homomorphisms. Originally, morphisms in Alg where required to be injective, however we give up this assumption for reasons which will be clear later on. Objects in Alg are interpreted as the algebras of observables of a quantum field theory, while morphisms provide relations between different algebras arising from causal embeddings between globally hyperbolic spacetimes.

Definition 2.3. A locally covariant quantum field theory (LCQFT) is a functor $\mathscr{A} : \mathsf{GHyp} \to \mathsf{Alg}$ fulfilling both *causality* and the *time slice axiom*.

Causality axiom: For each $f_1 : M_1 \to N$ and $f_2 : M_2 \to N$ in GHyp such that $f_1(M_1) \cap J_N(f_2(M_2)) = \varnothing$, we have $[\mathscr{A}(f_1)a_1, \mathscr{A}(f_2)a_2] = 0$ for each $a_1 \in \mathscr{A}(M_1)$ and $a_2 \in \mathscr{A}(M_2)$.

Time slice axiom: For each $f : M \to N$ in GHyp such that $f(M)$ includes a spacelike Cauchy surface for N, $\mathscr{A}(f) : \mathscr{A}(M) \to \mathscr{A}(N)$ is an isomorphism in Alg.

The functor $\mathscr{A}$ is interpreted in the following way: for each spacetime M, $\mathscr{A}$ assigns an algebra of observables $\mathscr{A}(M)$ defining the quantum field theory on M. Furthermore, whenever we have a causal embedding $f : M \to N$, $\mathscr{A}$ provides a $*$-homomorphism $\mathscr{A}(f)$ relating observables on the spacetime M to their counterparts on N.

The original restriction to injective morphisms in Alg (not considered here) was meant to interpret globally hyperbolic subregions of a given spacetime as subsystems at the algebraic level (this property is often called *isotony*). Actually, even for those examples where the requirement of injectivity is violated [Dappiaggi and Lang 2012; Sanders et al. 2014; Benini et al. 2014a], one can recover the Haag–Kastler axioms [Haag and Kastler 1964; Dimock 1980; Benini et al. 2013] (and their interpretation in terms of subsystems) regarding a fixed spacetime as the full system and regions of this spacetime as subsystems; see [Benini et al. 2014a, Section 5].

As we will see later, injectivity does not hold in the case of the vector potential too. For this reason in the present context we refrain from requiring injectivity for the morphisms in Alg.

Causality entails that observables in causally disjoint regions can be tested independently. This condition implements the requirement that no physical information can propagate faster than light, hence a *measurement* localized in some region cannot affect other measurements which are localized in causally disjoint regions.

To conclude, the time slice axiom can be interpreted as a statement about the content of the algebra of observables on a given spacetime. It means that all observables on a given spacetime N can be equivalently described by taking a globally hyperbolic neighborhood M of any spacelike Cauchy surface in N. This behavior mimics the one of an initial value problem, where each solution is completely determined by its values in the vicinity of some spacelike Cauchy surface.

For the last part of this subsection we focus the attention on the relative Cauchy evolution. For a locally covariant quantum field theory, such a construction is made possible by the time slice axiom. The notion was introduced in [Brunetti et al. 2003, Section 4], where its relation with the quantized energy-momentum tensor was explicitly computed in the case of the Klein–Gordon field.

Suppose a locally covariant quantum field theory $\mathscr{A} : \mathsf{GHyp} \to \mathsf{Alg}$ is given. Exploiting the time slice axiom, one can define the relative Cauchy evolution. We follow here the approach of [Fewster and Verch 2012, §3.4], where the construction presented below is described in full detail.

Given a globally hyperbolic spacetime $(M, g, \mathfrak{o}, \mathfrak{t})$, we introduce the set of hyperbolic perturbations $\mathsf{hp}(M)$ of M as the set of compactly supported symmetric

covariant 2-tensors h on M such that $g_h = g + h$ is a time-orientable Lorentzian metric on M and $(M, g_h, \mathfrak{o}, \mathfrak{t}_h)$ is a globally hyperbolic spacetime, where $\mathfrak{t}_h$ is the unique time-orientation for g_h agreeing with the original time-orientation $\mathfrak{t}$ outside $\mathrm{supp}(h)$.

Remark 2.4. As shown in [Beem et al. 1996, Section 7.1], $\mathsf{hp}(M)$ contains an open neighborhood of the zero section in the space of compactly supported covariant symmetric 2-tensors endowed with the test function topology. In particular it makes sense to endow $\mathsf{hp}(M)$ with the topology induced as a subset of the space of compactly supported covariant symmetric 2-tensors.

Given a globally hyperbolic spacetime M and a perturbation $h \in \mathsf{hp}(M)$, we indicate with $\widetilde{M}$ the globally hyperbolic spacetime obtained perturbing the metric of M as above. Denoting with K the support of h, we introduce two globally hyperbolic spacetimes $M_\pm = M \setminus J^\mp_M(K)$, which will act as intermediaries between M and $\widetilde{M}$ at the algebraic level, making it possible to account for the effect of the metric perturbation h on the space of observables $\mathscr{A}(M)$ associated to the original spacetime.

The construction proceeds observing that $M_\pm$ can be causally embedded in both M and $\widetilde{M}$ according to the following diagram:

$$\begin{array}{ccccc} & & M & & \\ & {\scriptstyle i_-}\nearrow & & \nwarrow{\scriptstyle i_+} & \\ M_- & & & & M_+ \\ & {\scriptstyle j_-}\searrow & & \swarrow{\scriptstyle j_+} & \\ & & \widetilde{M} & & \end{array} \tag{2-12}$$

This construction is pictorially represented in Figure 1. Spacelike Cauchy surfaces for $M_\pm$ are spacelike Cauchy surfaces for M and $\widetilde{M}$ too, as it can be checked directly from the definition of a Cauchy surface. Therefore the causal embeddings $i_\pm$ and $j_\pm$ fulfill the hypotheses in the statement of the time slice axiom, hence, applying the functor $\mathscr{A}$ to the diagram in (2-12), we get isomorphisms in Alg. This fact gives us the opportunity to define a special automorphism of $\mathscr{A}(M)$, namely the relative Cauchy evolution associated to the perturbation $h \in \mathsf{hp}(M)$:

$$R_h = \mathscr{A}(i_-)\mathscr{A}(j_-)^{-1}\mathscr{A}(j_+)\mathscr{A}(i_+)^{-1} : \mathscr{A}(M) \to \mathscr{A}(M)\,. \tag{2-13}$$

R_h is interpreted as the automorphic action induced by the metric perturbation h on the space of observables $\mathscr{A}(M)$, which is assigned to the globally hyperbolic spacetime M.

In some sense, the relative Cauchy evolution provides the feedback at the level of observables induced by a modification of the metric localized in a compact region.

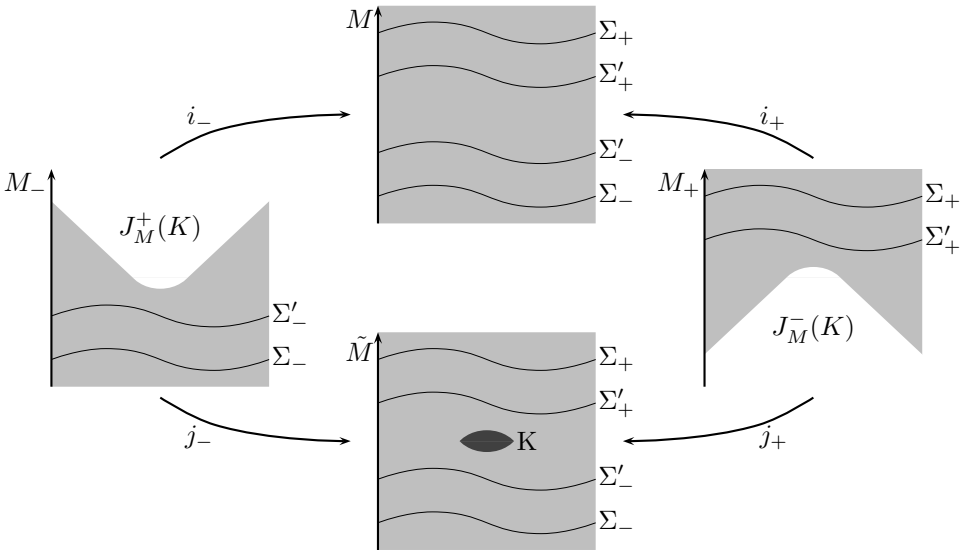


Figure 1. Pictorial representation of the globally hyperbolic spacetimes involved in the definition of the relative Cauchy evolution.

This is realized via $M_\pm$ in the following way. An observable on M is mapped to $\widetilde{M}$ via M_+ exploiting the time slice axiom. In $\widetilde{M}$ the observable is *propagated through* the region where the metric is perturbed, hence it is affected by the perturbation itself. This operation is performed in order to go back to the original spacetime M via M_-, instead of following the same path in the opposite direction via M_+ (which leads to a trivial result). Once back to the unperturbed spacetime M, one can compare the original observable with the one given by the relative Cauchy evolution in order to evaluate the effect of the metric perturbation.

2D. *$*$-algebras and states.* As stated in the previous subsection, locally covariant quantum field theories are functors taking values in an appropriate category of $*$-algebras.

However, as we will see in the case of the vector potential, such a functor can be obtained quantizing a classical analogue of a locally covariant quantum field theory. This operation is performed introducing a quantization functor.

For the case we are interested in, we use a Alg-valued quantization functor $\mathbb{Q}$ defined on the category PSym of presymplectic spaces. Objects of this category are pairs (V, σ), where V is a vector space and σ is a presymplectic form on V, that is to say a (possibly degenerate) antisymmetric bilinear form on V; while morphisms $L : (V_1, \sigma_1) \to (V_2, \sigma_2)$ are linear maps $L : V_1 \to V_2$ preserving the presymplectic structures σ_1 and σ_2, namely such that $\sigma_2(Lv, Lv') = \sigma_1(v, v')$ for each $v, v' \in V_1$.

Given a presymplectic space (V, σ), we define a unital $*$-algebra $\mathfrak{Q}(V, \sigma)$ using the Borchers–Uhlmann construction ([Borchers 1962; Uhlmann 1962]; see also [Brunetti et al. 2003, §2.6] or [Benini et al. 2014c, Appendix A]). To each element $v \in V$, we assign a hermitian symbol $\phi(v) = \phi(v)^*$. Then we take the unital $*$-algebra F freely generated over the field $\mathbb{C}$ by the symbols $\phi(v)$, $v \in V$, and $\mathbb{1}$, the unit of the resulting algebra F. Singling out the $*$-ideal generated by elements of the form

$$\phi(av + bw) - a\phi(v) - b\phi(w)\,, \quad v, w \in V\,,\ a, b \in \mathbb{C}\,, \tag{2-14}$$

$$\phi(v)\phi(w) - \phi(w)\phi(v) - \mathrm{i}\sigma(v, w)\mathbb{1}\,, \quad v, w \in V\,, \tag{2-15}$$

from the freely generated $*$-algebra F, we get $\mathfrak{Q}(V, \sigma)$, the *algebra of canonical commutation relations* associated to the presymplectic space (V, σ).

Remark 2.5. Note that (2-14) entails linearity of the implicitly defined *quantization map*

$$\phi : (V, \sigma) \to \mathfrak{Q}(V, \sigma)\,, \quad v \mapsto \phi(v), \tag{2-16}$$

while (2-15) is used to enforce the usual canonical commutation relations for bosonic field theories, which is also the case for the vector potential of electromagnetism.

A morphism $L : (V_1, \sigma_1) \to (V_2, \sigma_2)$ induces a unit-preserving $*$-homomorphism at the level of the freely generated $*$-algebras. This is specified on generators by setting $\phi(v_1) \mapsto \phi(Lv_1)$ for $v_1 \in V_1$, and $\mathbb{1}_1 \mapsto \mathbb{1}_2$. The obtained $*$-homomorphism naturally descends to the quotients by the $*$-ideals generated by (2-14) and (2-15); therefore we get a morphism $\mathfrak{Q}(L) : \mathfrak{Q}(V_1, \sigma_1) \to \mathfrak{Q}(V_2, \sigma_2)$ in Alg.

One can easily check that $\mathfrak{Q} : \mathsf{PSym} \to \mathsf{Alg}$ is a covariant functor, namely $\mathfrak{Q}(\mathrm{id}_{(V,\sigma)}) = \mathrm{id}_{\mathfrak{Q}(V,\sigma)}$ for each object (V, σ) in PSym and $\mathfrak{Q}(L'L) = \mathfrak{Q}(L')\mathfrak{Q}(L)$ for each pair of composable morphisms L, L' in PSym.

Remark 2.6. No topological information has been taken into account in the present construction. Actually, endowing a presymplectic space (V, σ) with a topology (coherently, the presymplectic form σ has to be continuous), one can consider on the $*$-algebra $\mathfrak{Q}(V, \sigma)$ the topology induced by the construction above. The freely generated algebra F is a direct limit; therefore it carries the topology canonically induced from that of V, thus becoming a topological $*$-algebra. That done, one should consider the quotient by the topological closure of the $*$-ideal generated by (2-14) and (2-15) in order to get a topological $*$-algebra $\mathfrak{Q}(V, \sigma)$.

Let us mention that one might also adopt other quantization procedures, leading to much more regular algebras such as C^*-algebras. This works even in the case of arbitrary presymplectic groups; see [Manuceau et al. 1973].

A $*$-algebra is not enough for the physical description of a quantum field theory. One needs also an algebraic notion of state in order to evaluate the expectation value of an observable.

Definition 2.7. A state ω on a unital $*$-algebra A is a normalized positive linear functional on A, namely $\omega : A \to \mathbb{C}$ is a linear map such that $\omega(a^*a) \geq 0$ for each $a \in A$ and $\omega(\mathbb{1}) = 1$.

A detailed analysis about algebraic states and their properties in relation to quantum field theory can be found, e.g., in [Bratteli and Robinson 1987; Bär et al. 2007; Bär and Becker 2009]. We would like to stress only one feature of algebraic states, that is the capability of reconstructing the usual Hilbert space representation of a quantum field theory exploiting the Gelfand–Naimark–Segal (GNS) theorem. Here we briefly recall this construction for a $*$-algebra without paying attention to topology. Some details for the case of topological $*$-algebras can be found in [Benini et al. 2013], while for the richer case of C^*-algebras see the references mentioned above.

Theorem 2.8. *Let A be a $*$-algebra and consider a state ω on A. Then there exist a Hilbert space H, a dense subspace $D \subseteq H$, a vector $\Omega \in D$ with norm 1 and a $*$-representation π of A by (possibly unbounded) linear maps on H such that $\pi(A)\Omega = D$ and $\langle \Omega, \pi(\cdot)\Omega \rangle = \omega$, where $\langle \cdot\,, \cdot \rangle$ denotes the inner product on H. Moreover, the triple (D, π, Ω) with the properties mentioned above, called a* GNS triple, *is unique up to unitary equivalence.*

Proof. It is possible to define a positive semidefinite sesquilinear form on A (here regarded as a vector space only) exploiting positivity of the state:

$$\langle \cdot\,, \cdot \rangle : A \times A \to \mathbb{C}\,, \quad (a, b) \mapsto \omega(b^*a)\,.$$

Yet, $\langle \cdot\,, \cdot \rangle$ is degenerate if $N = \{a \in A : \omega(a^*a) = 0\} \neq \{0\}$. Hermiticity follows from the fact that $\omega((a + \lambda b)^*(a + \lambda b)) \geq 0$ is a real number for each $a, b \in A$ and $\lambda \in \mathbb{C}$ (choose $\lambda = 1$ and $\lambda = \mathrm{i}$). In particular, a Cauchy–Schwarz inequality for $\langle \cdot\,, \cdot \rangle$ can be established, namely $|\langle a, b\rangle|^2 \leq \langle a, a\rangle\langle b, b\rangle$ for each $a, b \in A$. We deduce that $\langle a, a\rangle = 0$ entails $\langle a, b\rangle = 0$ for each $b \in A$. Therefore,

$$N = \{a \in A : \langle a, b\rangle = 0,\ \forall b \in A\}$$

is a vector subspace of A and we can consider the vector space $D = A/N$. By definition of N, the form $\langle \cdot\,, \cdot \rangle$ descends to the quotient D as a positive definite sesquilinear form. Thus D becomes a pre-Hilbert space. We denote its completion with H, which is a Hilbert space.

Notice that N is left-invariant, namely $aN \subseteq N$ for each $a \in A$. This follows from $\langle an, b\rangle = \omega(b^*an) = \langle n, a^*b\rangle = 0$ for each $a, b \in A$ and $n \in N$. This fact

makes it possible to represent elements of A by linear maps on H:

$$\pi : A \to L(H)\,, \quad \pi(a)[b] = [ab]\,,\ \forall a, b \in A\,.$$

It is easy to check the identities $\pi(\mathbb{1}) = \mathrm{id}_H$ and $\pi(ab) = \pi(a)\pi(b)$. Furthermore, $\langle [a^*b], [c]\rangle = \langle [b], [ac]\rangle$ entails that $\pi(a^*)$ defines the adjoint of $\pi(a)$, thus proving that π is a $*$-representation of A on H.

The unit $\mathbb{1} \in A$ defines a distinguished vector $\Omega = [\mathbb{1}] \in H$ of norm 1, which allows us to reconstruct the algebraic state, $\langle \Omega, \pi(a)\Omega\rangle = \omega(a)$ for each $a \in A$. Moreover, $\pi(A)\Omega = D$ by definition.

Suppose that another triple (D', π', Ω') satisfying the same properties is given. We define $U : D \to D'$ by $U\pi(a)\Omega = \pi'(a)\Omega'$ for each $a \in A$. The following identity, which holds true for each $a, b \in A$, entails that U is well-defined as a linear map and preserves the scalar product:

$$\langle \pi'(a)\Omega', \pi'(b)\Omega'\rangle' = \omega(b^*a) = \langle \pi(a)\Omega, \pi(b)\Omega\rangle\,.$$

In particular, this entails that U is bounded and thus it has a unique extension to the completions. In this way, we obtain $U : H \to H'$, which is linear and continuous. Similarly, one can define $V : D' \to D$ as $V\pi'(a)\Omega' = \pi(a)\Omega$ for each $a \in A$. Then, V has the same properties as U. In particular, it preserves the scalar products and it admits a unique linear and continuous extension $V : H' \to H$. From the definitions of U and V, it is easy to check that V is the inverse of U on the dense subspaces D and D'; hence the same is true everywhere on H and H'. We conclude that $U : H \to H'$ is a unitary equivalence such that $U\pi(\cdot) = \pi'(\cdot)U$. □

Remark 2.9. Even though the GNS construction can be carried out for $*$-algebras without any topology or taking into account a noncontinuous state, it turns out that operators representing elements of the $*$-algebra might be unbounded. This is not the case for more regular algebras (such as C^*-algebras) and continuous states.

Algebraic states provide the correct tool to evaluate expectation values of quantum observables. A quite large number of states is available for the algebra of a quantum field theory, yet not all of them exhibit a reasonable behavior from a physical perspective. A good criterion to select physically sensible states might be to mimic the properties shared by states for quantum field theories on Minkowski spacetime. Just to mention the most prominent states in this context, one encounters the vacuum, associated multi-particle states, coherent states and thermal equilibrium states as well. All these states share a peculiar behavior at very short distances, which plays a central role in the construction of Wick polynomials (which, in turn, provide an essential tool both to define physical quantities such as the quantized energy-momentum tensor and to discuss interacting models in a perturbative fashion). Therefore such short distance behavior seems to be vital for

quantum field theory on Minkowski spacetime. An ultraviolet behavior of this kind is mathematically described by the so-called *Hadamard condition*. Fortunately, this condition on the high frequency part of the 2-point correlation function associated to a state has a natural counterpart on curved spacetimes, even though there is no coordinate independent notion of the Fourier transform. Seminal papers about this topic are [Radzikowski 1996a; 1996b], where tools from *microlocal analysis* [Hörmander 2003, Chapter 8] were employed in order to circumvent the lack of a good notion of Fourier transform. Since then, several techniques to construct Hadamard states on globally hyperbolic spacetimes for various field theoretical models were developed. We refer the reader to the very rich literature cited in [Benini et al. 2013, §4.3], where a concise review of the Hadamard condition, as well as a rich collection of distinguished examples of Hadamard states, can be found.

3. Dynamics for the vector potential

In this section we describe the classical field theory of the electromagnetic vector potential over a d-dimensional globally hyperbolic spacetime M. An approach similar to the one presented below can be found in [Fewster and Pfenning 2003; Dappiaggi 2011; Benini 2014].

The relevant vector bundles for this model are the exterior tensor powers $\bigwedge^k T^*M$ of the cotangent bundle T^*M. For each $k \in \mathbb{N}$, $\bigwedge^k T^*M$ can be canonically endowed with a nondegenerate inner product induced by the metric and the orientation of M. Denoting the exterior product with $\wedge : \bigwedge^k T^*M \times \bigwedge^{k'} T^*M \to \bigwedge^{k+k'} T^*M$ and introducing the Hodge dual $* : \bigwedge^k T^*M \to \bigwedge^{d-k} T^*M$ using the background metric g and the orientation $\mathfrak{o}$, we get a nondegenerate inner product $\langle \cdot , \cdot \rangle = *^{-1}(\cdot \wedge *\cdot)$ on $\bigwedge^k T^*M$.

As is customary, we denote the space of sections of $\bigwedge^k T^*M$ (i.e., k-forms over M) by $\Omega^k(M) = \Gamma(\bigwedge^k T^*M)$, The inner product $\langle \cdot , \cdot \rangle$ on the vector bundle $\bigwedge^k T^*M$, together with the volume form $\mathrm{vol} = *1$, defines an inner product $(\cdot , \cdot)$ on k-forms. Explicitly, we have

$$(\alpha, \beta) = \int_M \alpha \wedge *\beta , \tag{3-1}$$

for $\alpha, \beta \in \Omega^k(M)$ with compact overlapping supports.

On k-forms one has the differential $\mathrm{d} : \Omega^k(M) \to \Omega^{k+1}(M)$ and the codifferential $\delta = (-1)^k *^{-1} \mathrm{d} * : \Omega^k(M) \to \Omega^{k-1}(M)$. It is important to notice that $\mathrm{dd} = 0$, hence $\delta\delta = 0$, too. Moreover, one can directly check that δ is the formal adjoint of d with respect to $(\cdot , \cdot)$.

Using d and δ, one can introduce the Laplace–de Rham operator $\Box = \delta\mathrm{d} + \mathrm{d}\delta$ on k-forms, which is a differential operator of Green-hyperbolic type. It is easy

to check that $\Box$ is formally self-adjoint with respect to $(\cdot\,,\cdot)$ on account of the properties of d and δ. This means that $\Box^* = \Box$, hence a similar relation holds true for the corresponding retarded/advanced Green operators, namely $G_\pm^* = G_\pm$. Therefore one also has

$$(G_+\alpha, \beta) = (\alpha, G_-\beta)\,, \tag{3-2}$$

for each k-form α with past-compact support and each k-form β with strictly future-compact support. A similar result can be obtained by interchanging future and past. From $\mathrm{dd} = 0$ and $\delta\delta = 0$, the identities for the Laplace–de Rham operator $\Box$ and the Green operators $G_\pm$ follow:

$$\Box\mathrm{d} = \mathrm{d}\Box, \qquad \Box\delta = \delta\Box, \tag{3-3}$$

$$\mathrm{d}G_\pm = G_\pm\mathrm{d}, \quad \delta G_\pm = G_\pm\delta\,, \tag{3-4}$$

where we use the same symbols to denote the operators acting on forms of different rank. As an example, we show how to prove the first identity involving Green operators. Take any $\alpha \in \Omega^k_{\mathrm{pc}}(M)$ and any $\beta \in \Omega^{k+1}_{\mathrm{c}}(M)$ and compute $(\mathrm{d}G_+\alpha, \beta)$, exploiting the properties of the Green operators, formal self-adjointness of $\Box$ as well as the identity $\Box\mathrm{d} = \mathrm{d}\Box$:

$$\begin{aligned}(\mathrm{d}G_+\alpha, \beta) &= (\mathrm{d}G_+\alpha, \Box G_-\beta) = (\Box\mathrm{d}G_+\alpha, G_-\beta)\\ &= (\mathrm{d}\Box G_+\alpha, G_-\beta) = (\mathrm{d}\alpha, G_-\beta) = (G_+\mathrm{d}\alpha, \beta)\,.\end{aligned}$$

Hence $\mathrm{d}G_+ = G_+\mathrm{d}$, since $\alpha \in \Omega^k_{\mathrm{pc}}(M)$ and $\beta \in \Omega^{k+1}_{\mathrm{c}}(M)$ are arbitrary.

The Lagrangian density of electromagnetism $\mathscr{L}$ is expressed in terms of the Faraday tensor, $F \in \Omega^2(M)$:

$$\mathscr{L} = *\langle F, F\rangle = F \wedge *F\,. \tag{3-5}$$

The Euler–Lagrange equations derived from $\mathscr{L}$ state that F is a closed and coclosed 2-form; that is to say, $\mathrm{d}F = 0$ and $\delta F = 0$. In the following we consider only exact Faraday tensors, namely we assume there exists a vector potential $A \in \Omega^1(M)$ such that $\mathrm{d}A = F$. As a consequence, the first equation $\mathrm{d}F = 0$ automatically holds true. The second equation remains to be checked, thus providing the dynamics of the vector potential, namely $\delta\mathrm{d}A = 0$. Nevertheless, since the relevant object in electromagnetism is the Faraday tensor, we are forced to consider equivalence classes of vector potentials. As a matter of fact, two vector potentials A and A' differing by $\mathrm{d}\varphi$, $\varphi \in \mathrm{C}^\infty(M)$, give rise to the same Faraday tensor $F = \mathrm{d}A$. For this reason, we consider gauge classes of vector potentials defined according to the equivalence relation

$$A \sim A' \iff \exists\varphi \in \mathrm{C}^\infty(M) \text{ such that } A' = A + \mathrm{d}\varphi\,; \tag{3-6}$$

that is, $A, A' \in \Omega^1(M)$ are considered the same whenever they differ by an exact 1-form.

Let us mention that there are several other approaches to electromagnetism on curved spacetimes. This model was analyzed directly from the perspective of the Faraday tensor in [Dappiaggi and Lang 2012]. In [Dappiaggi and Siemssen 2013] and [Fewster and Lang 2014] the approaches are similar to the present one, except for the notion of gauge equivalence, which is provided there by closed 1-forms instead of exact ones. The present setting is adopted in [Sanders et al. 2014], where also external source currents are dealt with. Some arguments can be found there to motivate our choice of gauge symmetry (3-6). A more geometrical perspective, much in the spirit of Yang–Mills theory, can be found in [Benini et al. 2014b], subsequently refined in [Benini et al. 2014a], in order to correctly address the Aharonov–Bohm effect as well as magnetic monopoles.

In Section 2B we collected much of the material needed to characterize spaces of solutions for Green-hyperbolic equations in terms of Green operators. However this is not enough in the present setting for two reasons. First, the linear differential operator $\delta\mathrm{d}$ ruling the dynamics of the vector potential is not Green-hyperbolic. Second, we have to deal with gauge equivalence too. To overcome such hindrances, we are going to exploit gauge symmetry in order to show that equivalence classes of vector potentials satisfying the dynamics, $\delta\mathrm{d}A = 0$, can be represented adopting the Lorenz gauge, that is to say $\delta A = 0$. Later, exploiting the fact that $\Box = \delta\mathrm{d} + \mathrm{d}\delta$ is Green-hyperbolic and realizing that on-shell vector potentials in the Lorenz gauge satisfy $\Box A = 0$, we provide a characterization of the space of gauge equivalence classes of solutions via the causal propagator $G = G_+ - G_-$ for $\Box$, slightly extending a result in [Dappiaggi 2011]. A more systematic treatment of gauge theories can be found in [Hack and Schenkel 2013].

Lemma 3.1. *Let M be a globally hyperbolic spacetime. Denote the space of solutions to the equation ruling the dynamics of the vector potential by*

$$\mathcal{S} = \ker(\delta\mathrm{d} : \Omega^1(M) \to \Omega^1(M)).$$

Moreover, use $\mathcal{G} = \mathrm{d}\mathrm{C}^\infty(M)$ to denote the space of gauge transformations. Then for each gauge class of solutions $[A] \in \mathcal{S}/\mathcal{G}$, there exists a representative $\hat{A} \in [A]$ in the Lorenz gauge; that is to say $\delta\hat{A} = 0$.

Proof. Let $A \in [A]$ be any representative and consider the equation $\Box\varphi + \delta A = 0$ for $\varphi \in \mathrm{C}^\infty(M)$. We can easily write down a solution of this equation by fixing a partition of unity $\{\chi_+, \chi_-\}$ such that $\chi_\pm = 1$ in a past- or future-compact region. With such a partition of unity, we get a solution $\varphi = -G_+(\chi_+\delta A) - G_-(\chi_-\delta A)$. Introducing $\hat{A} = A + \mathrm{d}\varphi$, we conclude that $\delta\mathrm{d}\hat{A} = 0$ and $\delta\hat{A} = \delta A + \Box\varphi = 0$ since $\Box = \delta\mathrm{d}$ on $\mathrm{C}^\infty(M)$. □

Since $\delta G_\pm = G_\pm \delta$ and taking into account (2-8) too, we realize that each coclosed $\alpha \in \Omega^1_{\mathrm{tc}}(M)$ gives rise to a solution $A = G\alpha \in \mathscr{S}$ in the Lorenz gauge. This gives us a hint how to represent the space $\mathscr{S}/\mathscr{G}$ of gauge classes of solutions. Note that in the next proof we will extensively make use of the exact sequence (2-8).

Theorem 3.2. *Let M be a globally hyperbolic spacetime and set*

$$\ker^k_{\mathrm{tc}}\,\delta = \ker(\delta : \Omega^k_{\mathrm{tc}}(M) \to \Omega^{k-1}_{\mathrm{tc}}(M)).$$

Then the causal propagator G for $\Box = \delta\mathrm{d} + \mathrm{d}\delta$ induces the following isomorphism of vector spaces:

$$I : \frac{\ker^1_{\mathrm{tc}}(\delta)}{\delta\mathrm{d}(\Omega^1_{\mathrm{tc}}(M))} \to \frac{\mathscr{S}}{\mathscr{G}}, \quad [\alpha] \mapsto [G\alpha].$$

Proof. As mentioned before the statement of the theorem, G maps $\ker^1_{\mathrm{tc}}(\delta)$ to $\mathscr{S}$. Given $\beta \in \Omega^1_{\mathrm{tc}}(M)$, $G\delta\mathrm{d}\beta = G(\Box - \mathrm{d}\delta)\beta = \mathrm{d}G(-\delta\beta) \in \mathscr{G}$. Therefore G induces a linear map from $\ker^1_{\mathrm{tc}}(\delta)/\delta\mathrm{d}(\Omega^1_{\mathrm{tc}}(M))$ to $\mathscr{S}/\mathscr{G}$.

This map is surjective on account of Lemma 3.1. Given $[A] \in \mathscr{S}/\mathscr{G}$, we find $\hat{A} \in [A]$ such that $\delta\hat{A} = 0$. Since $\Box\hat{A} = 0$, using (2-8) we find $\hat{\alpha} \in \Omega^1_{\mathrm{tc}}(M)$ such that $G\hat{\alpha} = \hat{A}$. From $\delta\hat{A} = 0$, (2-8) entails there exists $\psi \in \mathrm{C}^\infty_{\mathrm{tc}}(M)$ such that $\delta\hat{\alpha} = \Box\psi = \delta\mathrm{d}\psi$. We deduce that $\alpha = \hat{\alpha} - \mathrm{d}\psi \in \ker^1_{\mathrm{tc}}(\delta)$. Moreover, $[G\alpha] = [\hat{A} - \mathrm{d}G\psi] = [A]$.

Given $\alpha \in \ker^1_{\mathrm{tc}}(\delta)$, it remains only to check that $[G\alpha] = 0$ entails $\alpha \in \delta\mathrm{d}(\Omega^1_{\mathrm{tc}}(M))$. By definition, we find $\varphi \in \mathrm{C}^\infty(M)$ such that $\mathrm{d}\varphi = G\alpha$, which entails $\Box\varphi = 0$. Therefore there exists $\psi \in \mathrm{C}^\infty_{\mathrm{tc}}(M)$ such that $G\psi = \varphi$. Hence (2-8) ensures the existence of $\beta \in \Omega^1_{\mathrm{tc}}(M)$ such that $\Box\beta = \alpha - \mathrm{d}\psi$. Applying δ to both sides of the last identity, we get $\Box\delta\beta = -\Box\psi$, hence $\delta\beta = -\psi$. From this we conclude $\alpha = \delta\mathrm{d}\beta \in \delta\mathrm{d}(\Omega^1_{\mathrm{tc}}(M))$. □

Remark 3.3. One might be interested to solutions supported inside a spacelike compact region, namely consider $\mathscr{S}_{\mathrm{sc}} = \{A \in \Omega^1_{\mathrm{sc}}(M) : \delta\mathrm{d}A = 0\}$. In this case the corresponding notion of gauge symmetry is specified by $\mathscr{G}_{\mathrm{sc}} = \mathrm{d}\Omega^1_{\mathrm{sc}}(M)$. Following the same arguments used above, but using the exact sequence (2-7) in place of (2-8), one gets an isomorphism of vector spaces similar to the one presented in Theorem 3.2:

$$I_{sc} : \frac{\ker^1_c(\delta)}{\delta\mathrm{d}(\Omega^1_{\mathrm{c}}(M))} \to \frac{\mathscr{S}_{\mathrm{sc}}}{\mathscr{G}_{\mathrm{sc}}}, \quad [\alpha] \mapsto [G\alpha], \tag{3-7}$$

where $\ker^k_c(\delta) = \ker(\delta : \Omega^k_{\mathrm{c}}(M) \to \Omega^{k-1}_{\mathrm{c}}(M))$. In the next section we will encounter $\ker^1_c(\delta)/\delta\mathrm{d}(\Omega^1_{\mathrm{c}}(M))$ (enriched with more structure) as the space of classical observables for the vector potential. Hence, via I_{sc}, one can interpret $\mathscr{S}_{\mathrm{sc}}/\mathscr{G}_{\mathrm{sc}}$ as the space of observables of the theory. This approach was adopted in [Dimock 1992; Pfenning 2009].

4. Observables and quantization

In this section we first introduce a suitable observables for the vector potential at a classical level. Then we quantize the obtained space of observables adopting the scheme presented in Section 2D.

In order to define observables for the vector potential, we follow the spirit of [Brunetti et al. 2012], where observables are defined as functionals on field configurations. In the case under analysis it is sufficient to take into account only linear functionals, the dynamics being linear. For this reason our approach mimics the one in [Benini et al. 2014c], even though the situation is even simpler, the underlying bundle being a vector bundle. For the first part of this section it is enough to consider M to be a spacetime. When needed, we will also explicitly introduce the assumption of global hyperbolicity.

We start introducing a special class of linear functionals defined on the space of kinematically allowed field configurations, that is to say $\Omega^1(M)$:

$$\mathbb{O}_\alpha : \Omega^1(M) \to \mathbb{R}\,, \quad \mathbb{O}_\alpha(A) = (\alpha, A)\,,$$

where $\alpha \in \Omega^1_{\mathrm{c}}(M)$. We denote the space of such functionals with $\mathscr{E}^{\mathrm{kin}} \simeq \Omega^1_{\mathrm{c}}(M)$. The isomorphism $\alpha \mapsto \mathbb{O}_\alpha$ will be often used as an identification of $\mathscr{E}^{\mathrm{kin}}$ with $\Omega^1_{\mathrm{c}}(M)$. Since vector potentials differing by a gauge transformation are regarded to be equivalent, only gauge invariant functionals are relevant. For this reason we consider

$$\mathscr{E}^{\mathrm{inv}} = \{\mathbb{O}_\alpha \in \mathscr{E}^{\mathrm{kin}} : \mathbb{O}_\alpha(\mathrm{d}\varphi) = 0,\ \forall \varphi \in \mathrm{C}^\infty(M)\}\,.$$

Since $(\alpha, \mathrm{d}\varphi) = (\delta\alpha, \varphi)$ for each $\alpha \in \Omega^1_{\mathrm{c}}(M)$ and $\varphi \in \mathrm{C}^\infty(M)$, we conclude that $\mathscr{E}^{\mathrm{inv}} = \ker^1_c(\delta)$. Up to now, no dynamical information is encoded in the space of gauge invariant functionals. As a matter of fact, $\mathscr{E}^{\mathrm{inv}}$ provides gauge invariant functionals defined on all kinematically allowed field configurations, regardless of the equation of motion $\delta \mathrm{d}A = 0$. In order to the encode dynamics in a dual fashion on gauge invariant functionals, we proceed as follows. First, we consider the formal adjoint of the equation of motion operator $\delta\mathrm{d}$, which is formally self-adjoint since $(\delta\mathrm{d}\alpha, A) = (\alpha, \delta\mathrm{d}A)$ for each $A \in \Omega^1(M)$ and each $\alpha \in \Omega^1_{\mathrm{c}}(M)$. Then, we take the quotient of $\mathscr{E}^{\mathrm{inv}}$ by the image of $(\delta\mathrm{d})^* = \delta\mathrm{d} : \Omega^1_{\mathrm{c}}(M) \to \Omega^1_{\mathrm{c}}(M)$. In this way we obtain the vector space

$$\mathscr{E} = \frac{\mathscr{E}^{\mathrm{inv}}}{\delta\mathrm{d}(\Omega^1_{\mathrm{c}}(M))}\,,$$

which is interpreted as a space of classical observables for the vector potential, the interpretation being motivated by the fact that $\mathscr{E}$ comprises gauge invariant functionals which can be evaluated only on on-shell field configurations $[A] \in \mathscr{S}/\mathscr{G}$. In fact, the evaluation of $[\alpha] \in \mathscr{E}$ on $[A] \in \mathscr{S}/\mathscr{G}$ can be performed choosing arbitrary

representatives in the equivalence classes. Such evaluation is well-defined because each $A \in [A]$ is on-shell, namely such that $\delta \mathrm{d} A = 0$, and each $\alpha \in [\alpha]$ is gauge invariant.

The following theorem shows that $\mathscr{E}$ contains sufficiently many elements in order to distinguish vector potentials up to gauge. Moreover, if M admits a finite good cover, there are no redundant elements of $\mathscr{E}$, namely two different elements cannot take the same values on all field configurations. Therefore $\mathscr{E}$ is *optimal* in the sense of [Benini 2014] as a space of linear classical observables for the model we are considering (under the assumption of existence of a finite good cover for the spacetime M). This result is a special case of [Benini 2014, Theorem 7.6]. For a different, yet equivalent, approach to causally restricted de Rham cohomology, see [Khavkine 2014a].

Theorem 4.1. *Let M be a spacetime and let $[\alpha], [\alpha'] \in \mathscr{E}$ and $[A], [A'] \in \mathscr{S}/\mathscr{G}$.*

(1) *If $\mathbb{O}_{[\beta]}([A]) = \mathbb{O}_{[\beta]}([A'])$ for each $[\beta] \in \mathscr{E}$, then $[A] = [A']$.*

(2) *If M admits a finite good cover and $\mathbb{O}_{[\alpha]}([B]) = \mathbb{O}_{[\alpha']}([B])$ for each $[B] \in \mathscr{S}/\mathscr{G}$, then $[\alpha] = [\alpha']$.*

In view of our quantization prescription (see Section 2D) we want to endow $\mathscr{E}$ with a presymplectic structure. Assuming the spacetime M to be globally hyperbolic and denoting the causal propagator for $\Box$ with G, we define

$$\tau : \mathscr{E} \times \mathscr{E} \to \mathbb{R}, \quad \tau([\alpha], [\beta]) = (\alpha, G\beta), \tag{4-1}$$

where α and β are representatives of $[\alpha]$ and respectively $[\beta]$. The bilinear map $(\,\cdot\,, G\,\cdot\,) : \Omega^1_{\mathrm{c}}(M) \times \Omega^1_{\mathrm{c}}(M) \to \mathbb{R}$ is antisymmetric:

$$(\alpha, G\beta) = -(G\alpha, \beta) = -(\beta, G\alpha), \quad \forall \alpha, \beta \in \Omega^1_{\mathrm{c}}(M).$$

Moreover, for each $\alpha \in \ker^1_c(\delta)$ and $\omega \in \Omega^1_{\mathrm{c}}(M)$, one has

$$(\alpha, G\delta \mathrm{d}\omega) = (\alpha, G(\Box - \mathrm{d}\delta)\omega) = -(\delta\alpha, G\delta\omega) = 0.$$

This shows that τ is well-defined by (4-1), thus providing a presymplectic form on $\mathscr{E}$.

Remark 4.2. The presymplectic form τ is actually degenerate on certain globally hyperbolic spacetimes. Suppose that β lies in $\delta(\Omega^2_{\mathrm{c}}(M) \cap \mathrm{d}\Omega^1_{\mathrm{tc}}(M)) \setminus \delta \mathrm{d}\Omega^1_{\mathrm{c}}(M)$. This means that there exists $\gamma \in \Omega^1_{\mathrm{tc}}(M)$ such that $\delta \mathrm{d}\gamma = \beta$ and $\mathrm{d}\gamma$ has compact support, but there is no $\omega \in \Omega^1_{\mathrm{c}}(M)$ such that $\delta \mathrm{d}\omega = \beta$. Hence $[\beta] \neq 0$ in $\mathscr{E}$; however, by also exploiting (2-8), for each $[\alpha] \in \mathscr{E}$ we have

$$\tau([\alpha], [\beta]) = (\alpha, G\delta \mathrm{d}\gamma) = (\alpha, G(\Box - \mathrm{d}\delta)\gamma) = -(\delta\alpha, G\delta\gamma) = 0,$$

where the last equality follows from $\delta\alpha = 0$.

Obviously, for a globally hyperbolic spacetime with compact Cauchy surfaces such an α cannot exist, timelike compact regions being compact too. However, it is relatively easy to cook up examples of globally hyperbolic spacetimes with noncompact Cauchy surfaces where $\delta(\Omega^2_c(M) \cap \mathrm{d}\Omega^1_{tc}(M)) \setminus \delta\mathrm{d}\Omega^1_c(M)$ is not empty; see [Benini et al. 2014b, Remark 3.9].

Now we want to show that general local covariance (without injectivity) holds true for the field theoretical model considered here. This result will be achieved in two steps. First, we will construct a classical counterpart of a (noninjective) generally covariant quantum field theory for the vector potential. Then our quantization scheme will automatically provide a LCQFT according to Definition 2.3.

To each globally hyperbolic spacetime M we assign the presymplectic space $\mathscr{F}(M) = (\mathscr{E}_M, \tau_M)$ as defined above (note that we included a subscript to keep track of the underlying spacetime). Given a causal embedding $f : M \to N$, we consider the pullback $f^* : \Omega^k(N) \to \Omega^k(M)$ for k-forms and the pushforward $f_* : \Omega^k_c(M) \to \Omega^k_c(N)$ for compactly supported k-forms. f_* intertwines the differential d_M for forms on M with the differential d_N for forms on N, namely $f_*\mathrm{d}_M = \mathrm{d}_N f_*$. Moreover, since f is an isometry, $f_*\delta_M = \delta_N f_*$ as well. Therefore f_* induces a map $\mathscr{F}(f) : \mathscr{E}_M \to \mathscr{E}_N$ between the spaces of observables associated to M and N. It remains only to check that $\mathscr{F}(f)$ preserves the corresponding presymplectic structures τ_M and τ_N. This follows from the last part of Section 2B. Taking into account the present setting, from Section 2B we deduce $f^*G_{N\pm}f_* = G_{M\pm}$, where $G_{M\pm}$ and $G_{N\pm}$ denote the retarded/advanced Green operators for $\Box_M$ and respectively $\Box_N$. Given $[\alpha], [\beta] \in \mathscr{E}_M$, we compute

$$\begin{aligned}\tau_N(\mathscr{F}(f)[\alpha], \mathscr{F}(f)[\beta]) &= (f_*\alpha, G_N f_*\beta)_N = (\alpha, f^*G_N f_*\beta)_M \\ &= (\alpha, G_M\beta)_M = \tau_M([\alpha], [\beta]).\end{aligned}$$

This shows that $\mathscr{F}(f) : \mathscr{F}(M) \to \mathscr{F}(N)$ is a morphism in PSym. One can easily check that $\mathscr{F}(\mathrm{id}_M) = \mathrm{id}_{\mathscr{F}(M)}$ for each object M in GHyp and that $\mathscr{F}(f \circ f') = \mathscr{F}(f) \circ \mathscr{F}(f')$ for each pair of composable morphisms f, f' in GHyp. Therefore we conclude that $\mathscr{F} : \mathsf{GHyp} \to \mathsf{PSym}$ is a functor.

Theorem 4.3. *The functor $\mathscr{F} : \mathsf{GHyp} \to \mathsf{PSym}$ fulfills the classical counterparts of the causality and time slice axioms that are stated below. Yet injectivity of morphisms fails, namely there are morphisms f in GHyp for which $\mathscr{F}(f)$ is not injective.*

Causality axiom: *For each $f_1 : M_1 \to N$ and $f_2 : M_2 \to N$ in GHyp such that $f_1(M_1) \cap J_N(f_2(M_2)) = \varnothing$, we have $\tau_N(\mathscr{F}(f_1)[\alpha_1], \mathscr{F}(f_2)[\alpha_2]) = 0$ for each $[\alpha_1] \in \mathscr{F}(M_1)$ and $[\alpha_2] \in \mathscr{F}(M_2)$.*

Time slice axiom: *For each $f : M \to N$ in GHyp such that $f(M)$ includes a spacelike Cauchy surface for N, $\mathscr{F}(f) : \mathscr{F}(M) \to \mathscr{F}(N)$ is an isomorphism in PSym.*

Proof. A counterexample to injectivity is provided in Remark 5.6 of [Benini et al. 2014b]. In fact, taking into account only the linear part of the classical observables defined in the reference just cited for the case $G = \mathbb{R}$ as structure group, one gets the same space of classical observables which is considered here.

The causality property is a trivial consequence of the support properties of the causal propagator. Under the assumptions of the statement, we have the inclusion

$$\operatorname{supp}(f_{1*}\alpha_1) \cap J_N(\operatorname{supp}(f_{2*}\alpha_2)) \subseteq f_1(M_1) \cap J_N(f_2(M_2)) = \varnothing;$$

hence the supports of $f_{1*}\alpha_1$ and $G_N f_{2*}\alpha_2$ do not overlap. This shows that τ_N vanishes when evaluated on the pair $(\mathscr{F}(f_1)[\alpha_1], \mathscr{F}(f_2)[\alpha_2]) \in \mathscr{E}_M \times \mathscr{E}_M$.

To prove the time slice axiom we take $f : M \to N$ as in the statement and we look for an inverse of $\mathscr{F}(f)$.

As a preparatory step, we introduce a special partition of unity. Let Σ be a spacelike Cauchy surface for N included in $f(M)$. Since f is a causal embedding, $f(M)$ is a globally hyperbolic spacetime with Σ as a spacelike Cauchy surface. According to [Bernal and Sánchez 2005], we can foliate N as $\mathbb{R} \times \Sigma$ and regard $f(M)$ as an open neighborhood of $\{0\} \times \Sigma$ in N. In particular, there are spacelike Cauchy surfaces Σ_+, Σ_- for N of the form $\{t\} \times \Sigma$ which are contained in $f(M)$ and lie respectively inside the chronological future $I_M^+(\Sigma)$ and the chronological past $I_M^-(\Sigma)$ of Σ. We take a partition of unity $\{\chi_+, \chi_-\}$ on N such that $\chi_\pm = 1$ in $J_N^\pm(\Sigma_\pm)$.

Using $\{\chi_+, \chi_-\}$, we define a map $I : \mathscr{E}_N \to \mathscr{E}_M$ according to the following procedure. Given $[\beta] \in \mathscr{E}_N$ and fixing a representative $\beta \in [\beta]$, we can consider the 1-form $\delta\mathrm{d}(\chi_\pm G_\pm \beta) = \beta - \delta\mathrm{d}(\chi_\mp G_\pm \beta)$ and realize its support is compact. Here we exploited the compact support of β and the past-compact/future-compact support of $\chi_\pm$, together with $\delta\beta = 0$ and Definition 2.1. Moreover, note that the left side vanishes in $J_N^\mp(\Sigma_\mp)$. We can also consider the 1-form $\hat{\beta} = \delta\mathrm{d}(\chi_+ G\beta) = -\delta\mathrm{d}(\chi_- G\beta)$ (the second equality follows from $\delta\beta = 0$ and $\Box G\beta = 0$). As it can be easily checked, $\hat{\beta}$ has compact support inside the time slab $J_N^+(\Sigma_-) \cap J_N^-(\Sigma_+) \subseteq f(M)$. Setting $\omega = (\chi_- G_+\beta + \chi_+ G_-\beta) \in \Omega^1_{\mathrm{c}}(N)$, by a direct computation we find $\hat{\beta} + \delta\mathrm{d}\omega = \beta$. Hence $\hat{\beta}$ is a representative of $[\beta]$ and, as we already proved, its support lies inside $f(M)$ allowing us to introduce $\alpha = f^*\hat{\beta} \in \Omega^1_{\mathrm{c}}(M)$ such that $\delta\alpha = 0$. The same argument for another representative $\beta + \delta\mathrm{d}\gamma$ of $[\beta]$, $\gamma \in \Omega^1_{\mathrm{c}}(M)$, would give $\hat{\beta} + \delta\mathrm{d}(\chi_+ G\delta\mathrm{d}\gamma)$. Taking into account $G\Box\gamma = 0$, one gets

$$\delta\mathrm{d}(\chi_+ G\delta\mathrm{d}\gamma) = -\delta\mathrm{d}(\chi_+ \mathrm{d}\delta G\gamma) = -\delta(\mathrm{d}\chi_+ \wedge \mathrm{d}\delta G\gamma) = \delta\mathrm{d}(\mathrm{d}\chi_+ \wedge \delta G\gamma),$$

where $\hat{\gamma} = \mathrm{d}\chi_+ \wedge \delta G\gamma$ has compact support in $J_N^+(\Sigma_-) \cap J_N^-(\Sigma_+) \subseteq f(M)$. This fact follows from $\mathrm{d}\chi_+ = -\mathrm{d}\chi_-$ being supported inside a timelike compact region and $\operatorname{supp}(G\omega)$ being spacelike compact. Introducing $\theta = f^*\hat{\gamma} \in \Omega^1_{\mathrm{c}}(M)$, we conclude that, when starting from $\beta + \delta\mathrm{d}\gamma$, the procedure above provides $\alpha + \delta\mathrm{d}\theta$.

Since $[\alpha + \delta \mathrm{d}\theta] = [\alpha]$ in $\mathscr{E}_M$, we can define the linear map

$$I : \mathscr{E}_N \to \mathscr{E}_M\,, \quad [\beta] \mapsto [f^*\delta \mathrm{d}(\chi_+ G\beta)]\,. \tag{4-2}$$

One can directly check that I is actually the inverse of $\mathscr{F}(f)$ taking $[\alpha] \in \mathscr{E}_M$ and $[\beta] \in \mathscr{E}_N$ and computing $I\mathscr{F}(f)[\alpha]$ and $\mathscr{F}(f)I[\beta]$. In the first formula below we introduce subscripts on G to stress that both the causal propagators for $\Box$ on M and for $\Box$ on N are involved.

$$\begin{aligned} I\mathscr{F}(f)[\alpha] &= I[f_*\alpha] = [f^*\delta \mathrm{d}(\chi_+ G_N f_*\alpha)] = [\delta \mathrm{d}((f^*\chi_+)G_M\alpha)] \\ &= [\delta \mathrm{d}((f^*\chi_+)G_M\alpha) + \delta \mathrm{d}((f^*\chi_+)G_{M-}\alpha) + \delta \mathrm{d}((f^*\chi_-)G_{M+}\alpha)] \\ &= [\delta \mathrm{d}(G_{M+}\alpha)] = [\alpha]\,, \\ \mathscr{F}(f)I[\beta] &= \mathscr{F}(f)[f^*\delta \mathrm{d}(\chi_+ G\beta)] = [\delta \mathrm{d}(\chi_+ G\beta)] = [\beta]\,. \end{aligned}$$

For the first computation we exploited the fact that $(f^*\chi_\pm)G_{M\mp}\alpha$ is a compactly supported 1-form on M, while the second follows from the fact that $\delta \mathrm{d}(\chi_+ G\beta)$ is a representative of $[\beta]$ with support inside $f(M)$, as already shown above. Automatically I preserves the relevant presymplectic forms:

$$\tau_M(I[\beta], I[\beta']) = \tau_N(\mathscr{F}(f)I[\beta], \mathscr{F}(f)I[\beta']) = \tau_N([\beta], [\beta'])\,,$$

for each $[\beta], [\beta'] \in \mathscr{E}_N$. This shows that $I : \mathscr{F}(N) \to \mathscr{F}(M)$ is the inverse of $\mathscr{F}(f) : \mathscr{F}(M) \to \mathscr{F}(N)$ in PSym; therefore $\mathscr{F}(f)$ is an isomorphism in PSym. □

The last part of this section is devoted to the quantization of the functor $\mathscr{F}$ describing the classical field theory of the vector potential. This result is achieved composing $\mathscr{F} : \mathsf{GHyp} \to \mathsf{PSym}$ with the quantization functor $\mathfrak{Q} : \mathsf{PSym} \to \mathsf{Alg}$ for canonical commutation relations presented in Section 2D.

Theorem 4.4. *The functor $\mathscr{A} = \mathfrak{Q} \circ \mathscr{F} : \mathsf{GHyp} \to \mathsf{Alg}$ is a locally covariant quantum field theory according to Definition 2.3.*

Proof. $\mathscr{A}$ is defined by the composition of the covariant functors $\mathfrak{Q} : \mathsf{PSym} \to \mathsf{Alg}$ and $\mathscr{F} : \mathsf{GHyp} \to \mathsf{PSym}$; therefore it is a covariant functor from GHyp to Alg. Causality holds true on account of its classical counterpart fulfilled by $\mathscr{F}$ and the canonical commutation relations implemented by $\mathfrak{Q}$. The time slice axiom for $\mathscr{A}$ follows from the corresponding property of $\mathscr{F}$ and the fact that $\mathfrak{Q}$ is a functor, thus sending isomorphisms of PSym to isomorphisms of Alg. □

Remark 4.5. Up to now, neither $\mathscr{F}(M)$ nor $\mathscr{A}(M)$ were intended as topological spaces. Actually, one can endow $\mathscr{F}(M)$ with the topology induced by the test function topology on $\Omega^1_{\mathrm{c}}(M)$, $\mathscr{F}(M)$ being the quotient by $\delta \mathrm{d}(\Omega^1_{\mathrm{c}}(M))$ of the closed subspace $\mathscr{E}^{\mathrm{inv}}$ of $\Omega^1_{\mathrm{c}}(M)$. At least whenever M admits a finite good cover, the second statement of Theorem 4.1 means that the image of $\delta \mathrm{d} : \Omega^1_{\mathrm{c}}(M) \to \Omega^1_{\mathrm{c}}(M)$ coincides with the intersection of the kernels of the maps $\int_M (\,\cdot\,) \wedge *A : \Omega^1_{\mathrm{c}}(M) \to \mathbb{R}$,

$A \in \ker^1(\delta \mathrm{d})$. Since these maps are continuous, $\delta \mathrm{d}(\Omega^1_{\mathrm{c}}(M))$ is a closed subspace of $\Omega^1_{\mathrm{c}}(M)$. This entails that the topology induced on $\mathscr{F}(M)$ by the quotient is Hausdorff at least when M admits a finite good cover. Continuity of the push-forward along a smooth map and of the Green functions (see [Bär et al. 2007, Section 3.4]) entails that all our conclusions up to this point are compatible with the topological structure presented above, in particular $\mathscr{F}$ is a functor taking values in the category of topological presymplectic spaces. Therefore, Remark 2.6 entails that the functor $\mathscr{A}$ can be regarded as taking values in the category of unital topological $*$-algebras.

Theorem 4.4 provides a satisfactory description of the quantum field theory of the vector potential on each globally hyperbolic spacetime. One still needs Hadamard states for this model. A constructive result in this direction can be found in [Dappiaggi and Siemssen 2013] for asymptotically flat globally hyperbolic spacetimes at future null infinity. Furthermore, the existence of Hadamard states can be argued exploiting a deformation arguments involving ultrastatic spacetimes [Fulling et al. 1978; Fulling et al. 1981], where a complete timelike Killing vector field makes it possible to cook up Hadamard states by means of Fourier transform techniques. This approach was followed in [Fewster and Pfenning 2003]. An extension of the Gupta–Bleuler formalism to curved spacetimes is available too; see [Finster and Strohmaier 2013].

We recall the notion of a quasifree Hadamard state for the vector potential according to [Fewster and Pfenning 2003].

Definition 4.6. For a globally hyperbolic spacetime M, a state ω on the field algebra $\mathscr{A}(M)$ is quasifree and Hadamard if there exists a distribution $w \in \Omega^1_{\mathrm{c}}(M^2)'$ fulfilling the requirements listed below.

(1) w is a $\Box$-bisolution, i.e., $w(\Box\alpha, \beta) = 0 = w(\alpha, \Box\beta)$ for each $\alpha, \beta \in \Omega^1_{\mathrm{c}}(M)$.

(2) $w(\alpha, \beta) - w(\beta, \alpha) = i\tau([\alpha], [\beta])$ for each $\alpha, \beta \in \mathscr{E}^{\mathrm{inv}}$.

(3) The wavefront set of w has the form

$$WF(w) = \left\{(x, k; x', -k') \in \dot{T}^*M^2 : (x, k) \sim (x', k'),\ k \in V_x^+\right\},$$

where $\dot{T}^*$ denotes the cotangent bundle with the zero section removed, V_x^+ is the closed cone of lightlike covectors at x and $(x, k) \sim (x', k')$ means that x is joined to x' by a lightlike geodesic γ, k is the cotangent vector at x of γ and k' is the parallel transport of k along γ.

(4) The two-point function of the state ω is given by w, that is to say, for each $\alpha, \beta \in \mathscr{E}^{\mathrm{inv}}$,

$$\omega(\phi([\alpha])\phi([\beta])) = w(\alpha, \beta).$$

(5) All n-point functions vanish for n odd, while for n even they are completely determined by two-point function, namely, for $\alpha_1, \dots, \alpha_n \in \mathscr{E}^{\mathrm{inv}}$,

$$\omega\big(\phi([\alpha_1]) \cdots \phi([\alpha_n])\big) = \sum_{\pi \in P_n} \prod_{i=1}^{n/2} \omega\big(\phi([\alpha_{\pi(2i-1)}])\phi([\alpha_{\pi(2i)}])\big),$$

where P_n denotes the set of permutations π of $\{1, \dots, n\}$ such that $\pi(2i-1) < \pi(2i+1)$ and $\pi(2i-1) < \pi(2i)$ for each $i \in \{1, \dots, n/2\}$.

5. Relative Cauchy evolution

In this section we relate the relative Cauchy evolution to the energy-momentum tensor of the vector potential, thus extending a result which was originally established for the Klein–Gordon field in [Brunetti et al. 2003] and later shown to hold in the Dirac case as well; see [Sanders 2010].

As a starting point, we fix a globally hyperbolic spacetime M and we consider a representation π of the field algebra $\mathscr{A}(M)$ (meant here as a unital topological $*$-algebra according to Remark 4.5) on a Hilbert space H such that it makes sense to consider the functional derivative of the relative Cauchy evolution R_h with respect to the perturbation h. This is to be intended in the following sense: there exists a dense subspace S of H and a dense unital $*$-subalgebra B of $\mathscr{A}(M)$ such that, for each $\theta \in S$ and $b \in B$, there exists a symmetric contravariant 2-tensor t satisfying the following condition:

$$\int_M \Big(\frac{\mathrm{d}}{\mathrm{d}s}\Big|_0 h_{s\,\mu\nu}\Big)\, t^{\mu\nu}\,\mathrm{vol} = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_0 \big\langle \theta, \pi(R_s b)\theta \big\rangle, \tag{5-1}$$

for each compact set $K \subseteq M$ and each smooth 1-parameter family $s \in (-1, 1) \mapsto h_s \in \mathsf{hp}(M)$ of globally hyperbolic perturbations of M with support inside K, where R_s stands for R_{h_s}.[2] Uniqueness of t follows from Remark 2.4. This allows us to introduce

$$(\theta, b) \in S \times B \;\mapsto\; \Big\langle \theta, \Big(\frac{\delta}{\delta h}\pi(R_h b)\Big)\theta \Big\rangle \doteq t\,,$$

which implicitly defines, for each $b \in B$, the functional derivative $\delta\pi(R_h b)/\delta h$ of the relative Cauchy evolution as a quadratic form on V via the representation π.

As noted in [Brunetti et al. 2003], the GNS representation π_ω induced by a quasifree Hadamard state ω on the field algebra $\mathscr{A}(M)$ fulfills all requirements listed above for defining the functional derivative of the relative Cauchy evolution.

Remark 5.1. Using general arguments, in [Brunetti et al. 2003, Theorem 4.2] it is shown that $\delta\pi(R_h b)/\delta h$ is divergence-free with respect to the Levi–Civita

[2] We will use this notation whenever it is clear from the context which family of perturbations is being taken into account.

connection for the unperturbed metric g. This is a consistency check for the main theorem of this section since the final result consists of an equality between the functional derivative of the relative Cauchy evolution and a term involving the energy-momentum tensor of the electromagnetic field, which is divergence-free.

For convenience, we first state the final result and then we proceed step by step preparing the tools needed later for the proof.

Theorem 5.2. *Let M be a globally hyperbolic spacetime and consider a quasifree Hadamard state ω for the field algebra $\mathcal{A}(M)$ of the vector potential. Consider the GNS triple $(D_\omega, \pi_\omega, \Omega_\omega)$ associated to ω. Then the equality stated below holds true for each $[\alpha] \in \mathcal{F}(M)$ in the sense of quadratic forms on D_ω:*

$$\frac{\delta}{\delta h_{\mu\nu}} \pi_\omega\big(R_h \phi([\alpha])\big) = \frac{\mathrm{i}}{2}\big[\hat{T}^{\mu\nu}, \phi_\omega([\alpha])\big], \tag{5-2}$$

where $\phi_\omega([\alpha]) = \pi_\omega\big(\phi([\alpha])\big)$ is a generator of the field algebra $\mathcal{A}(M)$ represented via π_ω and $\hat{T}^{\mu\nu}$ is the quantized energy-momentum tensor (indices are raised using the background metric) obtained via point-splitting in the GNS representation induced by ω.

Remark 5.3. Instead of using the point-splitting prescription in order to quantize the classical energy-momentum tensor $T_{\mu\nu}$ of the electromagnetic field, defined as the functional derivative of the action with respect to the background metric, one could consider more refined quantization procedures, but the conclusions of Theorem 5.2 would not be affected. For details see the remarks after [Brunetti et al. 2003, Theorem 4.3].

5A. ***Energy-momentum tensor and point-splitting.*** As a starting point, we write down the classical energy-momentum tensor as the functional derivative with respect to the spacetime metric of the action S for the vector potential A, which is defined out of the Lagrangian (3-5):

$$T_{\mu\nu} = \frac{2}{\sqrt{|\det g|}} \frac{\delta S}{\delta g^{\mu\nu}} = F_{\mu\rho} F_\nu{}^\rho - \tfrac{1}{4} g_{\mu\nu} F_{\rho\sigma} F^{\rho\sigma}, \tag{5-3}$$

where indices are raised with respect to the spacetime metric g. The result is a divergence-free symmetric covariant 2-tensor T, as one can easily check taking into account the identities $F = \mathrm{d}A$ and $\delta \mathrm{d} A = 0$.

In Theorem 5.2 the quantized energy-momentum tensor $\hat{T}$ appears. This is obtained from the classical one, namely T, applying the point-splitting prescription [Wald 1994, Section 4.6]:

(1) Separate products of classical fields at the same spacetime point $p \in M$ introducing an auxiliary base point $q \in M$. This is to be intended in the limit

where the two points are close enough. To be more precise, one should consider a point $q \neq p$ in a normal neighborhood of p. This way there exists a unique geodesic connecting p to q. The product at different spacetime points is then properly defined in terms of the parallel transport operator Y along such geodesic;

(2) Replace products of classical fields with matrix elements of products of quantum fields in a Hilbert space representation;

(3) Deal with all computations in the point-split form. Only in the end take the coincidence limit, provided no singularity arises.

To start with, one has to define matrix elements for products of quantum fields in the GNS representation π_ω induced by a quasifree Hadamard state ω. According to Definition 4.6, we note that, given $\xi, \eta \in D_\omega$ and $\alpha_1, \ldots, \alpha_n \in \mathscr{E}^{\mathrm{inv}}$, the matrix element of a product of n fields $\langle \xi, \phi_\omega([\alpha_1]) \cdots \phi_\omega([\alpha_n])\eta\rangle$ can be written as a sum of products of a suitable bidistribution w evaluated on some test-sections in $\mathscr{E}^{\mathrm{inv}}$, among which one finds $\alpha_1, \ldots, \alpha_n$. Therefore, using w, one can define a n-distribution $\langle \xi, \hat{A}(p_1) \cdots \hat{A}(p_n)\eta\rangle \in \Omega^1_{\mathrm{c}}(M^n)'$ satisfying, for each $\alpha_1, \ldots, \alpha_n \in \mathscr{E}^{\mathrm{inv}}$, the identity

$$\int_{M^n} \big\langle \xi, \hat{A}_{\mu_1}(p_1) \cdots \hat{A}_{\mu_n}(p_n)\eta\big\rangle \alpha_1^{\mu_1}(p_1) \cdots \alpha_n^{\mu_n}(p_n)\ \mathrm{vol} = \big\langle \xi, \phi_\omega([\alpha_1]) \cdots \phi_\omega([\alpha_n])\eta\big\rangle, \quad (5\text{-}4)$$

where the integral denotes evaluation of a distribution on a test section and indices are raised with respect to g.

However, $\langle \xi, \hat{A}(p_1) \cdots \hat{A}(p_n)\eta\rangle$ is *not* the only distribution satisfying (5-4). Since only elements of $\mathscr{E}^{\mathrm{inv}} \subseteq \Omega^1_{\mathrm{c}}(M)$ can enter $\langle \xi, \phi_\omega([\alpha_1]) \cdots \phi_\omega([\alpha_n])\eta\rangle$, one can add exact n-distributions to $\langle \xi, \hat{A}(p_1) \cdots \hat{A}(p_n)\eta\rangle$ without affecting the identity (5-4). This ambiguity does not affect the quantization of T since only $F = \mathrm{d}A$ enters (5-3), and therefore Theorem 5.2 is not affected as well.

Remark 5.4. For $n = 1$, $\langle \xi, \hat{A}_\mu(p)\eta\rangle$ is a distribution generated by a smooth function, as observed in [Brunetti et al. 2003, p. 60].

Using (5-4), from items (1)–(2) above we get matrix elements for the quantized energy-momentum tensor:

$$\begin{aligned}
&\langle \xi, \hat{T}_{\mu\nu}(p,q)\eta\rangle \\
&= g^{\rho\sigma}(p) Y_{\nu\sigma}^{\nu'\sigma'}(p,q) \\
&\qquad \times \big[\nabla_\mu^p \nabla_{\nu'}^q \langle \xi, \hat{A}_\rho(p)\hat{A}_{\sigma'}(q)\eta\rangle - (\nu' \leftrightarrow \sigma') - (\mu \leftrightarrow \rho) + (\mu \leftrightarrow \rho \text{ and } \nu' \leftrightarrow \sigma')\big] \\
&\quad - \tfrac{1}{2} g_{\mu\nu}(p) g^{\rho\sigma}(p) g^{\tau\upsilon}(p) Y_{\sigma\upsilon}^{\sigma'\upsilon'}(p,q) \\
&\qquad \times \big[\nabla_\rho^p \nabla_{\sigma'}^q \langle \xi, \hat{A}_\tau(p)\hat{A}_{\upsilon'}(q)\eta\rangle - (\sigma' \leftrightarrow \upsilon')\big]. \quad (5\text{-}5)
\end{aligned}$$

Here the superscript on a covariant derivative indicates the spacetime dependence of the section upon which the covariant derivative is applied; $\mu \leftrightarrow \rho$ stands for a term equal to the one explicitly written before, with indices μ and ρ interchanged. Setting $\eta = \phi_\omega([\alpha])\theta,\ \xi = \theta$ and then $\xi = \theta,\ \eta = \phi_\omega([\alpha])\theta$ in (5-5) and taking the difference between the two outcomes, one gets $\langle \theta, \big[\hat{T}_{\mu\nu}(p,q), \phi_\omega([\alpha])\big]\theta\rangle$, which is the relevant term for Theorem 5.2. A closer look at (5-5) shows that this term can be evaluated once $\langle \theta, \big[\hat{A}_\mu(p)\hat{A}_{\nu'}(q), \phi_\omega([\alpha])\big]\theta\rangle$ is known. This can be obtained from the term $\langle \theta, \big[\phi_\omega([\beta])\phi_\omega([\gamma]), \phi_\omega([\alpha])\big]\theta\rangle$ by extracting a (nonunique) bidistribution as above. Using the canonical commutation relations (2-15), we obtain

$$\langle \theta, \big[\hat{A}_\mu(p)\hat{A}_{\nu'}(q), \phi_\omega([\alpha])\big]\theta\rangle = \mathrm{i}(G\alpha)_\mu(p)\langle \theta, \hat{A}_{\nu'}(q)\theta\rangle + \mathrm{i}(G\alpha)_{\nu'}(q)\langle \theta, \hat{A}_\mu(p)\theta\rangle .$$

As above, the one on the right side is not the only possible choice of a bidistribution representing the left side, yet this ambiguity amounts to an exact bidistribution, and therefore it disappears as soon as we take the appropriate derivatives in order to evaluate $\langle \theta, \big[\hat{T}_{\mu\nu}(p,q), \phi_\omega([\alpha])\big]\theta\rangle$. Already at this stage one realizes no singularity appears in the coincidence limit $p \to q$. Defining $A = G\alpha$, $F = \mathrm{d}A$, $\tilde{A} = \langle \theta, \hat{A}\theta\rangle$ and $\tilde{F} = \mathrm{d}\tilde{A}$, one is led to

$$\langle \theta, \big[\hat{T}_{\mu\nu}(p), \phi_\omega([\alpha])\big]\theta\rangle = \mathrm{i}g^{\rho\sigma}(p)\big(F_{\mu\rho}(p)\tilde{F}_{\nu\sigma}(p) + \tilde{F}_{\mu\rho}(p)F_{\nu\sigma}(p)\big) - \tfrac{\mathrm{i}}{2}g_{\mu\nu}(p)g^{\rho\sigma}(p)g^{\tau\upsilon}(p)F_{\rho\tau}(p)\tilde{F}_{\sigma\upsilon}(p). \quad (5\text{-}6)$$

5B. ***Classical relative Cauchy evolution.*** This subsection is devoted to finding a convenient formula for the relative Cauchy evolution at the classical level. This is defined replacing $\mathscr{A}$ in (2-13) with $\mathscr{F}$, the functor describing the classical field theory of the vector potential. This can be done on account of the time slice axiom Theorem 4.3. In view of the proof of Theorem 5.2, given a globally hyperbolic spacetime M, we fix a compact region $K \subseteq M$ and a 1-parameter family $s \in (-1,1) \mapsto h_s \in \mathsf{hp}(M)$ supported inside K. For each $s \in (-1,1)$, recalling the construction of Section 2C, we consider the globally hyperbolic spacetime $\widetilde{M}_s$, obtained perturbing M with h_s. Moreover, we take spacelike Cauchy surfaces Σ_+, Σ'_+ for $M_+ = M \setminus J^-_M(K)$ and Σ'_-, Σ_- for $M_- = M \setminus J^+_M(K)$ such that $\Sigma'_+ \subseteq I^-_{M_+}(\Sigma_+)$ and $\Sigma'_- \subseteq I^+_{M_-}(\Sigma_-)$; see Figure 1. Consider now the diagram in (2-12). For each causal embedding in this diagram, the functor $\mathscr{F}$ provides a morphism in PSym, which can be inverted according to the time slice axiom. In particular, we are interested in $\mathscr{F}(i_+)^{-1} : \mathscr{F}(M) \to \mathscr{F}(M_+)$ and $\mathscr{F}(j_-)^{-1} : \mathscr{F}(\widetilde{M}_s) \to \mathscr{F}(M_-)$. These maps can be explicitly defined following the proof of the time slice axiom in Theorem 4.3; see (4-2). For i_+ consider the spacelike Cauchy surfaces Σ_+, Σ'_+ and a partition of unity $\{\chi_+, \chi'_+\}$ on M such that $\chi_+ = 1$ in $J^+_M(\Sigma_+)$, while $\chi'_+ = 1$ in $J^-_M(\Sigma'_+)$. To define j_- consider instead Σ'_-, Σ_-, together with a partition of unity $\{\chi'_-, \chi_-\}$ on $\widetilde{M}_s$ such that $\chi'_- = 1$ in

$J^{+}_{\widetilde{M}_s}(\Sigma'_-)$ and $\chi_- = 1$ in $J^{-}_{\widetilde{M}_s}(\Sigma_-)$. Explicit formulas for $\mathscr{F}(j_+) : \mathscr{F}(M_+) \to \mathscr{F}(\widetilde{M}_s)$ and $\mathscr{F}(i_-) : \mathscr{F}(M_-) \to \mathscr{F}(M)$ are obtained simply via pushforward on compactly supported 1-forms as it is explained before Theorem 4.3.

We recall here the explicit form of the maps involved in the definition of the classical relative Cauchy evolution $r_s : \mathscr{F}(M) \to \mathscr{F}(M)$ for the perturbation h_s:

$$\begin{aligned}
\mathscr{F}(i_+)^{-1} &: \mathscr{F}(M) \to \mathscr{F}(M_+), \quad [\alpha] \mapsto [i_+^* \delta \mathrm{d}(\chi_+ G\alpha)],\\
\mathscr{F}(j_+) &: \mathscr{F}(M_+) \to \mathscr{F}(\widetilde{M}_s), \quad [\alpha] \mapsto [j_{+*}\alpha],\\
\mathscr{F}(j_-)^{-1} &: \mathscr{F}(\widetilde{M}_s) \to \mathscr{F}(M_-), \quad [\alpha] \mapsto [-j_-^* \delta_s \mathrm{d}(\chi_- G_s\alpha)],\\
\mathscr{F}(i_-) &: \mathscr{F}(M_-) \to \mathscr{F}(M), \quad [\alpha] \mapsto [i_{-*}\alpha],
\end{aligned}$$

where the subscript s means that the perturbation h_s plays a role. For example G_s is the causal propagator for $\Box_s = \delta_s \mathrm{d} + \mathrm{d}\delta_s$, where h_s enters δ_s via the Hodge dual on $\widetilde{M}_s$, which is defined out of the perturbed metric $g_s = g + h_s$ and the orientation of the underlying manifold. Composing the maps above, one gets a formula for the classical relative Cauchy evolution:

$$r_s : \mathscr{F}(M) \to \mathscr{F}(M), \quad [\alpha] \mapsto [-\delta_s \mathrm{d}(\chi_- G_s \delta \mathrm{d}(\chi_+ G\alpha))] \tag{5-7}$$

Remark 5.5. We are dealing with a family h_s of perturbations; therefore the above construction should be performed for each s. In particular, for each value of s, one should consider appropriate spacelike Cauchy surfaces. However, $\mathrm{supp}(h_s) \subseteq K$ for each s. Having under control the support of the whole family of perturbations h_s, it is possible to choose spacelike Cauchy surfaces and partitions of unity which do the job for each s. $\Sigma_\pm$, $\Sigma'_\pm$ were chosen exactly in this spirit.

For the proof of Theorem 5.2 we are interested in the functional derivative of $R_s = \mathfrak{Q}(r_s)$. Having this in mind, we compute $\mathrm{d}r_s[\alpha]/\mathrm{d}s\big|_0$ for an arbitrary, but fixed, $[\alpha] \in \mathscr{F}(M)$. This makes sense at least whenever the topology on $\mathscr{F}(M)$ is Hausdorff;[3] see Remark 4.5. Equation (5-7) and the Leibniz rule entail that

$$\frac{\mathrm{d}}{\mathrm{d}s} r_s[\alpha]\big|_0 = \Big[-\frac{\mathrm{d}}{\mathrm{d}s} \delta_s \mathrm{d}\big(\chi_- G \delta \mathrm{d}(\chi_+ G\alpha)\big)\big|_0 - \frac{\mathrm{d}}{\mathrm{d}s} \delta \mathrm{d}\big(\chi_- G_s \delta \mathrm{d}(\chi_+ G\alpha)\big)\big|_0 \Big].$$

Since $\mathrm{supp}(h_s) \cap \mathrm{supp}(\chi_-) \subseteq K \cap M_- = \varnothing$, the argument of the first derivative is constant in s. We deduce that the first contribution vanishes. Decomposing G_s in $G_{s+} - G_{s-}$ and noting that $\chi_- G_{s+} \delta \mathrm{d}(\chi_+ G\alpha)$ and $\chi'_- G_{s-} \delta \mathrm{d}(\chi_+ G\alpha)$ have compact supports, we get

[3] This property ensures uniqueness of limits; therefore $\mathrm{d}r_s[\alpha]/\mathrm{d}s|_0$ is uniquely defined as the limit for $s \to 0$ of $\big(r_s[\alpha] - [\alpha]\big)/s$.

$$\frac{\mathrm{d}}{\mathrm{d}s} r_s[\alpha]\big|_0 = \Big[-\frac{\mathrm{d}}{\mathrm{d}s}\delta\mathrm{d}\big(\chi_- G_{s\,+}\delta\mathrm{d}(\chi_+ G\alpha)\big)\big|_0 + \frac{\mathrm{d}}{\mathrm{d}s}\delta\mathrm{d}\big(\chi_- G_{s\,-}\delta\mathrm{d}(\chi_+ G\alpha)\big)\big|_0\Big]$$
$$= \Big[\frac{\mathrm{d}}{\mathrm{d}s}\delta\mathrm{d}G_{s\,-}\delta\mathrm{d}(\chi_+ G\alpha)\big|_0\Big] = \Big[-\frac{\mathrm{d}}{\mathrm{d}s}\delta\mathrm{d}G_{s\,-}\Box(\chi'_+ G\alpha)\big|_0\Big], \qquad (5\text{-}8)$$

due to the fact that $\delta\mathrm{d}\Omega^1_{\mathrm{c}}(M)$ is identified with 0 in $\mathscr{F}(M)$, $\mathrm{d}G_{s\,-} = G_{s\,-}\mathrm{d}$ on $\Omega^k_{\mathrm{c}}(M)$ and $\Box(\chi_+ G\alpha) = -\Box(\chi'_+ G\alpha)$. On account of the properties of the Green operators (2-6), and exploiting the Leibniz rule, one obtains the identity

$$\frac{\mathrm{d}}{\mathrm{d}s} G_{s\,-}\Box(\chi'_+ G\alpha)\big|_0 + G_-\frac{\mathrm{d}}{\mathrm{d}s}\Box_s(\chi'_+ G\alpha)\big|_0 = \frac{\mathrm{d}}{\mathrm{d}s} G_{s\,-}\Box_s(\chi'_+ G\alpha)\big|_0 = 0,$$

which can be plugged into (5-8). Keeping in mind that $\mathrm{supp}(h_s)$ does not meet $\mathrm{supp}(\chi_+)$, one concludes that $\Box_s(\chi_+ G\alpha) = \Box(\chi_+ G\alpha)$ for each s. Thus, taking into account also that $\Box(\chi_+ G\alpha) = -\Box(\chi_- G\alpha)$ has compact support, we get the equality

$$\frac{\mathrm{d}}{\mathrm{d}s} r_s[\alpha]\big|_0 = \Big[\delta\mathrm{d}G_-\frac{\mathrm{d}}{\mathrm{d}s}\Box_s(\chi'_+ G\alpha)\big|_0\Big] = \Big[\delta\mathrm{d}G_-\frac{\mathrm{d}}{\mathrm{d}s}\delta_s\mathrm{d}G\alpha\big|_0\Big]. \qquad (5\text{-}9)$$

From $\delta_s\delta_s = 0$, the Leibniz rule and $\delta\alpha = 0$ we deduce that

$$\delta\frac{\mathrm{d}}{\mathrm{d}s}\delta_s\mathrm{d}G\alpha\big|_0 = \frac{\mathrm{d}}{\mathrm{d}s}\delta\delta_s\mathrm{d}G\alpha\big|_0 + \frac{\mathrm{d}}{\mathrm{d}s}\delta_s\delta\mathrm{d}G\alpha\big|_0 = \frac{\mathrm{d}}{\mathrm{d}s}\delta_s\delta_s\mathrm{d}G\alpha\big|_0 = 0.$$

Taking into account this information, from (5-9) we come to the conclusion:

$$\frac{\mathrm{d}}{\mathrm{d}s} r_s[\alpha]\big|_0 = \Big[\Box G_-\frac{\mathrm{d}}{\mathrm{d}s}\delta_s\mathrm{d}G\alpha\big|_0\Big] = \Big[\frac{\mathrm{d}}{\mathrm{d}s}\delta_s\mathrm{d}G\alpha\big|_0\Big]. \qquad (5\text{-}10)$$

Explicitly, introducing $F = \mathrm{d}G\alpha$, one reads

$$\Big(\frac{\mathrm{d}}{\mathrm{d}s}\delta_s F\big|_0\Big)_\rho = -\frac{\mathrm{d}}{\mathrm{d}s} g_s^{\mu\nu}\nabla_{s\,\mu}F_{\nu\rho}\big|_0 \qquad (5\text{-}11)$$
$$= \dot{h}_{\mu\nu}\nabla^\mu F^\nu{}_\rho + \tfrac{1}{2}(F^\mu{}_\rho\nabla^\nu\dot{h}_{\mu\nu} - F^\mu{}_\nu\nabla^\nu\dot{h}_{\mu\rho}),$$

where ∇ and ∇_s are the Levi–Civita connections respectively for the unperturbed metric g and the perturbed one g_s. All indices in the result are raised using g and $\dot{h}$ denotes $\mathrm{d}h_s/\mathrm{d}s|_0$. This result follows from the subsequent identities, which are trivial consequences of $g_s^{\mu\nu}$ being the inverse of $g_{s\,\mu\nu}$ and ∇ (∇_s) being the Levi–Civita connection for g (respectively g_s):

$$\frac{\mathrm{d}}{\mathrm{d}s} g_s^{\mu\nu}\big|_0 = -g^{\mu\rho}g^{\nu\sigma}\dot{h}_{\rho\sigma}, \qquad (5\text{-}12)$$

$$\frac{\mathrm{d}}{\mathrm{d}s}(\nabla_{s\,\mu}X^\rho - \nabla_\mu X^\rho)\big|_0 = \tfrac{1}{2}X^\nu g^{\rho\sigma}\nabla_\mu\dot{h}_{\nu\sigma}, \qquad (5\text{-}13)$$

for each vector field X on M.

5C. ***Proof of Theorem 5.2.*** According to the hypotheses, let us consider a globally hyperbolic spacetime M and a quasifree Hadamard state ω on $\mathscr{A}(M)$ with associated GNS triple $(D_\omega, \pi_\omega, \Omega_\omega)$. Fixing $[\alpha] \in \mathscr{F}(M)$, $\theta \in D_\omega$, K compact in M, a 1-parameter family $s \in (-1, 1) \mapsto h_s \in \mathsf{hp}(M)$ supported inside K and taking into account Section 5A and Section 5B, the claim of the theorem boils down to the identity below:

$$\Big\langle \theta, \phi_\omega\Big(\frac{\mathrm{d}}{\mathrm{d}s} r_s[\alpha]\big|_0\Big)\theta\Big\rangle = \frac{\mathrm{i}}{2}\int_M \big\langle \theta, \big[\hat{T}^{\mu\nu}, \phi_\omega([\alpha])\big]\theta\big\rangle \dot{h}_{\mu\nu}\,\mathrm{vol}, \tag{5-14}$$

where $\dot{h}$ stands for $\mathrm{d}h_s/\mathrm{d}s|_0$. We rewrite the left side using (5-4) for $n = 1$, together with eqs. (5-10) and (5-11), and introducing the notation $A = G\alpha$, $F = \mathrm{d}A$ and $\tilde{A} = \langle \theta, \hat{A}\theta\rangle$. For the right side we consider (5-6), keeping in mind that $\dot{h}_{\mu\nu} = \dot{h}_{\nu\mu}$ and defining $\tilde{F} = \mathrm{d}\tilde{A}$. This turns (5-14) into the following identity:

$$\begin{aligned} L &\doteq \int_M \tilde{A}^\rho \big[\dot{h}_{\mu\nu}\nabla^\mu F^\nu{}_\rho \tfrac{1}{2}(F^\mu{}_\rho \nabla^\nu \dot{h}_{\mu\nu} - F^\mu{}_\nu \nabla^\nu \dot{h}_{\mu\rho})\big]\,\mathrm{vol} \\ &= -\tfrac{1}{2}\int_M \dot{h}_{\mu\nu}(2F^{\mu\rho}\tilde{F}^\nu{}_\rho - \tfrac{1}{2}g^{\mu\nu}F^{\rho\sigma}\tilde{F}_{\rho\sigma})\,\mathrm{vol} \doteq R\,. \end{aligned} \tag{5-15}$$

The proof will be complete as soon as one manages to check this identity.

We start by considering the right side. Here we integrate by parts all covariant derivatives acting on $\tilde{A}$. Note that several terms arising from partial integration vanish on account of $\delta F = \delta \mathrm{d} G\alpha = 0$. On account of the symmetry of $\dot{h}$ and the antisymmetry of F, the result is

$$\begin{aligned} R = \int_M \dot{h}_{\mu\nu}\tilde{A}_\rho \nabla^\nu F^{\mu\rho}\,\mathrm{vol} + \int_M F^{\mu\rho}(\tilde{A}_\rho \nabla^\nu \dot{h}_{\mu\nu} - \tilde{A}^\nu \nabla_\rho \dot{h}_{\mu\nu})\,\mathrm{vol} \\ + \frac{1}{2}\int_M g^{\mu\nu}F^{\rho\sigma}\tilde{A}_\rho \nabla_\sigma \dot{h}_{\mu\nu}\,\mathrm{vol}\,. \end{aligned}$$

Comparing R with the left side of (5-15), one reads

$$\begin{aligned} R &= L + \frac{1}{2}\int_M F^{\mu\rho}(\tilde{A}_\rho \nabla^\nu \dot{h}_{\mu\nu} - \tilde{A}^\nu \nabla_\rho \dot{h}_{\mu\nu})\,\mathrm{vol} + \frac{1}{2}\int_M g^{\mu\nu}F^{\rho\sigma}\tilde{A}_\rho \nabla_\sigma \dot{h}_{\mu\nu}\,\mathrm{vol} \\ &= L + \frac{1}{2}\int_M F^{\rho\sigma}\tilde{A}_\rho(g^{\mu\nu}\nabla_\sigma \dot{h}_{\mu\nu} - \nabla^\nu \dot{h}_{\sigma\nu})\,\mathrm{vol} - \frac{1}{2}\int_M \tilde{A}^\nu F^{\mu\rho}\nabla_\rho \dot{h}_{\mu\nu}\,\mathrm{vol} \\ &\doteq L + \frac{1}{2}\int_M F^{\rho\sigma}\tilde{A}_\rho X_\sigma\,\mathrm{vol} - \frac{1}{2}\int_M \tilde{A}^\nu Y_\nu\,\mathrm{vol}, \end{aligned}$$

X and Y being defined by

$$X_\sigma = g^{\mu\nu}\nabla_\sigma \dot{h}_{\mu\nu} - \nabla^\nu \dot{h}_{\sigma\nu}\,, \quad Y_\nu = F^{\mu\rho}\nabla_\rho \dot{h}_{\mu\nu}\,.$$

The rest of the proof is devoted to showing that both X and Y vanish everywhere on M. (5-13) entails that

$$\frac{\mathrm{d}}{\mathrm{d}s}\Gamma^{\rho}_{s\,\mu\nu}\big|_0 = \tfrac{1}{2}g^{\rho\sigma}\nabla_\mu \dot{h}_{\nu\sigma}\,,$$

Γ_s being the Christoffel symbols for the connection ∇_s. As a consequence, $\nabla_\mu \dot{h}_{\nu\sigma}$ is symmetric upon the interchange of μ and ν. Taking into account that F is antisymmetric, we get $X = 0$ and $Y = 0$, thus concluding the proof.

Acknowledgements

The author wishes to thank his supervisor Claudio Dappiaggi for the enlightening discussions and suggestions. This research was supported by a Ph.D. scholarship from the University of Pavia.

References

[Bär 2013] C. Bär, "Green-hyperbolic operators on globally hyperbolic spacetimes", preprint, 2013. arXiv 1310.0738

[Bär and Becker 2009] C. Bär and C. Becker, "C^*-algebras", pp. 1–37 in *Quantum field theory on curved spacetimes*, Lecture Notes in Phys. **786**, Springer, Berlin, 2009.

[Bär and Ginoux 2012a] C. Bär and N. Ginoux, "Classical and quantum fields on Lorentzian manifolds", pp. 359–400 in *Global differential geometry*, edited by C. Bär et al., Springer Proc. in Math. **17**, Springer, Berlin, 2012.

[Bär and Ginoux 2012b] C. Bär and N. Ginoux, "CCR- versus CAR-quantization on curved spacetimes", pp. 183–206 in *Quantum field theory and gravity*, edited by F. Finster et al., Birkhäuser, Basel, 2012.

[Bär et al. 2007] C. Bär, N. Ginoux, and F. Pfäffle, *Wave equations on Lorentzian manifolds and quantization*, European Mathematical Society (EMS), Zürich, 2007.

[Beem et al. 1996] J. K. Beem, P. E. Ehrlich, and K. L. Easley, *Global Lorentzian geometry*, 2nd ed., Pure Appl. Math. **202**, Marcel Dekker, New York, 1996.

[Benini 2014] M. Benini, "Optimal space of linear classical observables for Maxwell k-forms via spacelike and timelike compact de Rham cohomologies", preprint, 2014. arXiv 1401.7563

[Benini et al. 2013] M. Benini, C. Dappiaggi, and T.-P. Hack, "Quantum field theory on curved backgrounds—a primer", *Internat. J. Modern Phys. A* **28**:17 (2013), 1330023, 49.

[Benini et al. 2014a] M. Benini, C. Dappiaggi, T.-P. Hack, and A. Schenkel, "A C*-Algebra for Quantized Principal U(1)-Connections on Globally Hyperbolic Lorentzian Manifolds", *Communications in Mathematical Physics* (2014), 1–28.

[Benini et al. 2014b] M. Benini, C. Dappiaggi, and A. Schenkel, "Quantized abelian principal connections on Lorentzian manifolds", *Comm. Math. Phys.* **330**:1 (2014), 123–152.

[Benini et al. 2014c] M. Benini, C. Dappiaggi, and A. Schenkel, "Quantum field theory on affine bundles", *Ann. Henri Poincaré* **15**:1 (2014), 171–211.

[Bernal and Sánchez 2005] A. N. Bernal and M. Sánchez, "Smoothness of time functions and the metric splitting of globally hyperbolic spacetimes", *Comm. Math. Phys.* **257**:1 (2005), 43–50.

[Borchers 1962] H.-J. Borchers, "On structure of the algebra of field operators", *Nuovo Cimento* (10) **24** (1962), 214–236.

[Bott and Tu 1982] R. Bott and L. W. Tu, *Differential forms in algebraic topology*, Graduate Texts in Mathematics **82**, Springer, New York, 1982.

[Bratteli and Robinson 1987] O. Bratteli and D. W. Robinson, *Operator algebras and quantum statistical mechanics, 1: C*- and W*-algebras, symmetry groups, decomposition of states*, 2nd ed., Springer, New York, 1987.

[Brunetti et al. 2003] R. Brunetti, K. Fredenhagen, and R. Verch, "The generally covariant locality principle—a new paradigm for local quantum field theory", *Comm. Math. Phys.* **237**:1-2 (2003), 31–68.

[Brunetti et al. 2012] R. Brunetti, K. Fredenhagen, and P. L. Ribeiro, "Algebraic structure of classical field theory, I: Kinematics and linearized dynamics for real scalar fields", preprint, 2012. arXiv 1209.2148

[Dappiaggi 2011] C. Dappiaggi, "Remarks on the Reeh–Schlieder property for higher spin free fields on curved spacetimes", *Rev. Math. Phys.* **23**:10 (2011), 1035–1062.

[Dappiaggi and Lang 2012] C. Dappiaggi and B. Lang, "Quantization of Maxwell's equations on curved backgrounds and general local covariance", *Lett. Math. Phys.* **101**:3 (2012), 265–287.

[Dappiaggi and Siemssen 2013] C. Dappiaggi and D. Siemssen, "Hadamard states for the vector potential on asymptotically flat spacetimes", *Rev. Math. Phys.* **25**:1 (2013), 1350002, 31.

[Dimock 1980] J. Dimock, "Algebras of local observables on a manifold", *Comm. Math. Phys.* **77**:3 (1980), 219–228.

[Dimock 1992] J. Dimock, "Quantized electromagnetic field on a manifold", *Rev. Math. Phys.* **4**:2 (1992), 223–233.

[Fewster and Lang 2014] C. J. Fewster and B. Lang, "Dynamical locality of the free Maxwell field", preprint, 2014. arXiv 1403.7083

[Fewster and Pfenning 2003] C. J. Fewster and M. J. Pfenning, "A quantum weak energy inequality for spin-one fields in curved space-time", *J. Math. Phys.* **44**:10 (2003), 4480–4513.

[Fewster and Verch 2012] C. J. Fewster and R. Verch, "Dynamical locality and covariance: what makes a physical theory the same in all spacetimes?", *Ann. Henri Poincaré* **13**:7 (2012), 1613–1674.

[Finster and Strohmaier 2013] F. Finster and A. Strohmaier, "Gupta–Bleuler quantization of the Maxwell field in globally hyperbolic space-times", preprint, 2013. arXiv 1307.1632

[Fulling et al. 1978] S. A. Fulling, M. Sweeny, and R. M. Wald, "Singularity structure of the two-point function quantum field theory in curved spacetime", *Comm. Math. Phys.* **63**:3 (1978), 257–264.

[Fulling et al. 1981] S. A. Fulling, F. J. Narcowich, and R. M. Wald, "Singularity structure of the two-point function in quantum field theory in curved spacetime, II", *Ann. Physics* **136**:2 (1981), 243–272.

[Haag and Kastler 1964] R. Haag and D. Kastler, "An algebraic approach to quantum field theory", *J. Mathematical Phys.* **5** (1964), 848–861.

[Hack and Schenkel 2013] T.-P. Hack and A. Schenkel, "Linear bosonic and fermionic quantum gauge theories on curved spacetimes", *Gen. Relativity Gravitation* **45**:5 (2013), 877–910.

[Hörmander 2003] L. Hörmander, *The analysis of linear partial differential operators, I: Distribution theory and Fourier analysis*, 2nd ed., Springer, Berlin, 2003.

[Khavkine 2014a] I. Khavkine, "Cohomology with causally restricted supports", preprint, 2014. arXiv 1404.1932

[Khavkine 2014b] I. Khavkine, "Covariant phase space, constraints, gauge and the Peierls formula", *Internat. J. Modern Phys. A* **29**:5 (2014), #1430009.

[MacLane 1971] S. MacLane, *Categories for the working mathematician*, Graduate Texts in Mathematics **5**, Springer, New York, 1971.

[Manuceau et al. 1973] J. Manuceau, M. Sirugue, D. Testard, and A. Verbeure, "The smallest C^*-algebra for canonical commutations relations", *Comm. Math. Phys.* **32** (1973), 231–243.

[Pfenning 2009] M. J. Pfenning, "Quantization of the Maxwell field in curved spacetimes of arbitrary dimension", *Classical Quantum Gravity* **26**:13 (2009), 135017, 20.

[Radzikowski 1996a] M. J. Radzikowski, "Micro-local approach to the Hadamard condition in quantum field theory on curved space-time", *Comm. Math. Phys.* **179**:3 (1996), 529–553.

[Radzikowski 1996b] M. J. Radzikowski, "A local-to-global singularity theorem for quantum field theory on curved space-time", *Comm. Math. Phys.* **180**:1 (1996), 1–22.

[Sanders 2010] K. Sanders, "The locally covariant Dirac field", *Rev. Math. Phys.* **22**:4 (2010), 381–430.

[Sanders 2013] K. Sanders, "A note on spacelike and timelike compactness", *Classical Quantum Gravity* **30**:11 (2013), 115014, 10.

[Sanders et al. 2014] K. Sanders, C. Dappiaggi, and T.-P. Hack, "Electromagnetism, local covariance, the Aharonov-Bohm effect and Gauss' law", *Comm. Math. Phys.* **328**:2 (2014), 625–667.

[Uhlmann 1962] A. Uhlmann, "Über die Definition der Quantenfelder nach Wightman und Haag", *Wiss. Z. Karl-Marx-Univ. Leipzig Math.-Nat. Reihe* **11** (1962), 213–217.

[Wald 1994] R. M. Wald, *Quantum field theory in curved spacetime and black hole thermodynamics*, University of Chicago Press, 1994.

[Waldmann 2012] S. Waldmann, "Geometric wave equations", preprint, 2012. arXiv 1208.4706

Received 17 Feb 2014. Accepted 21 Apr 2014.

MARCO BENINI: marco.benini@pv.infn.it
Dipartimento di Fisica, Università di Pavia & INFN, Sezione di Pavia, Via Bassi 6, I-27100 Pavia, Italy
marco.benini@pv.infn.it


Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the submission page.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in MEMOCS are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and a Mathematics Subject Classification or a Physics and Astronomy Classification Scheme code for the article, and, for each author, postal address, affiliation (if appropriate), and email address if available. A home-page URL is optional.

Format. Authors are encouraged to use LaTeX and the standard amsart class, but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should normally be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages — Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc. — allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with as many details as you can about how your graphics were generated.

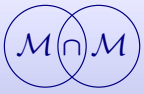
Bundle your figure files into a single archive (using zip, tar, rar or other format of your choice) and upload on the link you been provided at acceptance time. Each figure should be captioned and numbered so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript with all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables.

White Space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

Mathematics and Mechanics of Complex Systems
vol. 3 no. 2 2015

MEMOCS is a journal of the International Research Center for the Mathematics and Mechanics of Complex Systems at the Università dell'Aquila, Italy.




2326-7186(2015)3:2;1-D