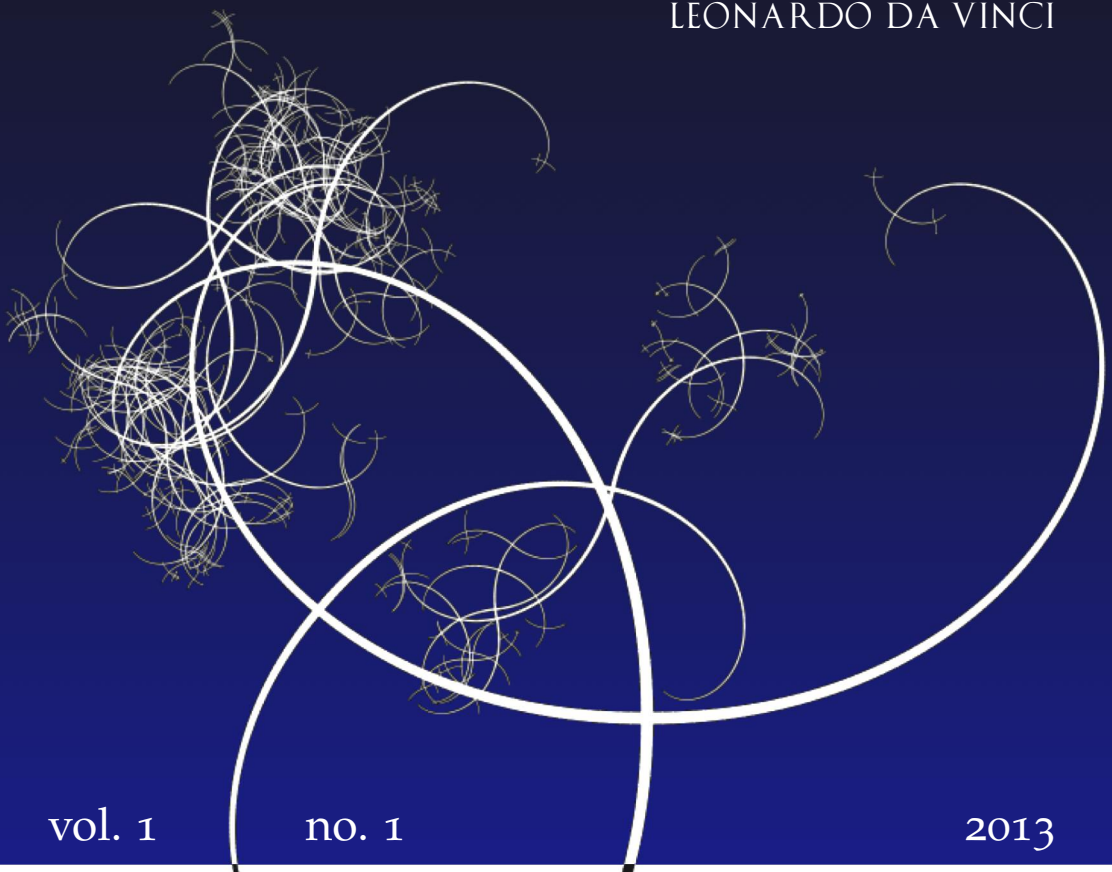
NISSUNA UMANA INVESTIGAZIONE SI PUO DIMANDARE
VERA SCIENZIA S'ESSA NON PASSA PER LE
MATEMATICHE DIMOSTRAZIONI
LEONARDO DA VINCI



vol. 1 no. 1 2013

Mathematics and Mechanics *of* Complex Systems

msp

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS

msp.org/memocs

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/

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 1, No. 1, 2013
dx.doi.org/10.2140/memocs.2013.1.1

DISLOCATIONS, IMPERFECT INTERFACES AND INTERFACE CRACKS IN ANISOTROPIC ELASTICITY FOR QUASICRYSTALS

XU WANG AND PETER SCHIAVONE

We derive the detailed structures of the 6×6 matrices $\boldsymbol{N}_i$ and $\boldsymbol{N}_i^{(-1)}$ $(i = 1, 2, 3)$ in the Stroh formalism of anisotropic elasticity for quasicrystals. All six matrices are expressed explicitly in terms of the sixty-six reduced elastic compliances. The Green's functions for bi-quasicrystals are also obtained. Next, we derive compliant and stiff interface models in anisotropic quasicrystalline bimaterials. It is observed that the phonon normal traction is always continuous across the stiff interface. Finally we present the asymptotic fields associated with a traction-free, semi-infinite interface crack in anisotropic quasicrystalline bimaterials and solve the collinear interface crack problem. The interface crack-tip field consists of three two-dimensional oscillatory singularities which are evaluated via the introduction of three complex stress intensity factors.

1. Introduction

The theory of anisotropic elasticity for crystals has been actively investigated for more than half a century; see, for example, [Lekhnitskii 1950; Eshelby et al. 1953; Stroh 1958; Willis 1964; Willis 1970; Willis 1971; Clements 1971; Barnett and Lothe 1973; Suo 1990; Suo et al. 1992; Gao et al. 1992; Ting 1986; Ting 1988; Ting 1996; Ru 2001; Cheng and Reddy 2002; Ting and Schiavone 2010; Wang and Pan 2010]. As pointed out in [Ting 1996], the Stroh formalism [1958], which is based on Stroh eigenvalues and eigenvectors, allows for an elegant and powerful analysis of two-dimensional deformations of anisotropic crystalline solids where as many as fifteen elastic constants are involved. The beauty of the Stroh formalism has indeed been observed by various researchers; see, for example, [Suo 1990; Ru 2001; Cheng and Reddy 2002; Lazar and Kirchner 2005; Wang and Pan 2010].

Quasicrystalline structures were first reported in [Shechtman et al. 1984]. The generalized anisotropic elasticity for quasicrystals developed in [Hu et al. 2000] requires that anisotropic quasicrystals have as many as one hundred and twenty elastic constants. Even for the study of two-dimensional deformations of quasicrystals,

Communicated by David Steigmann.
MSC2010: primary 74B05, 74E10; secondary 74E15.
Keywords: quasicrystal, anisotropic elasticity, Stroh formalism, dislocation, interface crack.

sixty-six pertinent elastic constants remain. Ding et al. [1995] extended the original six-dimensional Stroh formalism for crystals by developing a twelve-dimensional version in order to study dislocation problems in quasicrystals.

Outline of paper. In Section 2 we present the Stroh formalism for two-dimensional deformations in anisotropic quasicrystals. In Section 3, we derive the detailed structures of $\boldsymbol{N}_i$ and $\boldsymbol{N}_i^{(-1)}$ $(i = 1, 2, 3)$ appearing in the Stroh formalism for quasicrystals through the introduction of sixty-six reduced elastic compliances $S_{ij} = S_{ji}$ $(i, j = 1 \ldots 15$ and $i, j \neq 3, 9, 10, 14)$. In Section 4, the Green's functions for a line force and dislocation located

(1) in a homogeneous quasicrystal,

(2) along the interface of bi-quasicrystals, and

(3) in one of two bonded quasicrystalline half-planes

are investigated. Based on the structures obtained for $\boldsymbol{N}_i$ $(i = 1, 2, 3)$, we then develop in Section 5 two imperfect interface models in anisotropic quasicrystalline bimaterials. Finally, in Section 6, we derive the near-tip field of an interface crack in anisotropic bi-quasicrystals consisting of three two-dimensional coupled oscillatory singularities. Three complex intensity factors K_1, K_2 and K_3 are introduced to quantify the near-tip field. Also studied in Section 6 is the collinear interface crack problems in anisotropic bi-quasicrystals following the decoupling method proposed in [Suo 1990] and [Suo et al. 1992].

2. The Stroh formalism

In a fixed rectangular coordinate system x_i $(i = 1, 2, 3)$, let u_i and w_i be the phonon and phason displacements, $\sigma_{ij}(\sigma_{ij} = \sigma_{ji})$ and $H_{ij}(H_{ij} \neq H_{ji})$ be the phonon and phason stresses in an anisotropic quasicrystalline material. The stress-strain law and the equations of equilibrium are [Hu et al. 2000]:

$$\begin{aligned} \sigma_{ij} &= C_{ijkl}u_{k,l} + R_{ijkl}w_{k,l}, \quad \sigma_{ij,j} = 0, \\ H_{ij} &= R_{klij}u_{k,l} + K_{ijkl}w_{k,l}, \quad H_{ij,j} = 0, \end{aligned} \tag{1}$$

where the comma denotes differentiation, C_{ijkl} are the elastic constants in the phonon field, K_{ijkl}, are the elastic constants in the phason field and R_{ijkl} are the phonon-phason coupling constants. In addition C_{ijkl}, R_{ijkl} and K_{ijkl} possess the following symmetry:

$$C_{ijkl} = C_{jikl} = C_{klij} = C_{ijlk}, \quad R_{ijkl} = R_{jikl}, \quad K_{ijkl} = K_{klij}. \tag{2}$$

For two-dimensional deformations in which u_i and w_i depend only on x_1 and x_2, the general solutions can be expressed as

$$\begin{aligned} \boldsymbol{u} &= [\, u_1 \;\; u_2 \;\; u_3 \;\; w_1 \;\; w_2 \;\; w_3 \,]^T = \boldsymbol{A}\boldsymbol{f}(z) + \overline{\boldsymbol{A}}\,\overline{\boldsymbol{f}(z)}, \\ \boldsymbol{\Phi} &= [\, \Phi_1 \;\; \Phi_2 \;\; \Phi_3 \;\; \Psi_1 \;\; \Psi_2 \;\; \Psi_3 \,]^T = \boldsymbol{B}\boldsymbol{f}(z) + \overline{\boldsymbol{B}}\,\overline{\boldsymbol{f}(z)}, \end{aligned} \tag{3}$$

where

$$\begin{aligned} &\boldsymbol{A} = [\boldsymbol{a}_1 \;\; \boldsymbol{a}_2 \;\; \boldsymbol{a}_3 \;\; \boldsymbol{a}_4 \;\; \boldsymbol{a}_5 \;\; \boldsymbol{a}_6], \quad \boldsymbol{B} = [\boldsymbol{b}_1 \;\; \boldsymbol{b}_2 \;\; \boldsymbol{b}_3 \;\; \boldsymbol{b}_4 \;\; \boldsymbol{b}_5 \;\; \boldsymbol{b}_6], \\ &\boldsymbol{f}(z) = [f_1(z_1) \;\; f_2(z_2) \;\; f_3(z_3) \;\; f_4(z_4) \;\; f_5(z_5) \;\; f_6(z_6)]^T, \\ &z_i = x_1 + p_i x_2, \quad \operatorname{Im}\{p_i\} > 0 \quad (i = 1, \ldots, 6), \end{aligned} \tag{4}$$

with

$$\begin{bmatrix} \boldsymbol{N}_1 & \boldsymbol{N}_2 \\ \boldsymbol{N}_3 & \boldsymbol{N}_1^T \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_i \\ \boldsymbol{b}_i \end{bmatrix} = p_i \begin{bmatrix} \boldsymbol{a}_i \\ \boldsymbol{b}_i \end{bmatrix} \quad (i = 1, \ldots, 6), \tag{5}$$

$$\boldsymbol{N}_1 = -\boldsymbol{T}^{-1}\boldsymbol{R}^T, \quad \boldsymbol{N}_2 = \boldsymbol{T}^{-1}, \quad \boldsymbol{N}_3 = \boldsymbol{R}\boldsymbol{T}^{-1}\boldsymbol{R}^T - \boldsymbol{Q}, \tag{6}$$

$$\boldsymbol{Q} = \begin{bmatrix} C_{11} & C_{16} & C_{15} & R_{11} & R_{19} & R_{15} \\ C_{16} & C_{66} & C_{56} & R_{61} & R_{69} & R_{65} \\ C_{15} & C_{56} & C_{55} & R_{51} & R_{59} & R_{55} \\ R_{11} & R_{61} & R_{51} & K_{11} & K_{19} & K_{15} \\ R_{19} & R_{69} & R_{59} & K_{19} & K_{99} & K_{59} \\ R_{15} & R_{65} & R_{55} & K_{15} & K_{59} & K_{55} \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} C_{16} & C_{12} & C_{14} & R_{16} & R_{12} & R_{17} \\ C_{66} & C_{26} & C_{46} & R_{66} & R_{62} & R_{67} \\ C_{56} & C_{25} & C_{45} & R_{56} & R_{52} & R_{57} \\ R_{61} & R_{21} & R_{41} & K_{16} & K_{12} & K_{17} \\ R_{69} & R_{29} & R_{49} & K_{69} & K_{29} & K_{79} \\ R_{65} & R_{25} & R_{45} & K_{56} & K_{25} & K_{57} \end{bmatrix}, \tag{7}$$

$$\boldsymbol{T} = \begin{bmatrix} C_{66} & C_{26} & C_{46} & R_{66} & R_{62} & R_{67} \\ C_{26} & C_{22} & C_{24} & R_{26} & R_{22} & R_{27} \\ C_{46} & C_{24} & C_{44} & R_{46} & R_{42} & R_{47} \\ R_{66} & R_{26} & R_{46} & K_{66} & K_{26} & K_{67} \\ R_{62} & R_{22} & R_{42} & K_{26} & K_{22} & K_{27} \\ R_{67} & R_{27} & R_{47} & K_{67} & K_{27} & K_{77} \end{bmatrix}.$$

The matrices $\boldsymbol{Q}$ and $\boldsymbol{T}$ are symmetric and positive definite.

In (7), we have adopted the contracted notation

$$11 \leftrightarrow 1, \;\; 22 \leftrightarrow 2, \;\; 33 \leftrightarrow 3, \;\; 23 \leftrightarrow 4, \;\; 31 \leftrightarrow 5, \;\; 12 \leftrightarrow 6, \;\; 32 \leftrightarrow 7, \;\; 13 \leftrightarrow 8, \;\; 21 \leftrightarrow 9.$$

In addition the stress function vector $\boldsymbol{\Phi}$ is defined, in terms of the phonon and phason stresses, by

$$\sigma_{i1} = -\Phi_{i,2}, \quad \sigma_{i2} = \Phi_{i,1}, \quad H_{i1} = -\Psi_{i,2}, \quad H_{i2} = \Psi_{i,1} \quad (i = 1, 2, 3). \tag{8}$$

The 6×6 matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfy the normalized orthogonal relationship

$$\begin{bmatrix} \boldsymbol{B}^T & \boldsymbol{A}^T \\ \overline{\boldsymbol{B}}^T & \overline{\boldsymbol{A}}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \overline{\boldsymbol{A}} \\ \boldsymbol{B} & \overline{\boldsymbol{B}} \end{bmatrix} = \boldsymbol{I}. \tag{9}$$

Therefore three real Barnett–Lothe tensors $\boldsymbol{S}$, $\boldsymbol{H}$ and $\boldsymbol{L}$ can be introduced:

$$\boldsymbol{S} = \mathrm{i}(2\boldsymbol{A}\boldsymbol{B}^T - \boldsymbol{I}), \quad \boldsymbol{H} = 2\mathrm{i}\boldsymbol{A}\boldsymbol{A}^T, \quad \boldsymbol{L} = -2\mathrm{i}\boldsymbol{B}\boldsymbol{B}^T. \tag{10}$$

Here $\boldsymbol{H}$ and $\boldsymbol{L}$ are positive definite matrices. It can also be easily checked that

$$\begin{bmatrix} \boldsymbol{N}_1^{(-1)} & \boldsymbol{N}_2^{(-1)} \\ \boldsymbol{N}_3^{(-1)} & \boldsymbol{N}_1^{(-1)^T} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_i \\ \boldsymbol{b}_i \end{bmatrix} = \frac{1}{p_i} \begin{bmatrix} \boldsymbol{a}_i \\ \boldsymbol{b}_i \end{bmatrix} \quad (i = 1, \ldots, 6), \tag{11}$$

where

$$\boldsymbol{N}_1^{(-1)} = -\boldsymbol{Q}^{-1}\boldsymbol{R}, \quad \boldsymbol{N}_2^{(-1)} = -\boldsymbol{Q}^{-1}, \quad \boldsymbol{N}_3^{(-1)} = \boldsymbol{T} - \boldsymbol{R}^T\boldsymbol{Q}^{-1}\boldsymbol{R}. \tag{12}$$

In addition, both the 6×6 matrix $\boldsymbol{M}$ and its inverse $\boldsymbol{M}^{-1}$, defined by

$$\boldsymbol{M} = -\mathrm{i}\boldsymbol{B}\boldsymbol{A}^{-1} = \boldsymbol{H}^{-1} + \mathrm{i}\boldsymbol{H}^{-1}\boldsymbol{S}, \quad \boldsymbol{M}^{-1} = \mathrm{i}\boldsymbol{A}\boldsymbol{B}^{-1} = \boldsymbol{L}^{-1} - \mathrm{i}\boldsymbol{S}\boldsymbol{L}^{-1}, \tag{13}$$

are positive definite Hermitian matrices. In this work we make use of a positive definite Hermitian matrix $\boldsymbol{N}$ involving bimaterial elastic constants and defined as

$$\boldsymbol{N} = \boldsymbol{M}_1^{-1} + \overline{\boldsymbol{M}}_2^{-1} = \boldsymbol{L}_1^{-1} + \boldsymbol{L}_2^{-1} + \mathrm{i}(\boldsymbol{S}_2\boldsymbol{L}_2^{-1} - \boldsymbol{S}_1\boldsymbol{L}_1^{-1}). \tag{14}$$

It is clear from (7) that for two-dimensional deformations of quasicrystalline materials, there are in total sixty-six elastic constants: fifteen in the phonon field, twenty-one in the phason field and thirty phonon-phason coupling constants. In the next section we present the detailed structures of $\boldsymbol{N}_i$ and $\boldsymbol{N}_i^{(-1)}$ $(i = 1, 2, 3)$.

3. The structures of $\boldsymbol{N}_i$ and $\boldsymbol{N}_i^{(-1)}$

Consider first the structure of $\boldsymbol{N}_i$ $(i = 1, 2, 3)$. Since the second column of $\boldsymbol{Q}$ is identical to the first column of $\boldsymbol{R}$, and the second row of $\boldsymbol{R}$ is identical to the first row of $\boldsymbol{T}$, we have

$$\boldsymbol{N}_1 = \begin{bmatrix} * & -1 & * & * & * & * \\ * & 0 & * & * & * & * \\ * & 0 & * & * & * & * \\ * & 0 & * & * & * & * \\ * & 0 & * & * & * & * \\ * & 0 & * & * & * & * \end{bmatrix}, \quad \boldsymbol{N}_3 = \begin{bmatrix} * & 0 & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 \\ * & 0 & * & * & * & * \\ * & 0 & * & * & * & * \\ * & 0 & * & * & * & * \\ * & 0 & * & * & * & * \end{bmatrix}, \tag{15}$$

where $*$ denotes a possibly nonzero element.

Next, we introduce the reduced elastic compliances $S_{ij} = S_{ji}$ $(i, j = 1 \ldots 15$ and $i, j \neq 3, 9, 10, 14)$ such that

$$\begin{bmatrix} C_{11} & C_{12} & C_{14} & C_{15} & C_{16} & R_{11} & R_{12} & R_{15} & R_{16} & R_{17} & R_{19} \\ C_{12} & C_{22} & C_{24} & C_{25} & C_{26} & R_{21} & R_{22} & R_{25} & R_{26} & R_{27} & R_{29} \\ C_{14} & C_{24} & C_{44} & C_{45} & C_{46} & R_{41} & R_{42} & R_{45} & R_{46} & R_{47} & R_{49} \\ C_{15} & C_{25} & C_{45} & C_{55} & C_{56} & R_{51} & R_{52} & R_{55} & R_{56} & R_{57} & R_{59} \\ C_{16} & C_{26} & C_{46} & C_{56} & C_{66} & R_{61} & R_{62} & R_{65} & R_{66} & R_{67} & R_{69} \\ R_{11} & R_{21} & R_{41} & R_{51} & R_{61} & K_{11} & K_{12} & K_{15} & K_{16} & K_{17} & K_{19} \\ R_{12} & R_{22} & R_{42} & R_{52} & R_{62} & K_{12} & K_{22} & K_{25} & K_{26} & K_{27} & K_{29} \\ R_{15} & R_{25} & R_{45} & R_{55} & R_{63} & K_{15} & K_{25} & K_{55} & K_{56} & K_{57} & K_{59} \\ R_{16} & R_{26} & R_{46} & R_{56} & R_{66} & K_{16} & K_{26} & K_{56} & K_{66} & K_{67} & K_{69} \\ R_{17} & R_{27} & R_{47} & R_{57} & R_{67} & K_{17} & K_{27} & K_{57} & K_{67} & K_{77} & K_{79} \\ R_{19} & R_{29} & R_{49} & R_{59} & R_{69} & K_{19} & K_{29} & K_{59} & K_{69} & K_{79} & K_{99} \end{bmatrix}$$

$$\times \begin{bmatrix} S_{11} & S_{12} & S_{14} & S_{15} & S_{16} & S_{17} & S_{18} & S_{111} & S_{112} & S_{113} & S_{115} \\ S_{12} & S_{22} & S_{24} & S_{25} & S_{26} & S_{27} & S_{28} & S_{211} & S_{212} & S_{213} & S_{215} \\ S_{14} & S_{24} & S_{44} & S_{45} & S_{46} & S_{47} & S_{48} & S_{411} & S_{412} & S_{413} & S_{415} \\ S_{15} & S_{25} & S_{45} & S_{55} & S_{56} & S_{57} & S_{58} & S_{511} & S_{512} & S_{513} & S_{515} \\ S_{16} & S_{26} & S_{46} & S_{56} & S_{66} & S_{67} & S_{68} & S_{611} & S_{612} & S_{613} & S_{615} \\ S_{17} & S_{27} & S_{47} & S_{57} & S_{67} & S_{77} & S_{78} & S_{711} & S_{712} & S_{713} & S_{715} \\ S_{18} & S_{28} & S_{48} & S_{58} & S_{68} & S_{78} & S_{88} & S_{811} & S_{812} & S_{813} & S_{815} \\ S_{111} & S_{211} & S_{411} & S_{511} & S_{611} & S_{711} & S_{811} & S_{1111} & S_{1112} & S_{1113} & S_{1115} \\ S_{112} & S_{212} & S_{412} & S_{512} & S_{612} & S_{712} & S_{812} & S_{1112} & S_{1212} & S_{1213} & S_{1215} \\ S_{113} & S_{213} & S_{413} & S_{513} & S_{613} & S_{713} & S_{813} & S_{1113} & S_{1213} & S_{1313} & S_{1315} \\ S_{115} & S_{215} & S_{415} & S_{515} & S_{615} & S_{715} & S_{815} & S_{1115} & S_{1215} & S_{1315} & S_{1515} \end{bmatrix} = \boldsymbol{I}. \quad (16)$$

Remark. We adopt the convention that if three digits appear as subscripts of S_{ij}, the first digit is i and the remaining two form j. If four digits appear in the subscripts of S_{ij}, the first two digits are i and the remaining two will form j.

It can be easily deduced from (16) that

$$\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{R} \\ \boldsymbol{R}^T & \boldsymbol{T} \end{bmatrix} \begin{bmatrix} \boldsymbol{q}_2 & \boldsymbol{r}_2 \\ \boldsymbol{r}_2^T & \boldsymbol{t} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} - \boldsymbol{I}_2 & \boldsymbol{I}_{12}^T \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}, \quad (17)$$

where

$$\boldsymbol{q}_2 = \begin{bmatrix} S_{11} & 0 & S_{15} & S_{17} & S_{115} & S_{111} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ S_{15} & 0 & S_{55} & S_{57} & S_{515} & S_{511} \\ S_{17} & 0 & S_{57} & S_{77} & S_{715} & S_{711} \\ S_{115} & 0 & S_{515} & S_{715} & S_{1515} & S_{1115} \\ S_{111} & 0 & S_{511} & S_{711} & S_{1115} & S_{1111} \end{bmatrix},$$

$$\boldsymbol{r}_2 = \begin{bmatrix} S_{16} & S_{12} & S_{14} & S_{112} & S_{18} & S_{113} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ S_{56} & S_{25} & S_{45} & S_{512} & S_{58} & S_{513} \\ S_{67} & S_{27} & S_{47} & S_{712} & S_{78} & S_{713} \\ S_{615} & S_{215} & S_{415} & S_{715} & S_{815} & S_{1315} \\ S_{611} & S_{211} & S_{411} & S_{711} & S_{811} & S_{1113} \end{bmatrix},$$

$$\boldsymbol{t} = \begin{bmatrix} S_{66} & S_{26} & S_{46} & S_{612} & S_{68} & S_{613} \\ S_{26} & S_{22} & S_{24} & S_{212} & S_{28} & S_{213} \\ S_{46} & S_{24} & S_{44} & S_{412} & S_{48} & S_{413} \\ S_{612} & S_{212} & S_{412} & S_{1212} & S_{812} & S_{1213} \\ S_{68} & S_{28} & S_{48} & S_{812} & S_{88} & S_{813} \\ S_{613} & S_{213} & S_{413} & S_{1213} & S_{813} & S_{1313} \end{bmatrix},$$

$$\boldsymbol{I}_{12} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{I}_2 = \operatorname{diag}[0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0].$$

After some algebraic manipulations, we finally arrive at

$$\boldsymbol{N}_3 = -\boldsymbol{q}_2^{-1} = \frac{-1}{\Delta} \begin{bmatrix} \widehat{W}_{11} & 0 & \widehat{W}_{12} & \widehat{W}_{13} & \widehat{W}_{14} & \widehat{W}_{15} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \widehat{W}_{12} & 0 & \widehat{W}_{22} & \widehat{W}_{23} & \widehat{W}_{24} & \widehat{W}_{25} \\ \widehat{W}_{13} & 0 & \widehat{W}_{23} & \widehat{W}_{33} & \widehat{W}_{34} & \widehat{W}_{35} \\ \widehat{W}_{14} & 0 & \widehat{W}_{24} & \widehat{W}_{34} & \widehat{W}_{44} & \widehat{W}_{45} \\ \widehat{W}_{15} & 0 & \widehat{W}_{25} & \widehat{W}_{35} & \widehat{W}_{45} & \widehat{W}_{55} \end{bmatrix}, \tag{18}$$

$$\boldsymbol{N}_1 = \boldsymbol{r}_2^T \boldsymbol{q}_2^{-1} - \boldsymbol{I}_{12} = \begin{bmatrix} r_6 & -1 & s_6 & t_6 & u_6 & v_6 \\ r_2 & 0 & s_2 & t_2 & u_2 & v_2 \\ r_4 & 0 & s_4 & t_4 & u_4 & v_4 \\ r_{12} & 0 & s_{12} & t_{12} & u_{12} & v_{12} \\ r_8 & 0 & s_8 & t_8 & u_8 & v_8 \\ r_{13} & 0 & s_{13} & t_{13} & u_{13} & v_{13} \end{bmatrix}, \tag{19}$$

$$\boldsymbol{N}_2 = \boldsymbol{t} - \boldsymbol{r}_2^T \boldsymbol{q}_2^{-1} \boldsymbol{r}_2 = \begin{bmatrix} \kappa_{66} & \kappa_{26} & \kappa_{46} & \kappa_{612} & \kappa_{68} & \kappa_{613} \\ \kappa_{26} & \kappa_{22} & \kappa_{24} & \kappa_{212} & \kappa_{28} & \kappa_{213} \\ \kappa_{46} & \kappa_{24} & \kappa_{44} & \kappa_{412} & \kappa_{48} & \kappa_{413} \\ \kappa_{612} & \kappa_{212} & \kappa_{412} & \kappa_{1212} & \kappa_{812} & \kappa_{1213} \\ \kappa_{68} & \kappa_{28} & \kappa_{48} & \kappa_{812} & \kappa_{88} & \kappa_{813} \\ \kappa_{613} & \kappa_{213} & \kappa_{413} & \kappa_{1213} & \kappa_{813} & \kappa_{1313} \end{bmatrix}, \tag{20}$$

where $\boldsymbol{q}_2^{-1}$ is the pseudo-inverse of $\boldsymbol{q}_2$, Δ is the determinant of

$$\boldsymbol{W} = \begin{bmatrix} S_{11} & S_{15} & S_{17} & S_{115} & S_{111} \\ S_{15} & S_{55} & S_{57} & S_{515} & S_{511} \\ S_{17} & S_{57} & S_{77} & S_{715} & S_{711} \\ S_{115} & S_{515} & S_{715} & S_{1515} & S_{1115} \\ S_{111} & S_{511} & S_{711} & S_{1115} & S_{1111} \end{bmatrix}, \tag{21}$$

$\widehat{\boldsymbol{W}} = [\widehat{W}_{ij}]$ is the cofactor matrix of $\boldsymbol{W}$, and

$$r_\alpha = \frac{1}{\Delta}\begin{vmatrix} S_{1\alpha} & S_{5\alpha} & S_{7\alpha} & S_{15\alpha} & S_{11\alpha} \\ S_{15} & S_{55} & S_{57} & S_{515} & S_{511} \\ S_{17} & S_{57} & S_{77} & S_{715} & S_{711} \\ S_{115} & S_{515} & S_{715} & S_{1515} & S_{1115} \\ S_{111} & S_{511} & S_{711} & S_{1115} & S_{1111} \end{vmatrix}, \quad s_\alpha = \frac{1}{\Delta}\begin{vmatrix} S_{11} & S_{15} & S_{17} & S_{115} & S_{111} \\ S_{1\alpha} & S_{5\alpha} & S_{7\alpha} & S_{15\alpha} & S_{11\alpha} \\ S_{17} & S_{57} & S_{77} & S_{715} & S_{711} \\ S_{115} & S_{515} & S_{715} & S_{1515} & S_{1115} \\ S_{111} & S_{511} & S_{711} & S_{1115} & S_{1111} \end{vmatrix},$$

$$t_\alpha = \frac{1}{\Delta}\begin{vmatrix} S_{11} & S_{15} & S_{17} & S_{115} & S_{111} \\ S_{15} & S_{55} & S_{57} & S_{515} & S_{511} \\ S_{1\alpha} & S_{5\alpha} & S_{7\alpha} & S_{15\alpha} & S_{11\alpha} \\ S_{115} & S_{515} & S_{715} & S_{1515} & S_{1115} \\ S_{111} & S_{511} & S_{711} & S_{1115} & S_{1111} \end{vmatrix}, \quad u_\alpha = \frac{1}{\Delta}\begin{vmatrix} S_{11} & S_{15} & S_{17} & S_{115} & S_{111} \\ S_{15} & S_{55} & S_{57} & S_{515} & S_{511} \\ S_{17} & S_{57} & S_{77} & S_{715} & S_{711} \\ S_{1\alpha} & S_{5\alpha} & S_{7\alpha} & S_{15\alpha} & S_{11\alpha} \\ S_{111} & S_{511} & S_{711} & S_{1115} & S_{1111} \end{vmatrix},$$

$$v_\alpha = \frac{1}{\Delta}\begin{vmatrix} S_{11} & S_{15} & S_{17} & S_{115} & S_{111} \\ S_{15} & S_{55} & S_{57} & S_{515} & S_{511} \\ S_{17} & S_{57} & S_{77} & S_{715} & S_{711} \\ S_{115} & S_{515} & S_{715} & S_{1515} & S_{1115} \\ S_{1\alpha} & S_{5\alpha} & S_{7\alpha} & S_{15\alpha} & S_{11\alpha} \end{vmatrix} \qquad (\alpha = 6, 2, 4, 12, 8, 13),$$

$$\kappa_{\alpha\beta} = \frac{1}{\Delta}\begin{vmatrix} S_{11} & S_{1\alpha} & S_{15} & S_{17} & S_{115} & S_{111} \\ S_{1\beta} & S_{\alpha\beta} & S_{5\beta} & S_{7\beta} & S_{15\beta} & S_{11\beta} \\ S_{15} & S_{5\alpha} & S_{55} & S_{57} & S_{515} & S_{511} \\ S_{17} & S_{7\alpha} & S_{57} & S_{77} & S_{715} & S_{711} \\ S_{115} & S_{15\alpha} & S_{515} & S_{715} & S_{1515} & S_{1115} \\ S_{111} & S_{11\alpha} & S_{511} & S_{711} & S_{1115} & S_{1111} \end{vmatrix} \qquad (\alpha, \beta = 6, 2, 4, 12, 8, 13).$$

In view of (6), $\boldsymbol{N}_2$ is positive definite, while (18) and the fact that $\boldsymbol{W}$ defined by (21) is positive definite result in $-\boldsymbol{N}_3$ being positive semidefinite.

We next derive the structure of $\boldsymbol{N}_i^{(-1)}$ $(i = 1, 2, 3)$. It is not hard to check that

$$\boldsymbol{N}_1^{(-1)} = \begin{bmatrix} 0 & * & * & * & * & * \\ -1 & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \end{bmatrix}, \quad \boldsymbol{N}_3^{(-1)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & * & * & * & * & * \end{bmatrix}, \tag{22}$$

In addition, we have the identity

$$\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{R} \\ \boldsymbol{R}^T & \boldsymbol{T} \end{bmatrix} \begin{bmatrix} \boldsymbol{q} & \boldsymbol{r}_1 \\ \boldsymbol{r}_1^T & \boldsymbol{t}_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{I}_{12} & \boldsymbol{I} - \boldsymbol{I}_1 \end{bmatrix}, \tag{23}$$

where

$$\boldsymbol{q} = \begin{bmatrix} S_{11} & S_{16} & S_{15} & S_{17} & S_{115} & S_{111} \\ S_{16} & S_{66} & S_{56} & S_{67} & S_{615} & S_{611} \\ S_{15} & S_{56} & S_{55} & S_{57} & S_{515} & S_{511} \\ S_{17} & S_{67} & S_{57} & S_{77} & S_{715} & S_{711} \\ S_{115} & S_{615} & S_{515} & S_{715} & S_{1515} & S_{1115} \\ S_{111} & S_{611} & S_{511} & S_{711} & S_{1115} & S_{1111} \end{bmatrix}, \quad \boldsymbol{r}_1 = \begin{bmatrix} 0 & S_{12} & S_{14} & S_{112} & S_{18} & S_{113} \\ 0 & S_{26} & S_{46} & S_{612} & S_{68} & S_{613} \\ 0 & S_{25} & S_{45} & S_{512} & S_{58} & S_{513} \\ 0 & S_{27} & S_{47} & S_{712} & S_{78} & S_{713} \\ 0 & S_{215} & S_{415} & S_{715} & S_{815} & S_{1315} \\ 0 & S_{211} & S_{411} & S_{711} & S_{811} & S_{1113} \end{bmatrix},$$

$$\boldsymbol{t}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & S_{22} & S_{24} & S_{212} & S_{28} & S_{213} \\ 0 & S_{24} & S_{44} & S_{412} & S_{48} & S_{413} \\ 0 & S_{212} & S_{412} & S_{1212} & S_{812} & S_{1213} \\ 0 & S_{28} & S_{48} & S_{812} & S_{88} & S_{813} \\ 0 & S_{213} & S_{413} & S_{1213} & S_{813} & S_{1313} \end{bmatrix}, \qquad \boldsymbol{I}_1 = \text{diag}[1\ 0\ 0\ 0\ 0\ 0].$$

We can now arrive at

$$\boldsymbol{N}_3^{(-1)} = \boldsymbol{t}_1^{-1} = \frac{1}{\Delta'} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \widehat{W}'_{11} & \widehat{W}'_{12} & \widehat{W}'_{13} & \widehat{W}'_{14} & \widehat{W}'_{15} \\ 0 & \widehat{W}'_{12} & \widehat{W}'_{22} & \widehat{W}'_{23} & \widehat{W}'_{24} & \widehat{W}'_{25} \\ 0 & \widehat{W}'_{13} & \widehat{W}'_{23} & \widehat{W}'_{33} & \widehat{W}'_{34} & \widehat{W}'_{35} \\ 0 & \widehat{W}'_{14} & \widehat{W}'_{24} & \widehat{W}'_{34} & \widehat{W}'_{44} & \widehat{W}'_{45} \\ 0 & \widehat{W}'_{15} & \widehat{W}'_{25} & \widehat{W}'_{35} & \widehat{W}'_{45} & \widehat{W}'_{55} \end{bmatrix}, \tag{24}$$

$$\boldsymbol{N}_1^{(-1)} = \boldsymbol{r}_1 \boldsymbol{t}_1^{-1} - \boldsymbol{I}_{12}^T = \begin{bmatrix} 0 & r'_1 & s'_1 & t'_1 & u'_1 & v'_1 \\ -1 & r'_6 & s'_6 & t'_6 & u'_6 & v'_6 \\ 0 & r'_5 & s'_5 & t'_5 & u'_5 & v'_5 \\ 0 & r'_7 & s'_7 & t'_7 & u'_7 & v'_7 \\ 0 & r'_{15} & s'_{15} & t'_{15} & u'_{15} & v'_{15} \\ 0 & r'_{11} & s'_{11} & t'_{11} & u'_{11} & v'_{11} \end{bmatrix}, \tag{25}$$

$$\boldsymbol{N}_2^{(-1)} = -\boldsymbol{q} + \boldsymbol{r}_1 \boldsymbol{t}_1^{-1} \boldsymbol{r}_1^T = -\begin{bmatrix} \kappa'_{11} & \kappa'_{16} & \kappa'_{15} & \kappa'_{17} & \kappa'_{115} & \kappa'_{111} \\ \kappa'_{16} & \kappa'_{66} & \kappa'_{56} & \kappa'_{67} & \kappa'_{615} & \kappa'_{611} \\ \kappa'_{15} & \kappa'_{56} & \kappa'_{55} & \kappa'_{57} & \kappa'_{515} & \kappa'_{511} \\ \kappa'_{17} & \kappa'_{67} & \kappa'_{57} & \kappa'_{77} & \kappa'_{715} & \kappa'_{711} \\ \kappa'_{115} & \kappa'_{615} & \kappa'_{515} & \kappa'_{715} & \kappa'_{1515} & \kappa'_{1115} \\ \kappa'_{111} & \kappa'_{611} & \kappa'_{511} & \kappa'_{711} & \kappa'_{1115} & \kappa'_{1111} \end{bmatrix}, \tag{26}$$

where Δ' is the determinant of

$$\boldsymbol{W}' = \begin{bmatrix} S_{22} & S_{24} & S_{212} & S_{28} & S_{213} \\ S_{24} & S_{44} & S_{412} & S_{48} & S_{413} \\ S_{212} & S_{412} & S_{1212} & S_{812} & S_{1213} \\ S_{28} & S_{48} & S_{812} & S_{88} & S_{813} \\ S_{213} & S_{413} & S_{1213} & S_{813} & S_{1313} \end{bmatrix}, \tag{27}$$

$\widehat{\boldsymbol{W}}' = [\widehat{W}'_{ij}]$ is the cofactor of $\boldsymbol{W}'$, and

$$r'_\alpha = \frac{1}{\Delta'} \begin{vmatrix} S_{2\alpha} & S_{4\alpha} & S_{12\alpha} & S_{8\alpha} & S_{13\alpha} \\ S_{24} & S_{44} & S_{412} & S_{48} & S_{413} \\ S_{212} & S_{412} & S_{1212} & S_{812} & S_{1213} \\ S_{28} & S_{48} & S_{812} & S_{88} & S_{813} \\ S_{213} & S_{413} & S_{1213} & S_{813} & S_{1313} \end{vmatrix}, \quad s'_\alpha = \frac{1}{\Delta'} \begin{vmatrix} S_{22} & S_{24} & S_{212} & S_{28} & S_{213} \\ S_{2\alpha} & S_{4\alpha} & S_{12\alpha} & S_{8\alpha} & S_{13\alpha} \\ S_{212} & S_{412} & S_{1212} & S_{812} & S_{1213} \\ S_{28} & S_{48} & S_{812} & S_{88} & S_{813} \\ S_{213} & S_{413} & S_{1213} & S_{813} & S_{1313} \end{vmatrix},$$

$$t'_\alpha = \frac{1}{\Delta'} \begin{vmatrix} S_{22} & S_{24} & S_{212} & S_{28} & S_{213} \\ S_{24} & S_{44} & S_{412} & S_{48} & S_{413} \\ S_{2\alpha} & S_{4\alpha} & S_{12\alpha} & S_{8\alpha} & S_{13\alpha} \\ S_{28} & S_{48} & S_{812} & S_{88} & S_{813} \\ S_{213} & S_{413} & S_{1213} & S_{813} & S_{1313} \end{vmatrix}, \quad u'_\alpha = \frac{1}{\Delta'} \begin{vmatrix} S_{22} & S_{24} & S_{212} & S_{28} & S_{213} \\ S_{24} & S_{44} & S_{412} & S_{48} & S_{413} \\ S_{212} & S_{412} & S_{1212} & S_{812} & S_{1213} \\ S_{2\alpha} & S_{4\alpha} & S_{12\alpha} & S_{8\alpha} & S_{13\alpha} \\ S_{213} & S_{413} & S_{1213} & S_{813} & S_{1313} \end{vmatrix},$$

$$v'_\alpha = \frac{1}{\Delta'} \begin{vmatrix} S_{22} & S_{24} & S_{212} & S_{28} & S_{213} \\ S_{24} & S_{44} & S_{412} & S_{48} & S_{413} \\ S_{212} & S_{412} & S_{1212} & S_{812} & S_{1213} \\ S_{28} & S_{48} & S_{812} & S_{88} & S_{813} \\ S_{2\alpha} & S_{4\alpha} & S_{12\alpha} & S_{8\alpha} & S_{13\alpha} \end{vmatrix} \qquad (\alpha = 1, 6, 5, 7, 15, 11),$$

$$\kappa'_{\alpha\beta} = \frac{1}{\Delta'} \begin{vmatrix} S_{\alpha\beta} & S_{2\beta} & S_{4\beta} & S_{12\beta} & S_{8\beta} & S_{13\beta} \\ S_{2\alpha} & S_{22} & S_{24} & S_{212} & S_{28} & S_{213} \\ S_{4\alpha} & S_{24} & S_{44} & S_{412} & S_{48} & S_{413} \\ S_{12\alpha} & S_{212} & S_{412} & S_{1212} & S_{812} & S_{1213} \\ S_{8\alpha} & S_{28} & S_{48} & S_{812} & S_{88} & S_{813} \\ S_{13\alpha} & S_{213} & S_{413} & S_{1213} & S_{813} & S_{1313} \end{vmatrix} \qquad (\alpha, \beta = 1, 6, 5, 7, 15, 11).$$

In view of (12), the matrix $-\boldsymbol{N}_2^{(-1)}$ is positive definite, while, by (24) and the fact that $\boldsymbol{W}'$ defined by (27) is positive definite, we also have that $\boldsymbol{N}_3^{(-1)}$ is positive semidefinite.

4. Line force and line dislocation

4.1. ***Line force and dislocation in a homogeneous quasicrystal.*** Let a line of uniformly distributed force $\boldsymbol{p} = [p_1 \;\; p_2 \;\; p_3 \;\; q_1 \;\; q_2 \;\; q_3]^T$ per unit length be applied on the x_3-axis which also includes a line dislocation with Burgers vector

$\boldsymbol{b} = [b_1\ \ b_2\ \ b_3\ \ d_1\ \ d_2\ \ d_3]^T$. The analytic vector function $\boldsymbol{f}(z)$ due to the line force and dislocation is given by

$$\boldsymbol{f}(z) = \langle \ln z_\alpha \rangle \frac{\boldsymbol{A}^T \boldsymbol{p} + \boldsymbol{B}^T \boldsymbol{b}}{2\pi \mathrm{i}}, \tag{28}$$

where $\langle * \rangle$ is a 6×6 diagonal matrix in which each component varies with the index α (from 1 to 6). The elastic energy for the annular region $0 < r_0 \le r \le R_0$ per unit length of the line force and dislocation is

$$U = \frac{1}{4\pi} \ln\Big(\frac{R_0}{r_0}\Big)(\boldsymbol{p}^T \boldsymbol{H} \boldsymbol{p} + \boldsymbol{b}^T \boldsymbol{L} \boldsymbol{b}) > 0, \tag{29}$$

which also provides an indirect proof that $\boldsymbol{H}$ and $\boldsymbol{L}$ must be positive definite if the elastic energy is to remain positive.

4.2. ***Interfacial Green's function.*** We consider a bimaterial made of two bonded dissimilar anisotropic quasicrystalline half-planes, denoted by #1 ($x_2 > 0$) and #2 ($x_2 < 0$). The bimaterial is subjected to a line force $\boldsymbol{p}$ and line dislocation with Burgers vector $\boldsymbol{b}$ at the origin. In fact the solution can be found from [Ting 1996]. The elastic energy for the annular region $0 < r_0 \le r \le R_0$ of the quasicrystalline bimaterial is

$$U = \frac{1}{2\pi} \ln\Big(\frac{R_0}{r_0}\Big)(\boldsymbol{p}^T \tilde{\boldsymbol{H}} \boldsymbol{p} + \boldsymbol{b}^T \tilde{\boldsymbol{L}} \boldsymbol{b}) > 0, \tag{30}$$

where $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{L}}$ are two 6×6 positive definite real symmetric matrices given by

$$\tilde{\boldsymbol{H}} = \mathrm{Re}\{(\boldsymbol{M}_1 + \overline{\boldsymbol{M}}_2)^{-1}\}, \quad \tilde{\boldsymbol{L}} = \mathrm{Re}\{(\boldsymbol{M}_1^{-1} + \overline{\boldsymbol{M}}_2^{-1})^{-1}\}. \tag{31}$$

4.3. ***Green's function for quasicrystalline bimaterials.*** We consider a bimaterial made of two perfectly bonded dissimilar anisotropic quasicrystalline half-planes again denoted by #1 ($x_2 > 0$) and #2 ($x_2 < 0$). A line force $\boldsymbol{p}$ and line dislocation with Burgers vector $\boldsymbol{b}$ are applied at $(x_1, x_2) = (0, \delta)$ ($\delta > 0$) in material #1. The structure of the solution is similar to that in [Suo 1990] and [Ting 1996]. The image force tending to move a pure dislocation ($\boldsymbol{p} = \boldsymbol{0}$) away from the interface is described by

$$F = \frac{1}{4\pi\delta} \boldsymbol{b}^T (2\tilde{\boldsymbol{L}} - \boldsymbol{L}_1) \boldsymbol{b}, \tag{32}$$

where $\tilde{\boldsymbol{L}}$ is given by (31). For a sliding interface on which $\sigma_{12} = \sigma_{32} = 0$, the image force acting on the line dislocation is now characterized by

$$F = \frac{1}{4\pi\delta} (2\boldsymbol{b}_0^T \hat{\boldsymbol{L}} \boldsymbol{b}_0 - \boldsymbol{b}^T \boldsymbol{L}_1 \boldsymbol{b}), \tag{33}$$

where

$$\boldsymbol{b}_0 = [b_2 \;\; d_1 \;\; d_2 \;\; d_3]^T, \quad \hat{\boldsymbol{L}} = \mathrm{Re}\{\hat{\boldsymbol{N}}^{-1}\}, \quad \hat{\boldsymbol{N}} = \begin{bmatrix} N_{22} & N_{24} & N_{25} & N_{26} \\ \bar{N}_{24} & N_{44} & N_{45} & N_{46} \\ \bar{N}_{25} & \bar{N}_{45} & N_{55} & N_{56} \\ \bar{N}_{26} & \bar{N}_{46} & \bar{N}_{56} & N_{66} \end{bmatrix}, \tag{34}$$

N_{ij} being the components of the 6×6 Hermitian matrix $\boldsymbol{N}$ defined by (14). Both $\hat{\boldsymbol{N}}$ and $\hat{\boldsymbol{L}}$ are positive definite.

5. Imperfect interface models

In [Benveniste 2006], the author rigorously derives imperfect interface models for a three-dimensional curved interphase between two anisotropic crystalline solids by making use of the idea of Taylor expansion of the corresponding fields in thin regions. In this section we will derive compliant and stiff interface models in anisotropic quasicrystalline bimaterials using the Stroh formalism and the structures of $\boldsymbol{N}_i$ presented in Sections 2 and 3. To simplify the analysis, we consider here a straight imperfect interface. The stress-strain law for an interphase of constant thickness h between the upper and lower anisotropic quasicrystalline materials 1 and 2 is described by

$$\boldsymbol{\sigma}_1 = \boldsymbol{Q}_c \boldsymbol{u}_{,1} + \boldsymbol{R}_c \boldsymbol{u}_{,2}, \; \boldsymbol{\sigma}_2 = \boldsymbol{R}_c^T \boldsymbol{u}_{,1} + \boldsymbol{T}_c \boldsymbol{u}_{,2}, \tag{35}$$

where $\boldsymbol{Q}$, $\boldsymbol{R}$, $\boldsymbol{T}$ are defined in (7) with the subscript c being used to identify the quantities associated with the intermediate interphase, and

$$\boldsymbol{\sigma}_1 = [\sigma_{11} \;\; \sigma_{21} \;\; \sigma_{31} \;\; H_{11} \;\; H_{21} \;\; H_{31}]^T, \boldsymbol{\sigma}_2 = [\sigma_{12} \;\; \sigma_{22} \;\; \sigma_{32} \;\; H_{12} \;\; H_{22} \;\; H_{32}]^T. \tag{36}$$

(1) If we assume that $C^{(c)}_{ijkl} \ll C^{(1)}_{ijkl}, C^{(2)}_{ijkl}$; $R^{(c)}_{ijkl} \ll R^{(1)}_{ijkl}, R^{(2)}_{ijkl}$ and $K^{(c)}_{ijkl} \ll K^{(1)}_{ijkl}, K^{(2)}_{ijkl}$ (the so-called compliant interphase) and that the interphase is also very thin, then it follows from $(35)_2$ that

$$\boldsymbol{u}_1 - \boldsymbol{u}_2 = h\boldsymbol{N}_2^{(c)}\boldsymbol{\sigma}_2^{(1)} = h\boldsymbol{N}_2^{(c)}\boldsymbol{\sigma}_2^{(2)} \text{ on the compliant interface.} \tag{37}$$

This is, in fact, the anisotropic spring-type interface.

(2) If we assume that $C^{(c)}_{ijkl} \gg C^{(1)}_{ijkl}, C^{(2)}_{ijkl}$ and $K^{(c)}_{ijkl} \gg K^{(1)}_{ijkl}, K^{(2)}_{ijkl}$ (the so-called stiff interphase) and that the interphase is also very thin, then it follows from (35) and the equilibrium equation $\boldsymbol{\sigma}_{1,1} + \boldsymbol{\sigma}_{2,2} = \boldsymbol{0}$ that

$$\boldsymbol{u}_1 = \boldsymbol{u}_2, \quad \boldsymbol{\sigma}_2^{(1)} - \boldsymbol{\sigma}_2^{(2)} = h\boldsymbol{N}_3^{(c)}\boldsymbol{u}_{1,11} = h\boldsymbol{N}_3^{(c)}\boldsymbol{u}_{2,11} \text{ on the stiff interface.} \tag{38}$$

This is an extension of the Young–Laplace model to anisotropic quasicrystalline materials.

In view of the structure of $\boldsymbol{N}_3$ given by (18), the phonon normal traction component σ_{22} is *continuous* across the stiff interface.

6. Interface crack problems

First we derive the asymptotic fields associated with a traction-free semi-infinite crack which lies along the interface between the upper and lower anisotropic quasicrystalline half-planes #1 ($x_2 > 0$) and #2 ($x_2 < 0$). The portion $x_2 = 0$, $x_1 > 0$ of the interface is perfectly bonded, whereas the remaining portion $x_2 = 0$ and $x_1 < 0$ of the interface is fully debonded. In the following analysis, in view of the fact that $z_1 = z_2 = z_3 = z_4 = z_5 = z_6 = z$ on the interface $x_2 = 0$, we will replace the complex variables z_k ($k = 1, \ldots, 6$) by the common complex variable $z = x_1 + \mathrm{i}x_2$. After the analysis is completed, we can simply revert back to the corresponding complex variables.

We introduce an analytic vector function $\boldsymbol{h}(z)$ defined by

$$\boldsymbol{h}(z) = \boldsymbol{B}_1 \boldsymbol{f}_1'(z) = \boldsymbol{N}^{-1} \overline{\boldsymbol{N}} \boldsymbol{B}_2 \boldsymbol{f}_2'(z). \tag{39}$$

Consequently the traction and displacement jumps can be expressed in terms of $\boldsymbol{h}(z)$ as

$$\boldsymbol{\sigma}_2(x_1) = \boldsymbol{h}^+(x_1) + \overline{\boldsymbol{N}}^{-1} \boldsymbol{N} \boldsymbol{h}^-(x_1), \quad \mathrm{i}\boldsymbol{d}'(x_1) = \boldsymbol{N}[\boldsymbol{h}^+(x_1) - \boldsymbol{h}^-(x_1)], \tag{40}$$

We then arrive at the following homogeneous vector Riemann–Hilbert problem:

$$\begin{aligned} \boldsymbol{h}^+(z) - \boldsymbol{h}^-(z) &= \boldsymbol{0}, \quad z \notin C, \\ \overline{\boldsymbol{N}} \boldsymbol{h}^+(z) + \boldsymbol{N} \boldsymbol{h}^-(z) &= \boldsymbol{0}, \quad z \in C. \end{aligned} \tag{41}$$

Consider the eigenvalue problem

$$\overline{\boldsymbol{N}} \boldsymbol{w} = \mathrm{e}^{2\pi\epsilon} \boldsymbol{N} \boldsymbol{w}. \tag{42}$$

It can be concluded that:

(1) As a result of the positive definiteness of $\boldsymbol{N}$, the eigenvalue $\mathrm{e}^{2\pi\epsilon}$ is always positive; thus ϵ is real.

(2) If $(\epsilon, \boldsymbol{w})$ is an eigenpair, then $(-\epsilon, \overline{\boldsymbol{w}})$ is another eigenpair.

Three positive real numbers ϵ_1, ϵ_2, ϵ_3 and three complex vectors $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, $\boldsymbol{w}_3$ form six eigenpairs:

$$(\epsilon_1, \boldsymbol{w}_1),\ (-\epsilon_1, \overline{\boldsymbol{w}}_1),\ (\epsilon_2, \boldsymbol{w}_2),\ (-\epsilon_2, \overline{\boldsymbol{w}}_2),\ (\epsilon_3, \boldsymbol{w}_3),\ (-\epsilon_3, \overline{\boldsymbol{w}}_3). \tag{43}$$

The following orthogonal relationships can also be established:

$$\boldsymbol{w}_i^T \boldsymbol{N} \boldsymbol{w}_j = \overline{\boldsymbol{w}}_k^T \boldsymbol{N} \boldsymbol{w}_l = \boldsymbol{w}_k^T \boldsymbol{N} \overline{\boldsymbol{w}}_l = 0 \quad (i, j, k, l = 1, 2, 3 \text{ and } k \neq l), \tag{44}$$

The three positive real numbers (or oscillatory indices) ϵ_1, ϵ_2, ϵ_3 are given by

$$\epsilon_j = \frac{1}{2\pi} \ln \frac{1+\beta_j}{1-\beta_j} \quad (j = 1, 2, 3), \tag{45}$$

where β_j $(j = 1, 2, 3)$ are the three positive roots of the following cubic equation in β^2:

$$\beta^6 + \tfrac{1}{2}\mathrm{tr}(\check{\boldsymbol{S}}^2)\beta^4 - \tfrac{1}{2}|\check{\boldsymbol{S}}|\,\mathrm{tr}(\check{\boldsymbol{S}}^{-2})\beta^2 - |\check{\boldsymbol{S}}| = 0. \tag{46}$$

Here

$$\check{\boldsymbol{S}} = (\boldsymbol{L}_1^{-1} + \boldsymbol{L}_2^{-1})^{-1}(\boldsymbol{S}_1\boldsymbol{L}_1^{-1} - \boldsymbol{S}_2\boldsymbol{L}_2^{-1}). \tag{47}$$

The oscillatory indexes ϵ_1, ϵ_2, ϵ_3 are then explicitly determined.

Following Suo [1990], we can obtain for $\boldsymbol{h}(z)$ the expression

$$\boldsymbol{h}(z) = \sum_{j=1}^{3} \frac{\mathrm{e}^{\pi\epsilon_j} K_j z^{\mathrm{i}\epsilon_j} \boldsymbol{w}_j + \mathrm{e}^{-\pi\epsilon_j} \bar{K}_j z^{-\mathrm{i}\epsilon_j} \overline{\boldsymbol{w}}_j}{2(2\pi z)^{\frac{1}{2}} \cosh \pi\epsilon_j}, \tag{48}$$

where K_1, K_2 and K_3 are three complex stress intensity factors. $\boldsymbol{f}_1'(z)$ and $\boldsymbol{f}_2'(z)$ in the two half-planes are then given by

$$\begin{aligned} \boldsymbol{B}_1 \boldsymbol{f}_1'(z) &= \sum_{j=1}^{3} \frac{\mathrm{e}^{\pi\epsilon_j} K_j z^{\mathrm{i}\epsilon_j} \boldsymbol{w}_j + \mathrm{e}^{-\pi\epsilon_j} \bar{K}_j z^{-\mathrm{i}\epsilon_j} \overline{\boldsymbol{w}}_j}{2(2\pi z)^{\frac{1}{2}} \cosh \pi\epsilon_j} \quad (\mathrm{Im}\{z\} > 0), \\ \boldsymbol{B}_2 \boldsymbol{f}_2'(z) &= \sum_{j=1}^{3} \frac{\mathrm{e}^{-\pi\epsilon_j} K_j z^{\mathrm{i}\epsilon_j} \boldsymbol{w}_j + \mathrm{e}^{\pi\epsilon_j} \bar{K}_j z^{-\mathrm{i}\epsilon_j} \overline{\boldsymbol{w}}_j}{2(2\pi z)^{\frac{1}{2}} \cosh \pi\epsilon_j} \quad (\mathrm{Im}\{z\} < 0). \end{aligned} \tag{49}$$

The traction at the bonded interface a distance r ahead of the crack tip is

$$\boldsymbol{\sigma}_2(r) = \sum_{j=1}^{3} [t_j(r)\boldsymbol{w}_j + \bar{t}_j(r)\overline{\boldsymbol{w}}_j], \tag{50}$$

where

$$t_j(r) = t_{2j} + \mathrm{i}t_{1j} = \frac{\overline{\boldsymbol{w}}_j^T \boldsymbol{N}\boldsymbol{\sigma}_2(r)}{\overline{\boldsymbol{w}}_j^T \boldsymbol{N}\boldsymbol{w}_j} = \frac{K_j r^{\mathrm{i}\epsilon_j}}{\sqrt{2\pi r}}.$$

Equation (50) states that the interface traction can be decomposed into three components, each in the plane spanned by $\mathrm{Re}\{\boldsymbol{w}_j\}$ and $\mathrm{Im}\{\boldsymbol{w}_j\}$ $(j = 1, 2, 3)$.

The displacement jump a distance r behind the crack tip is

$$\boldsymbol{d}(r) = (\boldsymbol{N} + \overline{\boldsymbol{N}})\Big(\frac{r}{2\pi}\Big)^{\frac{1}{2}} \sum_{j=1}^{3} \left[\frac{K_j r^{\mathrm{i}\epsilon_j} \boldsymbol{w}_j}{(1 + 2\mathrm{i}\epsilon_j)\cosh \pi\epsilon_j} + \frac{\bar{K}_j r^{-\mathrm{i}\epsilon_j} \overline{\boldsymbol{w}}_j}{(1 - 2\mathrm{i}\epsilon_j)\cosh \pi\epsilon_j} \right]. \tag{51}$$

The energy release rate is therefore given by

$$G = \lim_{\delta\to 0} \frac{1}{2\delta} \int_0^\delta \boldsymbol{\sigma}_2^T(\delta - r)\boldsymbol{d}(r)\, dr = \sum_{j=1}^{3} \frac{\overline{\boldsymbol{w}}_j^T(\boldsymbol{N} + \overline{\boldsymbol{N}})\boldsymbol{w}_j}{4\cosh^2 \pi\epsilon_j} |K_j|^2 > 0. \tag{52}$$

Next, we consider a set of collinear cracks between two dissimilar anisotropic quasicrystalline half-planes with prescribed traction $\boldsymbol{t}_0(x_1)$ on the crack lines C. Suppose there are n finite cracks in the intervals (a_j, b_j) $(j = 1, 2, \ldots, n)$ and two semi-infinite cracks $(-\infty, b_0)$ and $(a_0, +\infty)$. The prescribed traction $\boldsymbol{t}_0(x_1)$ on the crack lines C will result in the inhomogeneous Riemann–Hilbert vector problem:

$$\boldsymbol{h}^+(x_1) + \overline{\boldsymbol{N}}^{-1}\boldsymbol{N}\boldsymbol{h}^-(x_1) = \boldsymbol{t}_0(x_1), \quad x_1 \in C. \tag{53}$$

In order to solve this, we follow the method in [Suo 1990] and [Suo et al. 1992] and write $\boldsymbol{h}(z)$ and $\boldsymbol{t}_0(x_1)$ in terms of their components using the eigenvector representation

$$\boldsymbol{h}(z) = h_1(z)\boldsymbol{w}_1 + h_2(z)\overline{\boldsymbol{w}}_1 + h_3(z)\boldsymbol{w}_2 + h_4(z)\overline{\boldsymbol{w}}_2 + h_5(z)\boldsymbol{w}_3 + h_6(z)\overline{\boldsymbol{w}}_3,$$
$$\boldsymbol{t}_0(x_1) = t_{01}(x_1)\boldsymbol{w}_1 + \bar{t}_{01}(x_1)\overline{\boldsymbol{w}}_1 + t_{02}(x_1)\boldsymbol{w}_2 + \bar{t}_{02}(x_1)\overline{\boldsymbol{w}}_2 + t_{03}(x_1)\boldsymbol{w}_3 + \bar{t}_{03}(x_1)\overline{\boldsymbol{w}}_3.$$

As a result, (53) can be decoupled as follows:

$$\left.\begin{aligned}
h_1^+(x_1) + \mathrm{e}^{-2\pi\epsilon_1} h_1^-(x_1) &= t_{01}(x_1)\\
h_2^+(x_1) + \mathrm{e}^{+2\pi\epsilon_1} h_2^-(x_1) &= \bar{t}_{01}(x_1)\\
h_3^+(x_1) + \mathrm{e}^{-2\pi\epsilon_2} h_3^-(x_1) &= t_{02}(x_1)\\
h_4^+(x_1) + \mathrm{e}^{+2\pi\epsilon_2} h_4^-(x_1) &= \bar{t}_{02}(x_1)\\
h_5^+(x_1) + \mathrm{e}^{-2\pi\epsilon_3} h_5^-(x_1) &= t_{03}(x_1)\\
h_6^+(x_1) + \mathrm{e}^{+2\pi\epsilon_3} h_6^-(x_1) &= \bar{t}_{03}(x_1)
\end{aligned}\right\} \quad \text{for } x_1 \in C, \tag{54}$$

whose solution can be given simply by

$$h_1(z) = \frac{\chi_1(z)}{2\pi\mathrm{i}} \int_C \frac{t_{01}(x_1)\, dx_1}{\chi_1^+(x_1)(x_1 - z)} + \chi_1(z) P_1(z),$$
$$h_2(z) = \frac{\bar{\chi}_1(z)}{2\pi\mathrm{i}} \int_C \frac{\bar{t}_{01}(x_1)\, dx_1}{\bar{\chi}_1^+(x_1)(x_1 - z)} + \bar{\chi}_1(z) P_2(z),$$
$$h_3(z) = \frac{\chi_2(z)}{2\pi\mathrm{i}} \int_C \frac{t_{02}(x_1)\, dx_1}{\chi_2^+(x_1)(x_1 - z)} + \chi_2(z) P_3(z),$$
$$h_4(z) = \frac{\bar{\chi}_2(z)}{2\pi\mathrm{i}} \int_C \frac{\bar{t}_{02}(x_1)\, dx_1}{\bar{\chi}_2^+(x_1)(x_1 - z)} + \bar{\chi}_2(z) P_4(z),$$
$$h_5(z) = \frac{\chi_3(z)}{2\pi\mathrm{i}} \int_C \frac{t_{03}(x_1)\, dx_1}{\chi_3^+(x_1)(x_1 - z)} + \chi_3(z) P_5(z),$$

$$h_6(z) = \frac{\bar{\chi}_3(z)}{2\pi \mathrm{i}} \int_C \frac{\bar{t}_{03}(x_1)\, dx_1}{\bar{\chi}_3^+(x_1)(x_1 - z)} + \bar{\chi}_3(z) P_6(z),$$

where $\chi_1(z)$, $\chi_2(z)$ and $\chi_3(z)$ are defined as

$$\chi_j(z) = \prod_{k=0}^{n} (z - a_k)^{-\frac{1}{2} - \mathrm{i}\epsilon_j} (z - b_k)^{-\frac{1}{2} + \mathrm{i}\epsilon_j} \quad (j = 1, 2, 3), \tag{55}$$

and $P_i(z)$ $(i = 1, \ldots, 6)$ are polynomials in z of order less than $n + 1$ [Ting 1996].

7. Conclusions

In this paper all six 6×6 matrices $\boldsymbol{N}_i$ and $\boldsymbol{N}_i^{(-1)}$ $(i = 1, 2, 3)$ in the Stroh formalism of anisotropic elasticity for quasicrystals have been explicitly expressed in terms of the sixty-six reduced elastic compliances $S_{ij} = S_{ji}$ $(i, j = 1 \ldots 15$ and $i, j \neq 3, 9, 10, 14)$. It is found that $\boldsymbol{N}_2$ and $-\boldsymbol{N}_2^{(-1)}$ are positive definite, whilst $-\boldsymbol{N}_3$ and $\boldsymbol{N}_3^{(-1)}$ are positive semidefinite.

In the study of Green's functions, we present the elastic energy expressions (29) for a line force and dislocation in a homogeneous quasicrystal and (30) for a line force and dislocation lying on a bi-quasicrystal interface, and obtain the image force on a dislocation near a perfect or sliding interface between two anisotropic quasicrystalline half-planes.

We also derive compliant and stiff interface models using the Stroh formalism. Green's function solutions for quasicrystalline bimaterials with imperfect interface can be further studied by using the method described in [Wang and Pan 2010].

Perhaps the most interesting conclusion from this research is that the interface crack-tip field consists of *three* two-dimensional oscillatory stress singularities $r^{-\frac{1}{2} \pm \mathrm{i}\epsilon_j}$ $(j = 1, 2, 3)$ characterized through the introduction of three complex stress intensity factors K_j $(j = 1, 2, 3)$. We end by again noting the beauty and power of the Stroh formalism which is fully demonstrated here.

Acknowledgements

Wang W. was supported by Innovation Program of Shanghai Municipal Education Commission (No. 12ZZ058). Schiavone acknowledges the support of the Natural Sciences and Engineering Research Council of Canada.

References

[Barnett and Lothe 1973] D. M. Barnett and J. Lothe, "Synthesis of the sextic and the integral formalism for dislocations, Green's functions and surface waves in anisotropic elastic solids", *Phys. Norv.* **7** (1973), 13–19.

[Benveniste 2006] Y. Benveniste, "A general interface model for a three-dimensional curved thin anisotropic interphase between two anisotropic media", *J. Mech. Phys. Solids* **54**:4 (2006), 708–734.

[Cheng and Reddy 2002] Z.-Q. Cheng and J. N. Reddy, "Octet formalism for Kirchhoff anisotropic plates", *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.* **458**:2022 (2002), 1499–1517.

[Clements 1971] D. L. Clements, "A crack between dissimilar anisotropic media", *Int. J. Eng. Sci.* **9**:2 (1971), 257–265.

[Ding et al. 1995] D.-H. Ding, R. Wang, W. Yang, C. Hu, and Y. Qin, "Elasticity theory of straight dislocations in quasicrystals", *Phil. Mag. Lett.* **72**:5 (1995), 353–359.

[Eshelby et al. 1953] J. D. Eshelby, W. T. Read, and W. Shockley, "Anisotropic elasticity with applications to dislocation theory", *Acta Metall.* **1**:3 (1953), 251–259.

[Gao et al. 1992] H. J. Gao, M. Abbudi, and D. M. Barnett, "On interfacial crack-tip field in anisotropic elastic solids", *J. Mech. Phys. Solids* **40**:2 (1992), 393–416.

[Hu et al. 2000] C. Hu, R. Wang, and D.-H. Ding, "Symmetry groups, physical property tensors, elasticity and dislocations in quasicrystals", *Rep. Progr. Phys.* **63**:1 (2000), 1–39.

[Lazar and Kirchner 2005] M. Lazar and H. O. K. Kirchner, "Cosserat (micropolar) elasticity in Stroh form", *Int. J. Solids Struct.* **42**:20 (2005), 5377–5398.

[Lekhnitskii 1950] S. G. Lekhnitskii, Теория упругости анизотропоного тела, Moscow, Gosud. Izdat. Tekhn.-Teor. Lit., 1950. Translated as *Theory of elasticity of an anisotropic body*, Holden-Day, San Francisco, 1963.

[Ru 2001] C. Q. Ru, "A two-dimensional Eshelby problem for two bonded piezoelectric half-planes", *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.* **457**:2008 (2001), 865–883.

[Shechtman et al. 1984] D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn, "Metallic phase with long-range orientational order and no translational symmetry", *Phys. Rev. Lett.* **53** (1984), 1951–1953.

[Stroh 1958] A. N. Stroh, "Dislocations and cracks in anisotropic elasticity", *Phil. Mag.* (8) **3** (1958), 625–646.

[Suo 1990] Z. Suo, "Singularities, interfaces and cracks in dissimilar anisotropic media", *Proc. Roy. Soc. London Ser. A* **427**:1873 (1990), 331–358.

[Suo et al. 1992] Z. Suo, C.-M. Kuo, D. M. Barnett, and J. R. Willis, "Fracture mechanics for piezoelectric ceramics", *J. Mech. Phys. Solids* **40**:4 (1992), 739–765.

[Ting 1986] T. C. T. Ting, "Explicit solution and invariance of the singularities at an interface crack in anisotropic composites", *Internat. J. Solids Structures* **22**:9 (1986), 965–983.

[Ting 1988] T. C. T. Ting, "Some identities and the structure of $\mathbf{N}_i$ in the Stroh formalism of anisotropic elasticity", *Quart. Appl. Math.* **46**:1 (1988), 109–120.

[Ting 1996] T. C. T. Ting, *Anisotropic elasticity: Theory and applications*, Oxford Engineering Science Series **45**, Oxford University Press, New York, 1996.

[Ting and Schiavone 2010] T. C. T. Ting and P. Schiavone, "Uniform antiplane shear stress inside an anisotropic elastic inclusion of arbitrary shape with perfect or imperfect interface bonding", *Internat. J. Engrg. Sci.* **48**:1 (2010), 67–77.

[Wang and Pan 2010] X. Wang and E. Pan, "Two-dimensional Eshelby's problem for two imperfectly bonded piezoelectric half-planes", *Int. J. Solids Struct.* **47**:1 (2010), 148–160.

[Willis 1964] J. R. Willis, "Anisotropic elastic inclusion problems", *Quart. J. Mech. Appl. Math.* **17** (1964), 157–174.

[Willis 1970] J. R. Willis, "Stress fields produced by dislocations in anisotropic media", *Philo. Mag.* **21**:173 (1970), 931–949.

[Willis 1971] J. R. Willis, "Fracture mechanics of interfacial cracks", *J. Mech. Phys. Solids* **19**:6 (1971), 353–368.

Received 23 Feb 2012. Accepted 5 Apr 2012.

XU WANG: xuwang_sun@hotmail.com
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, AB T6G 2G8, Canada
http://www.mece.ualberta.ca/∼schiavone/schiavon.htm

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 1, No. 1, 2013
dx.doi.org/10.2140/memocs.2013.1.19

LOCALIZATION OF POINT VORTICES UNDER CURVATURE PERTURBATIONS

ROBERTO GARRA

We discuss the effect of curvature on the dynamics of a two-dimensional inviscid incompressible fluid with initial vorticity concentrated in N small disjoint regions, that is, the classical point vortex system. We recall some results about point vortex dynamics on simply connected surfaces with constant curvature K, that is, plane, spherical, and hyperbolic surfaces. We show that the effect of curvature can be treated as a smooth perturbation to the Green's function of the equation related to the stream function in the planar case. Then we obtain as a main result that the localization property of point vortices, already proved for the plane, is preserved also under the effect of curvature perturbation.

1. Introduction

Vortex dynamics is a fundamental topic in fluid mechanics. In the framework of ideal incompressible fluid it is described by the Euler equation. A classical approximation made in order to study vortex dynamics analytically in two dimensions is to treat singular vorticity distributions. This means replacing a partial differential equation with infinite degrees of freedom with a system of ordinary differential equations with N degrees of freedom. This point vortex model was first introduced by Helmholtz in 1858 and Kirchhoff in 1876; it was also treated in classic textbooks like [Batchelor 1967] (for a detailed historical review see [Llewellyn Smith 2011]). The study of point vortex dynamics is still an important topic in mathematical physics, a "classical mathematics playground" as stated in [Aref 2007]. It finds its physical roots in the analysis of the dynamics of a two-dimensional inviscid incompressible fluid with initial vorticity sharply concentrated in N small disjoint regions. There are many papers devoted to the mathematical analysis of this model in the framework of dynamical systems (see, for example, [Newton 2001]) and mathematical fluid mechanics (see, for example, [Marchioro and Pulvirenti 1994]). However a critical point of this model is the divergence of the velocity computed in the point where the single vortex is localized. This (infinite) term could be skipped in a heuristic way from a physical point of view because it is a self-interaction term.

Communicated by Carlo Marchioro.
MSC2010: 76B47.
Keywords: point vortex dynamics, mathematical fluid mechanics, ideal fluid.

But from a mathematical point of view a rigorous connection between the Euler equation and the point vortex model is given by the proof that, if

$$\omega_0(x)\,dx \to \sum_{i=1}^{N} a_i \delta_{x_i}(dx),$$

then

$$\omega_t(x)\,dx \to \sum_{i=1}^{N} a_i \delta_{x_i(t)}(dx).$$

The proof of this fact was firstly given in [Marchioro and Pulvirenti 1993]. In more detail, it was proved there (and in [Marchioro 1998] that the time evolution of a system of vortices initially concentrated in N small disjoint regions of diameter ϵ remains concentrated in N disjoint vortices with diameter $d(\epsilon) \to 0$ as $\epsilon \to 0$. This property of the point vortex model is called *localization*.

Here we analyze the effect of curvature on the dynamics of N sharply concentrated vortices. There are a number of papers [Hally 1980; Kimura 1999] devoted to the analysis of vortex motion on surfaces with constant curvature, for example, on spheres [Crowdy 2006] and hyperbolic surfaces [Hwang and Kim 2009]. Moreover, Boatto [2008] treated the perturbative effect of curvature on the stability of a ring of vortices.

The main aim of this paper is to prove that the property of localization of the dynamics of point vortex motion is preserved under the effect of curvature perturbation. Actually, we show that the effect of the constant curvature K of the surface on the dynamics can be treated by means of a smooth perturbation to the Green's function of the plane case. Then we include this perturbation in an external mean field and we show that the localization of the vortices under the effect of curvature is essentially a corollary to the theorem of localization stated in [Marchioro and Pulvirenti 1993].

This result is really interesting from a physical point of view because it states that strong concentrated vortices remain concentrated under the effect of curvature. For example, in the spherical case we can apply it to the dynamics of vortices over the Earth's surface. Moreover, we can generalize this result to any regular surface that can be locally approximated with a Riemannian manifold with constant curvature K.

The plan of the paper is as follows: in Section 2 we introduce the constitutive equations of the point vortex model, in Section 3 we recall some useful results about point vortex dynamics on surfaces with constant curvature, and in Section 4 we discuss the main result, recalling the localization theorem in the planar case and proving that it also works taking into account the effect of curvature.

2. Point vortex motion in fluid mechanics

Here we introduce the constitutive equations of point vortex motion in the whole plane $\mathbb{R}^2$. Consider the Euler equation about a two-dimensional inviscid incompressible fluid with unitary density:

$$\partial_t \omega + (u \cdot \nabla)\omega = 0, \quad \nabla \cdot u = 0, \quad \omega = \operatorname{curl} u = \partial_1 u_2 - \partial_2 u_1, \tag{2-1}$$

with boundary condition $u \to 0$ as $|x| \to \infty$. Here $u \equiv (u_1, u_2)$ denotes the velocity field.

Then we define the stream function $\psi(x, t)$ such that $u(x, t) = \nabla^\perp \psi(x, t)$, with $\nabla^\perp \equiv (\partial_2, -\partial_1)$. It is immediate to see that

$$\omega(x, t) = -\nabla^2 \psi(x, t), \tag{2-2}$$

that is, a Poisson-type equation with ω as a source term. We notice that formally the stream function plays the role of a Hamiltonian; this explains the great interest in the point vortex system in the field of dynamical systems.

By using the definition of a stream function, we find the explicit form of the velocity field by means of the Green's function of (2-2):

$$u(x, t) = \nabla^\perp \psi = \int \nabla^\perp G(r, r')\omega(r')\, dr', \tag{2-3}$$

depending on the initial conditions on the vorticity and the domain. If the initial vorticity field is generated by N disjoint point vortices, we use an initial condition given by a measure

$$\omega(x, 0)\, dx = \sum_{i=1}^{N} a_i \delta_{x_i}(dx), \tag{2-4}$$

where a_i is the vortex intensity of the i-vortex situated at x_i. This is the so-called point vortex system. The dynamics of the N point vortex system is defined by the Green's function of (2-2). It clearly depends on the domain. For example, in the whole plane $\mathbb{R}^2$, the evolution equations for a system of point vortices is given by

$$\frac{dx_i(t)}{dt} = -\nabla^\perp \sum_{j=1;\, j\neq i}^{N} a_j G(x_i(t), x_j(t)), \quad x_i(t = 0) = x_i, \tag{2-5}$$

where $G(x_i, x_j) = 1/(2\pi) \ln|x_i(t) - x_j(t)|$ is the Green's function of (2-2) in the plane. It appears as a discrete solution of the Euler equation.

Starting from this mathematical formulation, there are a great number of possible investigations about the point vortex system. In the framework of mathematical fluid mechanics, a wide discussion of the properties of such systems and the rigorous relation with the Euler equation can be found in [Marchioro and Pulvirenti

1994]. In the framework of dynamical systems there are a great number of papers devoted to the analysis of integrability, relative equilibria, and applications; we refer to [Newton 2001].

In the next section we recall the explicit form of the Green's function in the planar, spherical, and hyperbolic cases.

3. Dynamics of point vortices on surfaces of constant curvature

Here we recall the main results about the Green's function of the Poisson equation over surfaces with constant curvature K. First of all, we recall that the three surfaces with constant curvature, a sphere ($K > 0$), a Euclidean plane ($K = 0$), and a hyperbolic plane ($K < 0$), can be considered as three different situations inside a family of Riemannian manifolds with the curvature K as a parameter. We refer to [Kimura 1999] for an unified geometrical setting of this problem. In this work the fundamental solution of the Poisson equation over a spherical surface is given as a function of the geodesic distance r from the north pole of the sphere, that is, $r = \theta R = \theta/\sqrt{K}$, with θ its colatitude and R the radius of the sphere. Then Kimura found, in a direct way, the Green's function for the hyperbolic case as a function of the same variables. We can prove that the Green's function G_K only depends upon the geodesic distance r and is given by

$$2\pi G_{K>0} = -\ln \sin \frac{\sqrt{K}\, r}{2} \qquad \text{for a spherical surface,} \qquad r \in \left(0, \frac{\pi}{\sqrt{K}}\right), \tag{3-1}$$

$$2\pi G_0 = -\ln r \qquad \text{for a plane,} \qquad r \in (0, \infty), \tag{3-2}$$

$$2\pi G_{K<0} = -\ln \tanh \frac{\sqrt{|K|}\, r}{2} \qquad \text{for a hyperbolic surface,} \qquad r \in (0, \infty). \tag{3-3}$$

Then if we take the difference $\Delta_{K>0}$ between (3-1) and (3-2), we obtain

$$\Delta_{K>0}(r) = -\ln \sin \frac{\sqrt{K}\, r}{2} + \ln r = \ln \frac{r}{\sin\left(\frac{1}{2}\sqrt{K} r\right)}. \tag{3-4}$$

This is a continuous function, with a bounded first derivative for $r \in (0, \pi/\sqrt{K})$ where the Green's function is defined, that is, it is also a Lipschitz function. This last statement has a central role in the following discussion. Actually, we can treat the effect of curvature as a Lipschitz perturbation to the Green's function of the planar case. The same reasoning can be applied in the hyperbolic case. In this case we find a Lipschitz function $\Delta_{K<0}$ for $r \in (0, \infty)$.

Moreover, it is simple to check by Taylor expansion that the planar case can be recovered in the limit $K \to 0$. In more detail, when considering the limit $K \to 0$,

we obtain:

$$2\pi G_{K>0} = -\ln[\sin(\sqrt{K}r/2)] + \ln(\sqrt{K}/2) \sim -\ln(r) - \frac{K}{24}r^2 + \dots, \tag{3-5}$$

$$2\pi G_{K<0} = -\ln[\tanh(\sqrt{|K|}r/2)] + \ln(\sqrt{|K|}/2) \sim -\ln(r) + \frac{|K|}{12}r^2 + \dots. \tag{3-6}$$

Then it's clear that the effect of the curvature on the dynamics can be parametrized as a smooth perturbation to the Green's function on the plane.

Finally we can write the Green's function of the Poisson equation over a surface with constant curvature as:

$$G(r) = G_0(r) + \Delta_K(r), \tag{3-7}$$

where $G_0 = -1/(2\pi)\ln(r)$ is the Green's function on the plane and $\Delta_K(r)$ is a Lipschitz perturbation dependent on the curvature K as previously defined.

This means that, from (2-3), the velocity field of the fluid over a surface with constant curvature is given by

$$u(x,t) = \nabla^\perp \psi(x,t) = u_0(x,t) + u_K(x,t), \tag{3-8}$$

where

$$u_0(x,t) = \int \nabla^\perp G_0(r,r')\omega(r')\,dr',$$

$$u_K(x,t) = \int \nabla^\perp \Delta_K(r,r')\omega(r')\,dr'.$$

As already discussed, we can treat the contribution u_K due to the curvature effect as a Lipschitz field. Then from a Lagrangian point of view the fluid particle satisfies the following equation:

$$\frac{dx(t)}{dt} = u_0(x,t) + u_K(x,t). \tag{3-9}$$

In the following we will use directly $u(x,t)$ for the velocity field of the planar case.

4. Localization of the vortices under curvature perturbation

In the planar case, we call "localization" the following property of the dynamics of a system of point vortices: the time evolution of N concentrated vortices, according to the Euler equation in the two-dimensional case, remains concentrated in N small disjoint regions of diameter $d(\epsilon) \to 0$ as $\epsilon \to 0$ [Marchioro and Pulvirenti 1993; Marchioro 1998]. This result provides a rigorous connection between the Euler equation and the point vortex model, giving a complete justification for skipping the divergent self-interaction term in the point vortex dynamics (for a full discussion of this point see [Marchioro and Pulvirenti 1994]).

In more detail, we recall the following localization theorem:

Theorem 4.1 [Marchioro 1998]. *Consider an initial datum*

$$\omega_\epsilon(x,0)=\sum_{i=1}^{N}\omega_{\epsilon;i}(x,0) \tag{4-1}$$

where $\omega_{\epsilon;i}(x,0)$ is a function with a definite sign supported in a region $\Lambda_{\epsilon;i}$ such that

$$\Lambda_{\epsilon;i}=\operatorname{supp}\omega_{\epsilon;i}(x,0)\subset\Sigma(z_i|\epsilon),\quad \Sigma(z_i|\epsilon)\cap\Sigma(z_j|\epsilon)=0 \text{ if } i\neq j, \tag{4-2}$$

for ϵ small enough. Here $\Sigma(z|r)$ denotes the circle of center z and radius r. The intensity of any single vortex is

$$\int dx\,\omega_{\epsilon;i}(x,0)\equiv a_i\in\mathbb{R}, \tag{4-3}$$

independent of ϵ and we assume

$$|\omega_{\epsilon;i}(x,0)|\le M\epsilon^{-\gamma},\quad M>0,\quad \gamma>0. \tag{4-4}$$

Denote by $\omega_\epsilon(x,t)$ the time evolution of (4-1) *according to the Euler equation with boundary condition $u\to 0$ as $|x|\to\infty$. Then, for any fixed time T, for any $\alpha\in\left[0,\frac{1}{3}\right)$ and $0\le t\le T$, we have*:

- *For all $d>0$, there exists $\epsilon_0(d,T)$ such that, if $\epsilon<\epsilon_0$, then $\omega_\epsilon(x,t)=\sum_{i=1}^{N}\omega_{\epsilon,i}(x,t)$. Moreover, $\operatorname{supp}\omega_{\epsilon;i}(x,t)\subset\Sigma(z_i(t)|d)$, where $d\to 0$ as $\epsilon\to 0$ and $z_i(t)$ is the solution of the differential system*

$$\dot z_i(t)=\sum_{j=1;i\neq j}^{N}a_i\nabla^\perp G(|z_i-z_j|),\quad \nabla^\perp=(\partial_2,-\partial_1),\quad z_i(0)=z_i, \tag{4-5}$$

  where $G(\cdot)$ is the Green's function of the Poisson equation in the planar case with vanishing boundary condition at infinity.

- *For any continuous bounded function $f(x)$*

$$\lim_{\epsilon\to 0}\int\omega_\epsilon(x,t)f(t)=\sum_{i}^{N}a_i f(z_i(t)). \tag{4-6}$$

The value of $T>0$ must be such that there are no collapses for any $t<T$; a complete discussion of the existence of such a T is given in [Marchioro and Pulvirenti 1994].

Note that this formulation is an improvement of the previous result stated in [Marchioro and Pulvirenti 1993], giving a much better estimate of the support $d(\epsilon)$ of the vortices.

The main step for the proof of this theorem is to study the localization of a single vortex, simulating the effect of the other $N-1$ vortices with a Lipschitz external field $F(x,t)$. In this case the motion of the vortex is described by the Euler equation in the weak form:

$$\frac{d}{dt}\omega(f) = \omega[(u+F)\cdot\nabla f], \tag{4-7}$$

where $\omega(f(x)) = \int dx\,\omega(x,t)f(x)$ and $f(x)$ is a bounded smooth function. From a Lagrangian point of view, we have

$$\frac{dx}{dt} = u(x,t) + F(x,t). \tag{4-8}$$

Then, defining the center of vorticity as

$$B_\epsilon(t) \equiv \int x\omega_\epsilon(x,t)\,dt, \tag{4-9}$$

we state the following theorem about the localization of a single blob:

Theorem 4.2. *Suppose that*

$$\mathrm{supp}|\omega_\epsilon(x,0)| \subset \Sigma(x^*|\epsilon) \tag{4-10}$$

and

$$|\omega_\epsilon(x,0)| \le M\epsilon^{-\gamma}, \quad M>0, \quad \gamma>0, \quad \int dx\,\omega_\epsilon(x,0) = 1. \tag{4-11}$$

Then, there exists $C(\beta,T)>0$, with $\beta>0$, such that for $0\le t\le T$

$$\mathrm{supp}|\omega_\epsilon(x,t)| \subset \Sigma(B(t)|d) \tag{4-12}$$

where

$$d = C(\beta,T)\epsilon^\beta, \tag{4-13}$$

and $B(t)$ is the solution of the ordinary differential equation

$$\frac{dB(t)}{dt} = F(B(t),t), \tag{4-14}$$

$$B(0) = x^*. \tag{4-15}$$

We refer to [Marchioro 1998] for the complete proof of this theorem, and in the Appendix we sketch the proof for the utility of the reader. Here we again remark that one of the central assumptions is about the Lipschitz continuity of the simulating external field.

Starting from these results, we can finally state our main result: the localization property of point vortices is preserved in surfaces with constant curvature.

Theorem 4.3. *Consider an initial datum*

$$\omega_\epsilon(x,0) = \sum_{i=1}^{N} \omega_{\epsilon;i}(x,0), \tag{4-16}$$

where $\omega_{\epsilon;i}(x,0)$ *is a function with a definite sign supported in a region* $\Lambda_{\epsilon;i}$ *such that*

$$\Lambda_{\epsilon;i} = \operatorname{supp} \omega_{\epsilon;i}(x,0) \subset \Sigma(z_i|\epsilon), \quad \Sigma(z_i|\epsilon) \cap \Sigma(z_j|\epsilon) = 0 \ \text{ if } i \neq j, \tag{4-17}$$

for ϵ *small enough. The intensity of any single vortex is*

$$\int dx\, \omega_{\epsilon;i}(x,0) \equiv a_i \in \mathbb{R}, \tag{4-18}$$

independent of ϵ *and we assume*

$$|\omega_{\epsilon;i}(x,0)| \leq M\epsilon^{-\gamma}, \quad M > 0, \quad \gamma > 0. \tag{4-19}$$

Denote by $\omega_\epsilon(x,t)$ *the time evolution of* (4-1) *on a surface of constant curvature* K *according to the Euler equation, then Theorem 4.1 holds.*

The proof of this theorem is similar to that of the planar case, considering first of all the localization of a single vortex. We have shown that the effect of curvature on the Green's function can be treated as a Lipschitz perturbation to the Green's function of the planar case. Then the localization of a single vortex is a corollary of Theorem 4.2

Corollary 4.4. *Consider a single point vortex such that*

$$\begin{gathered} \operatorname{supp}|\omega_\epsilon(x,0)| \subset \Sigma(x^*|\epsilon), \quad |\omega_\epsilon(x,0)| \leq M\epsilon^{-\gamma}, \\ M > 0, \quad \gamma > 0, \quad \int dx\, \omega_\epsilon(x,0) = 1. \end{gathered} \tag{4-20}$$

Denote by $\omega_\epsilon(x,t)$ *the time evolution on a surface with constant curvature* K*, according to the Euler equation. Then, there exists* $C(\beta,T) > 0$*, with* $\beta > 0$*, such that for* $0 \leq t \leq T$

$$\operatorname{supp}|\omega_\epsilon(x,t)| \subset \Sigma(B(t)|d), \tag{4-21}$$

where

$$d = C(\beta,T)\epsilon^\beta, \tag{4-22}$$

and $B(t)$ *is the solution of the ordinary differential equation*

$$\frac{dB(t)}{dt} = F_K(B(t),t), \quad B(0) = x^*. \tag{4-23}$$

where $F_K(x,t)$ is a Lipschitz field including in a single term the effect of the curvature (depending on K) and the effect of the other $N-1$ vortices on the dynamics of the single vortex.

The main improvement was proved in Section 3: the velocity term linked to the curvature effect is a Lipschitz function. Then we include the effect of curvature on the motion in a single Lipschitz term in (4-7). It is simple then to come back to the general case of the N vortices and to prove the main result.

We conclude that the localization theorem for point vortices moving on surfaces with constant curvature is a consequence of the analysis given in Section 3 about the effect of the curvature on vortex dynamics. Then its proof is exactly the same as that of the planar case discussed in [Marchioro 1998].

This result is valid for any regular surface. Actually, it is always possible to approximate locally these surfaces with manifolds with constant curvature K. Then we can develop exactly the same reasoning, including the curvature effect in an external Lipschitz continuous field. Moreover, the localization holds also in the presence of internal frontiers such as continents on the Earth's surface. Again, the physical meaning of this rigorous result is that it permits one to skip the singular part of the self-interacting term in the point vortex model, previously neglected in the basis of heuristic physical reasoning.

Appendix: Proof of Theorem 4.2

Here we give a synthetic idea of the rather technical proof of the localization theorem stated in [Marchioro 1998], recalling the fundamental steps. The main difficulty is due to the singularity of the kernel in the velocity expression

$$u(x,t) = \int K(x-y)\omega_\epsilon(y,t)\,dy, \tag{A.1}$$

where

$$K(x-y) = \nabla^\perp G(x-y) = \frac{\nabla^\perp \ln|x-y|}{2\pi}$$

in the planar case without boundaries.

First we introduce the moment of inertia I_ϵ with respect to the center of vorticity defined in (4-9):

$$I_\epsilon = \int \omega_\epsilon(x,t)(x - B_\epsilon(t))^2\,dx. \tag{A.2}$$

We want to show that the main part of the vorticity is concentrated around the center of vorticity. It is simple to prove that if $F = 0$ then B_ϵ and I_ϵ are constant along the motion, bringing us to (4-14).

If $F \neq 0$ then

$$\frac{dI}{dt} = 2\int (x - B_\epsilon(t))F(x,t)\omega_\epsilon(x,t)\,dx. \tag{A.3}$$

Then, by using the Lipschitz condition on $F(x,t)$ we find

$$\left|\frac{dI}{dt}\right| \leq 2L\int (x - B_\epsilon(t))^2\omega_\epsilon(x,t)\,dx = 2LI_\epsilon(t), \tag{A.4}$$

and integrating we obtain

$$\left|\frac{dI}{dt}\right| \leq I_0 e^{2Lt}, \tag{A.5}$$

so that

$$\lim_{\epsilon\to 0} I_\epsilon(t) = 0 \text{ at least as } \epsilon^2. \tag{A.6}$$

Hence we find that the main part of the vorticity remains concentrated around the center of vorticity. However we have to give an estimate of the mass and velocity of the filaments of vorticity generated by fluid particles near the boundaries and spreading out from the initially concentrated field. With this purpose we prove that the mass of vorticity near the boundary of the support is very small when $\epsilon \to 0$. Here the main technical complication is due to the singularity of the kernel in (A.1). First we introduce a nonnegative function $W_R \in C^\infty(\mathbb{R}^2)$ satisfying the following conditions, for a fixed $C_1 > 0$:

$$W_R(r) = \begin{cases} 1 & \text{if } |r| < R, \\ 0 & \text{if } |r| > 2R, \end{cases} \tag{A.7}$$

$$|\nabla W_R(r)| < \frac{C_1}{R}, \tag{A.8}$$

$$|\nabla W_R(r) - \nabla W_R(r')| < \frac{C_1}{R^2}|r - r'|. \tag{A.9}$$

Then we define a regularized measure of the mass of vorticity outside $\Sigma(B_\epsilon(t)|r)$:

$$\mu_t(R) = 1 - \int dx\, W_R(x - B_\epsilon(t))\omega_\epsilon(x,t), \tag{A.10}$$

such that if $\operatorname{supp}\omega_\epsilon(x,t) \subset \Sigma(B_\epsilon(t)|r)$ then $\mu_t(R) = 0$. Hence it gives a direct measure of the localization of the vorticity field.

We evaluate the growth in time of such a measure:

$$\frac{d\mu_t}{dt} = -\int dx\, \nabla W_R(x - B_\epsilon(t))\Big(u(x,t) + F(x,t) - \frac{dB}{dt}\Big)\omega_\epsilon(x,t). \tag{A.11}$$

Using (4-14) and (A.1), we obtain

$$\frac{d\mu_t}{dt} = -\int dx\, \omega_\epsilon(x,t)\nabla W_R(x-B_\epsilon(t)) \int dy\, K(x-y)\omega_\epsilon(y,t) \\ -\int dx\, \omega_\epsilon(x,t)\nabla W_R(x-B_\epsilon(t)) \int dy\, \omega_\epsilon(y,t)(F(x,t)-F(y,t)). \quad \text{(A.12)}$$

To give an estimate to the first term of (A.12), we split the integration domain into many different rings, defined by the following sets:

- if $h<n$, $T_h \equiv \{(x,y)|x \notin \Sigma(B_\epsilon(t)|R),\, y \in \Sigma(B_\epsilon(t)|a_h) - \Sigma(B_\epsilon(t)|a_{h-1})\}$,
- if $h=n$, $T_n \equiv \{(x,y)|x \notin \Sigma(B_\epsilon(t)|R),\, y \notin \Sigma(B_\epsilon(t)|a_{n-1})\}$,
- if $h<n$, $S_h \equiv \{(x,y)|y \notin \Sigma(B_\epsilon(t)|R),\, x \in \Sigma(B_\epsilon(t)|a_h) - \Sigma(B_\epsilon(t)|a_{h-1})\}$,
- if $h=n$, $S_n \equiv \{(x,y)|y \notin \Sigma(B_\epsilon(t)|R),\, x \notin \Sigma(B_\epsilon(t)|a_{n-1})\}$,

where $a_0=0$, $a_1=1$, and $a_k = 2a_{k-1}$.

Starting from the set T_h, where $\nabla W_R(y)=0$, we obtain

$$\left| \int_D dy \omega_\epsilon(x,t)\omega_\epsilon(y,t)\nabla W_R(x-B_\epsilon(t))K(x-B_\epsilon(t)) \right. \\ \left. + \int_D dy \omega_\epsilon(x,t)\omega_\epsilon(y,t)\nabla W_R(x-B_\epsilon(t))\big[K(x-y)-K(x-B_\epsilon(t))\big] \right|, \quad \text{(A.13)}$$

where $D \equiv \Sigma(B_\epsilon(t)|a_h) - \Sigma(B_\epsilon(t)|a_{h-1})$.

In (A.13) the first term is null because $\nabla W_R(x)\cdot K(x) = 0$. Moreover, we observe that

$$|K(x-y)-K(x)| < \text{const.}\, \frac{\rho}{|x|(|x|-\rho)} \quad \text{if} \quad |y|<\rho<|x|, \quad \text{(A.14)}$$

so the contribution to (A.13) due to T_h is bounded by

$$\text{(A.13)} \le \text{const.}\, \frac{m_t(R)}{R}\left\{ \frac{\epsilon}{R^2} + \sum_{h=2}^{n-1} \frac{a_h \epsilon^2}{R(R-a_h)a_{h-1}^2} \right\} \le \frac{\epsilon}{R^3} m_t(R), \quad \text{(A.15)}$$

where $m_t(R) = 1 - \int \omega_\epsilon(y,t)\,dy$ is the vorticity mass outside $\Sigma(B_\epsilon(t)|R)$.

The contribution due to T_n is simply bounded by using the fact that $\nabla W_R(r)$ removes the singularity of the kernel, because

$$|(\nabla W_R(x) - \nabla W_R(y))K(x-y)| \le \frac{1}{R^2} \quad \text{where} \quad R = |x - B_\epsilon(t)|. \quad \text{(A.16)}$$

Then it is also simple to give an estimate to the second term of (A.12) by using the Lipschitz continuity of $F(x,t)$. Recollecting all the terms we have

$$\left| \frac{d\mu(R)}{dt} \right| \le A(R,\epsilon) m_t(R), \quad \text{(A.17)}$$

where the explicit expression of $A(R,\epsilon)$ comes directly from the previous reasoning by simple calculation. We observe, from the definition of regularized mass, that

$$m_t(R) \le \mu_t\Big(\frac{R}{2}\Big); \tag{A.18}$$

using this inequality in (A.17) and integrating we obtain

$$\mu_t(R) \le \mu_0(R) + A(R)\int_0^t \mu_t\Big(\frac{R}{2}\Big)dt. \tag{A.19}$$

Then we can use an iterative procedure:

$$\begin{aligned}\mu_t(R) &\le \mu_0(R) + A(R)\int_0^t \mu_t\Big(\frac{R}{2}\Big)dt \\ &\le \mu_0(R) + \mu_0\Big(\frac{R}{2}\Big)A(R)\int dt + A(R)A\Big(\frac{R}{2}\Big)\int_0^t dt_1\int_0^{t_1} dt\,\mu_t\Big(\frac{R}{4}\Big),\end{aligned} \tag{A.20}$$

choosing the number n of iterations so that $n\to\infty$ as $\epsilon\to 0$ and $\mu_0(R2^{-n})=0$. We finally have that

$$m_t(R) \le \frac{(\text{const.})^n}{n!} \to 0 \text{ as } \epsilon\to 0 \text{ faster than any power of } \epsilon. \tag{A.21}$$

This means that the vorticity mass becomes very small near the boundary if we take strong concentrations, that is, $\epsilon\to 0$. Then it is also simple to prove that the velocity field generated by the fluid particles near the boundary vanishes for strong concentrations. Finally the main theorem is achieved essentially from these results.

Here we have just recalled the main steps of the proof, leaving to the reader the most part of calculations. As just seen, the property of the solution about the Lipschitz continuity of the field generated by the other $N-1$ vortices is central for the proof of the localization.

References

[Aref 2007] H. Aref, "Point vortex dynamics: a classical mathematics playground", *J. Math. Phys.* **48**:6 (2007), 065401, 23.

[Batchelor 1967] G. K. Batchelor, *An introduction to fluid dynamics*, Cambridge University Press, 1967.

[Boatto 2008] S. Boatto, "Curvature perturbations and stability of a ring of vortices", *Discrete Contin. Dyn. Syst. Ser. B* **10**:2-3 (2008), 349–375.

[Crowdy 2006] D. Crowdy, "Point vortex motion on the surface of a sphere with impenetrable boundaries", *Phys. Fluids* **18**:3 (2006), 036602, 7.

[Hally 1980] D. Hally, "Stability of streets of vortices on surfaces of revolution with a reflection symmetry", *J. Math. Phys.* **21**:1 (1980), 211–217.

[Hwang and Kim 2009] S. Hwang and S.-C. Kim, "Point vortices on hyperbolic sphere", *J. Geom. Phys.* **59**:4 (2009), 475–488.

[Kimura 1999] Y. Kimura, "Vortex motion on surfaces with constant curvature", *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.* **455**:1981 (1999), 245–259.

[Llewellyn Smith 2011] S. G. Llewellyn Smith, "How do singularities move in potential flow?", *Phys. D* **240**:20 (2011), 1644–1651.

[Marchioro 1998] C. Marchioro, "On the localization of the vortices", *Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat.* (8) **1**:3 (1998), 571–584.

[Marchioro and Pulvirenti 1993] C. Marchioro and M. Pulvirenti, "Vortices and localization in Euler flows", *Comm. Math. Phys.* **154**:1 (1993), 49–61.

[Marchioro and Pulvirenti 1994] C. Marchioro and M. Pulvirenti, *Mathematical theory of incompressible nonviscous fluids*, Appl. Math. Sciences **96**, Springer, New York, 1994.

[Newton 2001] P. K. Newton, *The N-vortex problem: analytical techniques*, Appl. Math. Sciences **145**, Springer, New York, 2001.

Received 27 Feb 2012. Revised 14 Mar 2012. Accepted 17 Apr 2012.

ROBERTO GARRA: rolinipame@yahoo.it
Dipartimento di Scienze di Base e Applicate per l'Ingegneria, Università degli Studi di Roma 'La Sapienza', via Antonio Scarpa 16, I-00161 Roma, Italy

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 1, No. 1, 2013
dx.doi.org/10.2140/memocs.2013.1.33

CONTRACTION OF THE PROXIMAL MAP AND GENERALIZED CONVEXITY OF THE MOREAU–YOSIDA REGULARIZATION IN THE 2-WASSERSTEIN METRIC

ERIC A. CARLEN AND KATY CRAIG

We investigate the Moreau–Yosida regularization and the associated proximal map in the context of discrete gradient flow for the 2-Wasserstein metric. Our main results are a stepwise contraction property for the proximal map and an "above the tangent line" inequality for the regularization. Using the latter, we prove a Talagrand inequality and an HWI inequality for the regularization, under appropriate hypotheses. In the final section, the results are applied to study the discrete gradient flow for Rényi entropies. As Otto showed, the gradient flow for these entropies in the 2-Wasserstein metric is a porous medium flow or a fast diffusion flow, depending on the exponent of the entropy. We show that a striking number of the remarkable features of the porous medium and fast diffusion flows are present in the discrete gradient flow and do not simply emerge in the limit as the time-step goes to zero.

1. Introduction

Given a complete metric space (X, d), a functional $E : X \to \mathbb{R} \cup \{\infty\}$, and $\tau > 0$, the *Moreau–Yosida regularization* of E is

$$E_\tau(y) := \inf_{x \in X} \left\{ \frac{1}{2\tau} d(x, y)^2 + E(x) \right\}.$$

The corresponding *proximal set* $J_\tau : X \to 2^X$ is

$$J_\tau(y) := \operatorname*{argmin}_{x \in X} \left\{ \frac{1}{2\tau} d(x, y)^2 + E(x) \right\}.$$

If there is a unique element in $J_\tau(y)$, we denote it by y_τ and call it the *proximal point*. We call $y \mapsto y_\tau$ the *proximal map*.

Communicated by Raffaele Esposito.
Carlen's work was partially supported by U.S. National Science Foundation grant DMS 0901632.
Craig's work was partially supported by a Presidential Fellowship at Rutgers University.

MSC2010: 49-XX.
Keywords: Wasserstein metric, gradient flow, Moreau–Yosida regularization.

When $X = \mathcal{H}$ is a Hilbert space, a suitable context in which to develop the theory of the Moreau–Yosida regularization is the class of functionals that are proper, lower semicontinuous, and convex. For all such E and $\tau > 0$, the Moreau–Yosida regularization E_τ is convex and Fréchet differentiable [Moreau 1965]. Furthermore, its derivative is Lipschitz continuous, and, as $\tau \to 0$, $E_\tau \nearrow E$ pointwise [Brézis 1973]. The Moreau–Yosida regularization provides a way to regularize E that preserves convexity.

The proximal map is similarly well-behaved for functionals that are proper, lower semicontinuous, and convex. For each $y \in \mathcal{H}$ and $\tau > 0$, there is a unique proximal point y_τ, so that the proximal map $y \mapsto y_\tau$ is well-defined on all of $\mathcal{H}$. As shown in [Moreau 1965], the proximal map is a contraction in the Hilbert space norm:

$$\|x_\tau - y_\tau\| \le \|x - y\| \quad \text{for all } x, y \in \mathcal{H}.$$

One of the main reasons for interest in the Moreau–Yosida regularization and proximal map is their relation to gradient flow. The *gradient flow* of a functional E is the Cauchy problem

$$\frac{d}{dt}y(t) = -\nabla E(y(t)), \quad y(0) \in \overline{D(E)} = \overline{\{z \in \mathcal{H} : E(z) < \infty\}}, \tag{1-1}$$

which is well-defined as long as ∇E exists along the flow $y(t)$.[1] The Moreau–Yosida regularization plays a key role in the proof of existence for solutions to the gradient flow [Brézis 1971]. First, one uses the additional regularity of E_τ to find solutions to the related gradient flow problem

$$\frac{d}{dt}y_\tau(t) = -\nabla E_\tau(y_\tau(t)), \quad y_\tau(0) \in \overline{D(E)}.$$

Then, as $\tau \to 0$, the curves $y_\tau(t)$ converge to a curve $y(t)$ that solves (1-1) in an appropriate sense.

The proximal map expresses the discrete dynamics of gradient flow. Specifically, one may use the proximal map to define the *discrete gradient flow* sequence

$$y_n = (y_{n-1})_\tau, \quad y_0 \in \overline{D(E)},$$

as in [Martinet 1970; 1972]. Whenever the proximal map $y \mapsto y_\tau$ is well-defined, we may identify the proximal set $J_\tau(y)$ with its unique element y_τ and write J_τ^n to indicate n repeated applications of the proximal map. The exponential formula quantifies the sense in which the discrete gradient flow is a discretized version of gradient flow [Brézis 1973]. If $y(t)$ is a gradient flow with initial conditions $y(0)$, then

$$y(t) = \lim_{n\to\infty} (J_{t/n})^n(y(0)). \tag{1-2}$$

[1] Alternatively, one may define the gradient flow in terms of the subdifferential [Brézis 1971].

More recently, the Moreau–Yosida regularization and proximal map have been applied outside of the Hilbert space context to gradient flow in the 2-Wasserstein metric. Briefly, we recall some facts about this metric, mainly to establish our notation; see [Ambrosio et al. 2008] and [Villani 2003] for more background. We present these facts both in the most general setting, without restrictions on the type of probability measures we consider, and in a simpler setting, focusing our attention on probability measures with finite second moment that are absolutely continuous with respect to Lebesgue measure. While our results hold in the most general setting, many interesting applications concern only the simpler setting, in which the exposition and notation is more straightforward.

Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of Borel probability measures on $\mathbb{R}^d$. Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, a Borel map $T : \mathbb{R}^d \to \mathbb{R}^d$ *transports* μ *onto* ν if $\nu(B) = \mu(T^{-1}(B))$ for all Borel sets $B \subseteq \mathbb{R}^d$. We call ν the *push-forward of* μ *under* T and write $\nu = T\#\mu$.

Now consider a measure $\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. (We will distinguish probability measures on $\mathbb{R}^d \times \mathbb{R}^d$, from probability measures on $\mathbb{R}^d$ by writing them in bold font.) Let π_1 be the projection onto the first component of $\mathbb{R}^d \times \mathbb{R}^d$, and let π_2 be the projection onto the second component. The first and second *marginals* of $\boldsymbol{\mu}$ are $\pi_1\#\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d)$ and $\pi_2\#\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d)$.

Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the set of *transport plans* from μ to ν is

$$\Gamma(\mu, \nu) := \{\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi_1\#\boldsymbol{\mu} = \mu, \pi_2\#\boldsymbol{\mu} = \nu\}.$$

The *2-Wasserstein distance* between μ and ν is

$$W_2(\mu, \nu) := \left(\inf\left\{\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 d\boldsymbol{\mu}(x, y) : \boldsymbol{\mu} \in \Gamma(\mu, \nu)\right\}\right)^{1/2}. \tag{1-3}$$

When $W_2(\mu, \nu) < \infty$, this infimum is attained, and we refer to the plans that attain the infimum as *optimal transport plans*. We denote the set of optimal transport plans by $\Gamma_0(\mu, \nu)$.

The 2-Wasserstein distance satisfies the triangle inequality and is non-negative, non-degenerate, and symmetric. However, $\mathcal{P}(\mathbb{R}^d)$ endowed with the 2-Wasserstein distance is not a metric space, since there exist measures that are infinite distances apart. Let $\mathcal{P}_{\mu_0}(\mathbb{R}^d)$ be the subset of $\mathcal{P}(\mathbb{R}^d)$ consisting of measures that are a finite distance from some fixed Borel probability measure μ_0, so that, by the triangle inequality, $(\mathcal{P}_{\mu_0}(\mathbb{R}^d), W_2)$ is a metric space. As indicated by the notation, one may take μ_0 to be the initial conditions of a gradient flow. Note that when $\mu_0 = \delta_0$, the Dirac mass at the origin, $\mathcal{P}_{\delta_0}(\mathbb{R}^d)$ is the subset of $\mathcal{P}(\mathbb{R}^d)$ with finite second moment.

We now define the 2-Wasserstein distance in a simpler setting. Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the set of probability measures with finite second moment and $\mathcal{P}_2^a(\mathbb{R}^d)$ denote the set of probability measures with finite second moment that are absolutely

continuous with respect to Lebesgue measure. If $\mu \in \mathcal{P}_2^a(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance between μ and ν reduces to the form

$$W_2(\mu, \nu) := \left(\inf \left\{ \int |x - T(x)|^2 d\mu(x) : T\#\mu = \nu \right\} \right)^{1/2}. \tag{1-4}$$

The Brenier–McCann theorem guarantees that the infimum in (1-4) is attained by $T = \nabla\varphi$, where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is convex and $\nabla\varphi$ is unique μ-almost everywhere [McCann 1995]. In particular,

$$W_2^2(\mu, \nu) = \int |x - \nabla\varphi(x)|^2 d\mu(x),$$

and we call $\nabla\varphi$ the *optimal transport map from μ to ν*. To emphasize its dependence on μ and ν, we denote the optimal transport map from μ to ν by $\boldsymbol{t}_\mu^\nu$.

Given $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$ with $W_2^2(\mu_1, \mu_2) < \infty$ and $\boldsymbol{\mu} \in \Gamma_0(\mu^1, \mu^2)$, a *geodesic* connecting μ_1 and $\mu_2 \in \mathcal{P}(\mathbb{R}^d)$ is a curve of the form

$$\mu_\alpha^{1\to 2} : [0, 1] \to \mathcal{P}(\mathbb{R}^d), \quad \mu_\alpha^{1\to 2} = ((1-\alpha)\pi_1 + \alpha\pi_2)\#\boldsymbol{\mu}.$$

As shown in [Ambrosio et al. 2008, Theorem 7.2.2], this definition agrees with the metric space definition of a geodesic, i.e., a curve $\mu_\alpha : [0, 1] \to \mathcal{P}(\mathbb{R}^d)$ with $W_2(\mu_0, \mu_1) < \infty$ such that $W_2(\mu_\alpha, \mu_\beta) = |\alpha - \beta| W_2(\mu_0, \mu_1)$. If $\mu_1 \in \mathcal{P}_2^a(\mathbb{R}^d)$, $\mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, then the geodesic connecting μ_1 and μ_2 is unique and of the form

$$\mu_\alpha^{1\to 2} : [0, 1] \to \mathcal{P}_2(\mathbb{R}^d), \quad \mu_\alpha^{1\to 2} = \left((1-\alpha)\mathrm{id} + \alpha \boldsymbol{t}_{\mu_1}^{\mu_2}\right)\#\mu_1,$$

where $\mathrm{id}(x) = x$ is the identity transformation.

A functional $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ is λ-*convex* in the 2-Wasserstein metric if, for all $\mu_1, \mu_2 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$, there exists a geodesic connecting μ_1 and μ_2 along which E is λ-convex:

$$E(\mu_\alpha^{1\to 2}) \le (1-\alpha)E(\mu_1) + \alpha E(\mu_2) - \alpha(1-\alpha)\frac{\lambda}{2} W_2^2(\mu_1, \mu_2). \tag{1-5}$$

If a functional is 0-convex, we simply call it *convex*.[2] If a functional is 0-convex and strict inequality holds in (1-5) for all $\alpha \in (0, 1)$, we call it *strictly convex*.

Given a functional $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ and $\tau > 0$, its Moreau–Yosida regularization is

$$E_\tau(\mu) := \inf_{\nu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} W_2^2(\mu, \nu) + E(\nu) \right\} \tag{1-6}$$

[2]It is also common to refer to convex functionals in the 2-Wasserstein metric as *displacement convex* [McCann 1997].

and the corresponding *proximal set* $J_\tau : \mathscr{P}_{\mu_0}(\mathbb{R}^d) \to 2^{\mathscr{P}_{\mu_0}(\mathbb{R}^d)}$ is

$$J_\tau(\mu) := \operatorname*{argmin}_{\nu \in \mathscr{P}_{\mu_0}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} W_2^2(\mu, \nu) + E(\nu) \right\}. \tag{1-7}$$

As before, if there is a unique element in $J_\tau(\mu)$, we denote it by μ_τ and call it the *proximal point*. Similarly, we call $\mu \mapsto \mu_\tau$ the *proximal map*. The properties of the Moreau–Yosida regularization and proximal map in the 2-Wasserstein metric will be the main focus of this paper.

As in the Hilbertian case, one of the main reasons for interest in the Moreau–Yosida regularization and the proximal map in the 2-Wasserstein metric is their relation to gradient flow. When E and μ are sufficiently smooth, the *2-Wasserstein gradient of E at $\mu \in D(E)$* is

$$\nabla_W E(\mu) = -\nabla \cdot \left(\mu \nabla \frac{\delta E}{\delta \rho}(\mu) \right), \tag{1-8}$$

where $\delta E/\delta\rho$ is the functional derivative of E [Otto 2001; Ambrosio et al. 2008, Chapters 8 and 10].[3] The *gradient flow* of E is the Cauchy problem

$$\frac{d}{dt}\mu(t) = -\nabla_W E(\mu(t)), \quad \mu(0) \in \overline{D(E)} = \overline{\{\mu \in P_{\mu_0}(\mathbb{R}^d) : E(\mu) < \infty\}},$$

which is well-defined as long as $\nabla_W E(\mu(t))$ exists along the flow $\mu(t)$.[4] We will sometimes refer to this as the *continuous gradient flow* in order to distinguish it from the *discrete gradient flow* we define below.

Otto [1996; 2001] observed that the right-hand side of (1-8) may be viewed as the gradient vector field on the "Riemannian manifold of probability densities on $\mathbb{R}^d$" associated to the functional E, where the Riemannian metric is the infinitesimal form of the 2-Wasserstein metric. (It is one of his insights that the 2-Wasserstein metric is induced by a Riemannian metric.) In this metric, the length of the gradient of E at μ is given by

$$|\nabla_W E(\mu)| = \left(\int \left| \nabla \frac{\delta E}{\delta \rho}(\mu) \right|^2 d\mu \right)^{1/2}. \tag{1-9}$$

As in the Hilbertian case, the proximal map expresses the dynamics for discrete gradient flow. When the proximal map $\mu \mapsto \mu_\tau$ is well-defined (which occurs under

[3] Some authors, including Ambrosio et al., identify the tangent vector $\nabla_W E(\mu)$ with the gradient vector field $-\nabla \frac{\delta E}{\delta \rho}(\mu)$ on $\mathbb{R}^d$. One gets Otto's representative from this by multiplying by μ and taking the divergence. The choice of representatives is merely notational.

[4] Alternatively, one may define the gradient flow in terms of the subdifferential [Ambrosio et al. 2008, Definition 11.1.1].

much weaker assumptions on E and μ than are needed to define the gradient, as we describe before (1-14) below) we may define the *discrete gradient flow* sequence

$$\mu_n = (\mu_{n-1})_\tau, \quad \mu_0 \in \overline{D(E)}. \tag{1-10}$$

As before, we identify the proximal set $J_\tau(\mu)$ with its unique element μ_τ and write J_τ^n to indicate n repeated applications of the proximal map.

One of the advantages of discrete gradient flow is that it is not necessary to make precise the sense in which (1-8) defines a gradient vector field. This fact was emphasized by De Giorgi [1993] in his theory of the metric derivative and extensively developed by Ambrosio, Gigli, and Savaré [Ambrosio et al. 2008, Chapter 8]. We follow De Giorgi's lead, and all of the estimates we use involve only the length of the gradient $|\nabla_W E(\mu)|$. In the case that E and μ lack sufficient smoothness for (1-9) to be well-defined, we will interpret the symbol $|\nabla_W E(\mu)|$ as the *metric slope*

$$\limsup_{\nu\to\mu} \frac{(E(\mu) - E(\nu))^+}{W_2(\mu, \nu)}. \tag{1-11}$$

We use the heuristic notation $|\nabla_W E(\mu)|$ since, as demonstrated by Otto [1996; 2001], it is often enlightening to think of $|\nabla_W E(\mu)|$ as coming from a Riemannian metric on $\mathcal{P}(\mathbb{R}^d)$.

The book [Ambrosio et al. 2008] contains a detailed study of gradient flow and discrete gradient flow in the 2-Wasserstein metric for large classes of functionals, developing the analogy with the Hilbert space theory. It would be too much to hope for a perfect analogy. For example, in the Hilbert space context, if a functional E is proper, lower semicontinuous, and convex, then its Moreau–Yosida regularization E_τ is also convex. However, in the 2-Wasserstein metric, it is well-known that even when E satisfies analogous assumptions, E_τ is not always convex.[5] The key technical difference between the two metrics is that while

$$x \mapsto \tfrac{1}{2}\|x - y\|^2 \tag{1-12}$$

is 1-convex along geodesics,

$$\mu \mapsto \tfrac{1}{2} W_2^2(\mu, \nu) \tag{1-13}$$

is not λ-convex along geodesics, for any $\lambda \in \mathbb{R}$, if the dimension of the underlying space is greater than or equal to 2 [ibid., Example 9.1.5]. Since much of De Giorgi's "minimizing steps" approach to gradient flow relies on the 1-convexity of (1-12), this lack of convexity in the 2-Wasserstein case complicates the implementation of De Giorgi's scheme.

[5] For the reader's convenience, we include an example in Section 3.

Ambrosio et al. circumvent this difficulty with their observation that, though $\mu \mapsto \frac{1}{2}W_2^2(\mu, \nu)$ is not 1-convex along all geodesics, it is 1-convex along a different class of curves. They define the set of *generalized geodesics* to be the union of these classes of curves over all $\nu \in \mathcal{P}(\mathbb{R}^d)$ (see Section 2A). By considering functionals that are *convex along generalized geodesics* — a stronger condition than merely being convex along geodesics (see Section 2B) — they deduce a priori estimates that provide detailed control over the gradient flow and discrete gradient flow.

The key results that we will use concern functionals $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ that are proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics (see Section 2B).[6] With these assumptions, Ambrosio, Gigli, and Savaré show that if $\tau > 0$ is small enough so that $\lambda\tau > -1$, then for all $\mu \in \overline{D(E)}$ the proximal map

$$\mu \mapsto \mu_\tau \tag{1-14}$$

and the *discrete gradient flow* sequence

$$\mu_n = (\mu_{n-1})_\tau, \quad \mu_0 \in \overline{D(E)},$$

are well-defined. They go on to prove the 2-Wasserstein analogue of the exponential formula (1-2) relating the discrete gradient flow to the continuous gradient flow [Ambrosio et al. 2008, Theorem 4.0.4]. Specifically, they show that if $\mu(t)$ is the solution to the continuous gradient flow of E with initial conditions $\mu(0) \in \overline{D(E)}$, then

$$\mu(t) = \lim_{n\to\infty} (J_{t/n})^n(\mu(0)). \tag{1-15}$$

Using the assumption of convexity along generalized geodesics, Ambrosio, et al. comprehensively develop the theory of continuous gradient flow. While this assumption is stronger than (standard) convexity along geodesics, it is not restrictive: all important examples of functionals that are convex along geodesics are also convex along generalized geodesics [Ambrosio et al. 2008, Section 9.3].

In this paper, we take a closer look at the Moreau–Yosida regularization and the proximal map in the 2-Wasserstein metric for functionals that are convex along generalized geodesics. We show that, while the Moreau–Yosida regularization does not preserve E's convexity along all geodesics (as in the Hilbertian case), if E attains its minimum at $\bar{\mu}$, the Moreau–Yosida regularization does satisfy an "above the tangent line" inequality at $\bar{\mu}$. This type of inequality is a necessary condition for convexity — in particular, a function from $\mathbb{R}$ to $\mathbb{R}$ is convex if and only if it lies above its tangent line at every point.

[6] The results in [Ambrosio et al. 2008] are often stated in the context when $\mu_0 = \delta_0$, the Dirac mass at the origin, so $\mathcal{P}_{\mu_0}(\mathbb{R}^d) = \mathcal{P}_2(\mathbb{R}^d)$. We quote these results in broader generality, since the proofs are easily adapted to this case.

Theorem 1.1 (generalized convexity of E_τ). *Given $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics with $\lambda \geq 0$, assume that E attains its minimum at $\bar{\mu}$. For $\tau > 0$, define*

$$\lambda_\tau := \frac{\lambda}{1 + \lambda\tau}.$$

Then, for all $\mu \in \overline{D(E)}$, there exists a geodesic $\mu_\alpha^{\bar{\mu}\to\mu}$ from $\bar{\mu}$ to μ such that

$$E_\tau(\mu_\alpha^{\bar{\mu}\to\mu}) \leq (1-\alpha)E_\tau(\bar{\mu}) + \alpha E_\tau(\mu) - \alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\bar{\mu}, \mu). \tag{1-16}$$

In Section 4A, we show that (1-16) is sharp by presenting an example in which E is λ-convex and E_τ is no more than λ_τ-convex

As a consequence of Theorem 1.1, we show E_τ satisfies a Talagrand inequality and an HWI inequality.

Theorem 1.2 (Talagrand and HWI inequalities). *Under the assumptions of the Theorem 1.1, for all $\mu \in \overline{D(E)}$, we have the Talagrand inequality*

$$E_\tau(\mu) - E_\tau(\bar{\mu}) \geq \frac{\lambda_\tau}{2}W_2^2(\mu, \bar{\mu}) \tag{1-17}$$

and the HWI inequality

$$E_\tau(\mu) - E_\tau(\bar{\mu}) \leq |\nabla_W E_\tau(\mu)| W_2(\mu, \bar{\mu}) - \frac{\lambda_\tau}{2}W_2^2(\mu, \bar{\mu}). \tag{1-18}$$

These inequalities capture E_τ's behavior at $\bar{\mu}$ from both ends of the "above the tangent line" inequality.

We also develop the analogy between Hilbertian metrics and the 2-Wasserstein metric by proving a contraction inequality for the proximal map. In a Hilbert space, if E is proper, lower semicontinuous, and convex, Moreau [1965] showed that the proximal map satisfies

$$\|x_\tau - y_\tau\| \leq \|x - y\| \quad \text{for all } x, y \in \mathcal{H}. \tag{1-19}$$

This turns out to be a rather miraculous property of the Hilbertian norm that fails even in simple Banach spaces. For example, consider the ℓ^∞ norm on $\mathbb{R}^2$. Fix two points $a = (0, 0)$ and $b = (1, 1)$, and let K be the closed half-space lying beneath the line $3x_2 = x_1 - 4$. Let E be the indicator function for K,

$$E(x) := \begin{cases} 0 & \text{if } x = (x_1, x_2) \in K, \\ \infty & \text{otherwise.} \end{cases}$$

Then

$$J_\tau(y) := \operatorname*{argmin}_{x\in\mathbb{R}^2} \left\{ \frac{1}{2\tau}\|x - y\|_\infty^2 + E(x) \right\} = \operatorname*{argmin}_{x\in K} \left\{ \frac{1}{2\tau}\|x - y\|_\infty^2 \right\}.$$

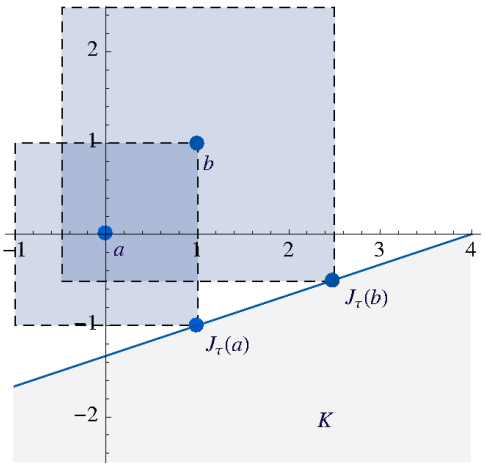


Figure 1. In the Banach space $\mathbb{R}^2$, endowed with the ℓ^∞ norm, the proximal map is not a contraction.

Therefore, $J_\tau(a) = (1, -1)$ and $J_\tau(b) = (\frac{5}{2}, -\frac{1}{2})$ for all $\tau > 0$. This is not a contraction since $\|a - b\|_\infty = 1 < \frac{3}{2} = \|J_\tau(a) - J_\tau(b)\|_\infty$ (see figure).

The situation for general metric spaces is even more involved than the situation for metrics induced by norms, and one does not expect a contraction to hold. Nevertheless, if E is appropriately convex, the continuous-time gradient flow defined by (1-15) is contractive [Otto 2001; Ambrosio et al. 2008, Theorem 4.0.4]. This gives hope that some contraction property of the proximal map is present at the discrete level and does not merely emerge in the limit.

Our next result shows that this is the case. In particular, we achieve contraction of the proximal map by making a small modification to the squared distance: given $\tau > 0$, we consider the functional $\Lambda_\tau : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ defined by

$$\Lambda_\tau(\mu, \nu) := W_2^2(\mu, \nu) + \frac{\tau^2}{2}|\nabla_W E(\mu)|^2 + \frac{\tau^2}{2}|\nabla_W E(\nu)|^2. \tag{1-20}$$

As before, we interpret $|\nabla_W E(\mu)|$ as the metric slope (1-11) when E and μ lack sufficient smoothness for the norm of the 2-Wasserstein gradient (1-9) to be well-defined.

Though we state the following theorem in the context of the 2-Wasserstein metric, it continues to hold in a more abstract setting: given a functional E on a complete metric space (X, d), if E is proper, coercive, lower semicontinuous, and satisfies [Ambrosio et al. 2008, Assumption 4.0.1] for some $\lambda \in \mathbb{R}$, then the result remains true by replacing W_2 with d.

Theorem 1.3 (contraction of proximal map). *Given $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics,*

fix $\tau > 0$ small enough so that $\lambda\tau > -1$. Consider $\mu, \nu \in \overline{D(E)}$ and let $\Lambda_\tau : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ be given by (1-20)*. Then, if $\lambda \geq 0$, the proximal map is contracting in Λ_τ,*

$$\Lambda_\tau(\mu_\tau, \nu_\tau) \leq \Lambda_\tau(\mu, \nu). \tag{1-21}$$

More generally, for $\lambda \in \mathbb{R}$,

$$\begin{aligned}&\Lambda_\tau(\mu_\tau, \nu_\tau) - \Lambda_\tau(\mu, \nu)\\ &\leq -\tfrac{1}{2}(\tau|\nabla_W E(\nu)| - W_2(\nu, \nu_\tau))^2 - \tfrac{1}{2}(\tau|\nabla_W E(\mu)| - W_2(\mu, \mu_\tau))^2\\ &\quad -\tfrac{1}{2}\lambda\tau\big[2W_2^2(\mu_\tau, \nu_\tau)+W_2^2(\mu, \nu_\tau)+W_2^2(\nu, \mu_\tau)+W_2^2(\nu, \nu_\tau)+W_2^2(\mu, \mu_\tau)\big].\end{aligned} \tag{1-22}$$

In Section 4A, we show that the inequality (1-22) is sharp. Then, in Section 4B, we apply (1-21) together with scaling properties of the W_2 metric to derive sharp polynomial rates of convergence to Barenblatt profiles for certain fast diffusion and porous medium equations. Otto originally deduced these results in [Otto 2001] by considering a modified gradient flow problem for λ-convex functionals with $\lambda > 0$. The contraction inequality (1-21) provides a simple route to such results. The fast diffusion and porous media equations also provide examples of strictly convex functionals for which the proximal map is strictly contracting in Λ_τ but not in W_2.

Remark 1.4. While [Ambrosio et al. 2008] does not explicitly consider monotonicity results for modifications of the squared distance along the discrete gradient flow, such a result (for a different modification) can be found by reading between the lines in Lemma 4.2.4 of that reference. Consider the alternative modification to the squared distance function defined by

$$\widetilde{\Lambda}_\tau(\mu, \nu) := W_2^2(\mu, \nu) + \tau E(\mu) + \tau E(\nu). \tag{1-23}$$

If one takes the final inequality on page 92 of [Ambrosio et al. 2008] for $\lambda = 0$ and $n = 1$, rearranges terms, and symmetrizes in μ and ν, one obtains (1-21) with $\widetilde{\Lambda}_\tau$ in place of Λ_τ. A key difference between $\widetilde{\Lambda}_\tau$ and our functional Λ_τ is that, for measures μ and ν with $|\nabla_W E(\mu)|$ and $|\nabla_W E(\mu)| < \infty$, Λ_τ involves only an $\mathcal{O}(\tau^2)$ correction to $W_2^2(\mu, \nu)$, while $\widetilde{\Lambda}_\tau$ involves an $\mathcal{O}(\tau)$ correction to $W_2^2(\mu, \nu)$.

Remark 1.5. While one might first suppose that Λ_τ could only be used to study discrete gradient flows with initial data μ, ν satisfying $|\nabla_W E(\mu)|$, $|\nabla_W E(\mu)| < \infty$, when E is strictly convex, the discrete gradient flow produces this regularity in one step (see Lemma 2.2). We shall see an example of this in Section 4B when we apply Theorem 1.3 to the discrete gradient flow for the Rényi entropies.

For $\lambda > 0$, one can extract from (1-22) a useful inequality that implies, among other things, an optimal exponential rate of decrease of $\Lambda_\tau(\mu, \bar{\mu})$ when E has a minimizer $\bar{\mu}$ (necessarily unique due to the strict convexity).

Corollary 1.6 (the case $\lambda > 0$). *Consider $\lambda > 0$ and $\tau > 0$ sufficiently small so that $\tau\lambda \le 1$. Then for all E satisfying the hypotheses of Theorem* 1.3 *and $\mu, \nu \in \overline{D(E)}$,*

$$(1+\tau\lambda)\Lambda_\tau(\mu_\tau, \nu_\tau) \le (1-\tau\lambda)\Lambda_\tau(\mu, \nu) + 3\lambda\tau\Lambda_\tau^{1/2}(\mu, \nu)[W_2(\mu, \mu_\tau) + W_2(\nu, \nu_\tau)]. \quad (1\text{-}24)$$

We give the proof of this corollary in Section 3. However, to explain its consequences, we state and prove a simple discrete Gronwall-type inequality. It is a discrete version of the continuous-time inequality [Ambrosio et al. 2008, Lemma 4.1.8]. (See [Baiocchi 1989; Emmrich 1999] for related discrete Gronwall inequalities.)

Lemma 1.7 (a discrete Gronwall-type inequality). *Let $\lambda, \tau > 0$, and let $\{a_n\}$ and $\{b_n\}$ be two sequences of non-negative numbers such that for all $n \ge 0$,*

$$(1+\tau\lambda)a_n \le (1-\tau\lambda)a_{n-1} + \tau a_{n-1}^{1/2} b_n. \quad (1\text{-}25)$$

Then,

$$a_n^{1/2} \le (1+\lambda\tau)^{-n} a_0^{1/2} + \sqrt{\frac{\tau}{2\lambda}}(1+\lambda\tau)\left(\sum_{k=1}^{n} b_k^2\right)^{1/2}.$$

Consider the discrete gradient flow of E starting from $\mu \in D(E)$ with $\tau > 0$ and $\tau\lambda \le 1$. Let $\mu_0 := \mu$ and inductively define $\{\mu_n\}$ by repeated application of the proximal map. Define $\{\nu_n\}$ in the same way, starting from $\nu \in D(E)$. Now, apply Lemma 1.7 and Corollary 1.6 to these discrete gradient flows of E, taking

$$a_n := \Lambda_\tau(\mu_n, \nu_n) \quad \text{and} \quad b_n := 3\lambda\sqrt{2W_2^2(\mu_{n-1}, \mu_n) + 2W_2^2(\nu_{n-1}, \nu_n)}\,.$$

Since

$$W_2^2(\mu, \mu_\tau) \le 2\tau[E(\mu) - E(\mu_\tau)], \quad (1\text{-}26)$$

$\sum_{k=1}^n b_k^2$ is bounded by a telescoping sum:

$$\sum_{k=1}^{n} b_k^2 \le \tau 36\lambda^2[(E(\mu) - E(\mu_n)) + (E(\nu) - E(\nu_n))].$$

In case E is bounded below, we may assume without loss of generality that E is non-negative. Then,

$$\Lambda_\tau^{1/2}(\mu_n, \nu_n) \le (1+\lambda\tau)^{-n}\Lambda_\tau^{1/2}(\mu, \nu) + \lambda\tau\frac{6(1+\lambda\tau)}{\sqrt{2\lambda}}\sqrt{E(\mu) + E(\nu)}. \quad (1\text{-}27)$$

Thus, for positive λ and sufficiently small τ, $\Lambda_\tau^{1/2}(\mu_n, \nu_n)$ decays "exponentially fast" at rate λ up to the time that this quantity becomes $\mathbb{O}(\tau)$.[7]

[7] At this point, we may use the bound $E(\mu_n) \le (1+\lambda\tau)^{-2n}E(\mu)$ [Ambrosio et al. 2008, Theorem 3.1.6] and apply (1-27) iteratively.

The proof of Lemma 1.7 is elementary, so we provide it here, closing this section.

Proof of Lemma 1.7. Multiply both sides of (1-25) by $(1+\tau\lambda)^{2n-1}$ to obtain

$$(1+\tau\lambda)^{2n}a_n \le (1-(\tau\lambda)^2)(1-\tau\lambda)^{2n-2}a_{n-1} + \tau\left((1+\tau\lambda)^{2n-2}a_{n-1}\right)^{1/2}(1+\tau\lambda)^n b_n.$$

Defining

$$\tilde{a}_n := (1+\tau\lambda)^{2n}a_n \qquad \text{and} \qquad \tilde{b}_n := \tau(1+\tau\lambda)^n b_n,$$

we have $\tilde{a}_n \le \tilde{a}_{n-1} + \tilde{a}_{n-1}^{1/2}\tilde{b}_n$, and therefore $\tilde{a}_n \le a_0 + \sum_{k=1}^{n} \tilde{a}_{k-1}^{1/2}\tilde{b}_k$. Defining

$$c_n := \max\{\tilde{a}_k : 0 \le k \le n\},$$

we have $c_n \le a_0 + c_n^{1/2}\sum_{k=1}^{n}\tilde{b}_k$. This quadratic inequality implies that $c_n^{1/2} \le a_0^{1/2} + \sum_{k=1}^{n}\tilde{b}_k$. By the Cauchy–Schwarz inequality and the fact that

$$\sum_{k=1}^{n}\alpha^k \le \frac{\alpha}{\alpha-1}\alpha^n \quad \text{for } \alpha := (1+\lambda\tau)^2 \ge 1,$$

we have

$$\sum_{k=1}^{n}\tilde{b}_k \le \frac{\sqrt{\tau}(1+\lambda\tau)^{n+1}}{\sqrt{2\lambda}}\left(\sum_{k=1}^{n} b_k^2\right)^{1/2}. \qquad \square$$

2. Generalized convexity and the proximal map

2A. ***Generalized geodesics.*** In a Hilbert space, $x \mapsto \frac{1}{2}\|x-y\|^2$ is 1-convex along geodesics. However, the same is not true for the squared 2-Wasserstein distance when the dimension of the underlying space exceeds 1, as pointed out by Ambrosio et al. [2008, Example 9.1.5]. Instead, these authors observe that $\mu \mapsto \frac{1}{2}W_2^2(\mu,\nu)$ is convex along a different set of curves, which we now describe.

Fix $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$ with optimal plans $\boldsymbol{\mu}_{1,2} \in \Gamma_0(\mu_1,\mu_2)$, $\boldsymbol{\mu}_{1,3} \in \Gamma_0(\mu_1,\mu_3)$. For $1 \le i < j \le 3$, let $\pi_{i,j}$ be the projection onto the i-th and j-th components of $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$. Fix $\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ so that $\pi_{1,2}\#\boldsymbol{\mu} = \boldsymbol{\mu}_{1,2}$ and $\pi_{1,3}\#\boldsymbol{\mu} = \boldsymbol{\mu}_{1,3}$ [ibid., Lemma 5.3.2]. (We use bold font to distinguish probability measures on $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ or $\mathbb{R}^d \times \mathbb{R}^d$ from probability measures on $\mathbb{R}^d$.) As in [ibid., Definition 9.2.2], a *generalized geodesic joining μ_2 to μ_3 with base μ_1* is a curve of the form

$$\mu_\alpha^{2\to 3} : [0,1] \to \mathcal{P}(\mathbb{R}^d), \quad \mu_\alpha^{2\to 3} := ((1-\alpha)\pi_2 + \alpha\pi_3)\#\boldsymbol{\mu}.$$

In the case $\mu_1 \in \mathcal{P}_2^a(\mathbb{R}^d)$ and $\mu_2, \mu_3 \in \mathcal{P}_2(\mathbb{R}^d)$, this reduces to

$$\mu_\alpha^{2\to 3} : [0,1] \to \mathcal{P}(\mathbb{R}^d), \quad \mu_\alpha^{2\to 3} = \left((1-\alpha)\boldsymbol{t}_{\mu_1}^{\mu_2} + \alpha\boldsymbol{t}_{\mu_1}^{\mu_3}\right)\#\mu_1.$$

Ambrosio, Gigli, and Savaré demonstrate that $\mu \mapsto \frac{1}{2}W_2^2(\mu, \mu_1)$ is 1-convex along any generalized geodesic $\mu_\alpha^{2\to 3}$ with base μ_1, for all $\mu^2, \mu^3 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$ [ibid., Lemma 9.2.1]. Note that if the base μ_1 equals either μ_2 or μ_3, $\mu_\alpha^{2\to 3}$ is a (standard) geodesic joining μ_2 and μ_3. Thus, while $\mu \to \frac{1}{2}W_2^2(\mu, \mu_1)$ is not convex along geodesics (in the sense that it is not convex along *all* geodesics), it is convex along *some* geodesics.

2B. ***The functionals*** $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$**.** Fix a Borel probability measure μ_0. We consider functionals $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ that satisfy the following conditions:

- *proper:* $D(E) := \{\mu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d) : E(\mu) < \infty\} \neq \varnothing$.
- *coercive*:[8] There exists $\tau^* > 0$ such that for all $0 < \tau < \tau^*$, $\mu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$,

$$E_\tau(\mu) = \inf_{\nu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} W_2^2(\mu, \nu) + E(\nu) \right\} > -\infty.$$

  As noted in [Ambrosio et al. 2008, Lemma 2.2.1], by a triangle inequality argument, it is enough to check that there exists $\tau_0 > 0$ such that

$$E_{\tau_0}(\mu_0) = \inf_{\nu \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)} \left\{ \frac{1}{2\tau_0} W_2^2(\mu_0, \nu) + E(\nu) \right\} > -\infty. \tag{2-1}$$

- *lower semicontinuous:* For all $\mu_n, \mu \in P_{\mu_0}(\mathbb{R}^d)$ such that $\mu_n \to \mu$ in W_2,

$$\liminf_{n\to\infty} E(\mu_n) \geq E(\mu).$$

- λ*-convex along generalized geodesics*: For any $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$, there exists a generalized geodesic $\mu_\alpha^{2\to 3}$ from μ_2 to μ_3 with base μ_1 such that, for all $\alpha \in [0, 1]$,

$$E(\mu_\alpha^{2\to 3}) \leq (1-\alpha)E(\mu_2) + \alpha E(\mu_3) - \alpha(1-\alpha)\frac{\lambda}{2} \int |x_2 - x_3|^2 d\boldsymbol{\mu}(x). \tag{2-2}$$

  Note that, for $\lambda > 0$, this condition is stronger than requiring that $E(\mu_\alpha^{2\to 3})$, considered as a real-valued function of $\alpha \in [0, 1]$, be $\lambda W_2^2(\mu_2, \mu_3)$ convex, since

$$\int |x_2 - x_3|^2 d\boldsymbol{\mu} \geq W_2^2(\mu_2, \mu_3).$$

[8] In the case $\mu_0 = \delta_0$, the Dirac mass at the origin, this is equivalent to the definition of coercivity in [Ambrosio et al. 2008], which requires that there exist some $\tau_* > 0$ and $\mu_* \in \mathcal{P}_2(\mathbb{R}^d)$ such that

$$\inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \frac{1}{2\tau_*} W_2^2(\mu_*, \nu) + E(\nu) \right\} > -\infty.$$

If E is λ-convex along generalized geodesics, then in particular it is λ-*convex:* for any $\mu_1, \mu_2 \in \mathcal{P}_{\mu_0}(\mathbb{R}^d)$, there exists a geodesic $\mu_\alpha^{1\to 2}$ from μ_1 to μ_2 such that for all $\alpha \in [0, 1]$,

$$E(\mu_\alpha^{1\to 2}) \le (1-\alpha)E(\mu_1) + \alpha E(\mu_2) - \alpha(1-\alpha)\frac{\lambda}{2}W_2^2(\mu_1, \mu_2).$$

This is equivalent to $E(\mu_\alpha^{1\to 2})$, considered as a real-valued function of $\alpha \in [0, 1]$, being $\lambda W_2^2(\mu_1, \mu_2)$ convex [Ambrosio et al. 2008, Remark 9.1.2].

The requirement that a functional $E : \mathcal{P}_{\mu_0} \to \mathbb{R} \cup \{\infty\}$ be proper, coercive, lower semicontinuous, and convex along generalized geodesics is the natural analogue of the Hilbertian requirement that a functional $E : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ be proper, lower semicontinuous, and convex. The two differences are the addition of the coercivity assumption and the strengthening of the convexity assumption. In a Hilbert space $\mathcal{H}$, all functionals that are proper, lower semicontinuous, and convex are also coercive (in this sense), so the addition of the coercivity assumption is a natural way to ensure that the 2-Wasserstein Moreau–Yosida regularization is not identically $-\infty$. The convexity assumption is strengthened because convexity along generalized geodesics is the useful 2-Wasserstein analogue of Hilbertian convexity. While in a Hilbert space, $x \mapsto \frac{1}{2}\|x - y\|^2$ is 1-convex along all geodesics, the same does not hold for the 2-Wasserstein metric. Requiring convexity of the functional on a larger class of curves compensates for the weaker convexity W_2^2.

2C. ***Further results about the proximal map.*** The following theorem collects some key results regarding the proximal map.

Theorem 2.1 [Ambrosio et al. 2008, Theorem 4.1.2 and Corollary 4.1.3]. *Given $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics, fix $\tau > 0$ small enough so that $\tau\lambda > -1$. Then, for $\mu \in \overline{D(E)}$, the proximal map*

$$\mu \mapsto \mu_\tau$$

is well-defined. Furthermore, the following variational inequality holds:

$$\frac{1}{2\tau}\left(W_2^2(\mu_\tau, \nu) - W_2^2(\mu, \nu)\right) + \frac{\lambda}{2}W_2^2(\mu_\tau, \nu) \le E(\nu) - E(\mu_\tau) - \frac{1}{2\tau}W_2^2(\mu, \mu_\tau) \quad \text{for all } \nu \in D(E). \tag{2-3}$$

When the proximal map is well-defined, it satisfies an Euler–Lagrange equation — a fact originally observed by Otto [1996; 2001]. We state this result in the framework of [Ambrosio et al. 2008, Lemma 10.1.2].

Lemma 2.2. *Given $E : \mathcal{P}_{\mu_0}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics, fix $\tau > 0$ small enough so that*

$\tau\lambda > -1$. *Assume that* $\mu \in \overline{D(E)}$ *so* $\mu \mapsto \mu_\tau$ *is well-defined by Theorem 2.1. Then*

$$\tau |\nabla_W E(\mu_\tau)| \le W_2(\mu, \mu_\tau). \tag{2-4}$$

We may interpret $|\nabla_W E(\mu_\tau)|$ *as the metric slope* (1-11) *when* E *and* μ *lack sufficient smoothness for the norm of the* 2*-Wasserstein gradient* (1-9) *to be well-defined.*

On the other hand, if $\mu \in \mathcal{P}_2^a(\mathbb{R}^d)$ *and both* E *and* μ_τ *are smooth enough so that the* 2*-Wasserstein gradient* $\nabla_W E(\mu_\tau)$ *is well-defined by* (1-8)*, then*

$$\boldsymbol{t}_{\mu_\tau}^{\mu} = \mathrm{id} + \tau \nabla \frac{\delta E}{\delta \rho}(\mu_\tau) \tag{2-5}$$

μ_τ*-almost everywhere and*

$$\tau |\nabla_W E(\mu_\tau)| = W_2(\mu, \mu_\tau). \tag{2-6}$$

Proof. (2-4) follows from [Ambrosio et al. 2008, Theorem 3.1.6].

(2-5) follows from [ibid., Lemma 10.1.2] and the fact that, when E is differentiable, $\nabla \frac{\delta E}{\delta \rho}(\mu_\tau)$ is the unique element of its subdifferential at μ_τ.

(2-6) follows from (2-5) by considering the $L^2(\mu_\tau)$ norm of $\boldsymbol{t}_{\mu_\tau}^{\mu} - \mathrm{id} = \tau \nabla \frac{\delta E}{\delta \rho}(\mu_\tau)$. □

3. Proofs of Theorems 1.1, 1.2, and 1.3 and Corollary 1.6

We now prove the theorems and corollaries announced in the introduction, turning first to the generalized convexity of E_τ. In a Hilbert space, if E is proper, lower semicontinuous, and convex, then its Moreau–Yosida regularization E_τ is also convex. It is well known that the exact analogue in the 2-Wasserstein metric is false. For lack of a reference, we provide the following example.

Fix $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and define $E : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ by

$$E(\mu) := \begin{cases} 0 & \text{if } \mu = \mu_0, \\ \infty & \text{otherwise.} \end{cases} \tag{3-1}$$

E is proper, coercive, lower semicontinuous, and convex along all curves in $\mathcal{P}_2(\mathbb{R}^d)$. In particular, E is convex along generalized geodesics. By definition,

$$E_\tau(\mu) = \inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} W_2^2(\mu, \nu) + E(\nu) \right\} = \frac{1}{2\tau} W_2^2(\mu, \mu_0).$$

By [Ambrosio et al. 2008, Example 9.1.5], when the dimension of the underlying space satisfies $d \ge 2$, E_τ is not λ-convex along geodesics for any $\lambda \in \mathbb{R}$.

As demonstrated by the previous example, the convexity of E_τ is related to the convexity of the squared 2-Wasserstein distance. This also holds in the Hilbertian case, where the convexity of E_τ is a consequence of the 1-convexity of the map

$x \mapsto \frac{1}{2}\|x-y\|^2$ [Moreau 1967]. Therefore, it is natural that our proof of the convexity inequality for E_τ requires the following convexity inequality for W_2^2.

Lemma 3.1 (convexity inequality for W_2^2). *Fix three measures $\mu_1, \mu_2, \mu_3 \in \mathcal{P}(\mathbb{R}^d)$ that are a finite 2-Wasserstein distance apart. Let $\mu_\alpha^{1\to 3}$ be a generalized geodesic from μ_1 to μ_3 with base point μ_2,*

$$\mu_\alpha^{1\to 3} := ((1-\alpha)\pi_1 + \alpha\pi_3)\#\boldsymbol{\mu},$$

where $\boldsymbol{\mu} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ satisfies $\boldsymbol{\mu}_{1,2} := \pi_{1,2}\#\boldsymbol{\mu} \in \Gamma_0(\mu_1, \mu_2)$ and $\boldsymbol{\mu}_{2,3} := \pi_{2,3}\#\boldsymbol{\mu} \in \Gamma_0(\mu_2, \mu_3)$. Let $\mu_\alpha^{1\to 2}$ be the geodesic from μ_1 to μ_2 defined by

$$\mu_\alpha^{1\to 2} := ((1-\alpha)\pi_1 + \alpha\pi_2)\#\boldsymbol{\mu}_{1,2}.$$

Then,

$$W_2^2(\mu_\alpha^{1\to 2}, \mu_\alpha^{1\to 3}) \le (1-\alpha)W_2^2(\mu_1, \mu_1) + \alpha W_2^2(\mu_2, \mu_3) - \alpha(1-\alpha)W_2^2(\mu_2, \mu_3). \quad (3\text{-}2)$$

Proof. Note that

$$\mu_\alpha^{1\to 2} = ((1-\alpha)\pi_1 + \alpha\pi_2)\#\boldsymbol{\mu}_{1,2} = ((1-\alpha)\pi_1 + \alpha\pi_2)\#\boldsymbol{\mu}.$$

Then by [Ambrosio et al. 2008, Equation 7.1.6],

$$\begin{aligned} W_2^2(\mu_\alpha^{1\to 2}, \mu_\alpha^{1\to 3}) &\le \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |[(1-\alpha)\pi_1 + \alpha\pi_3] - [(1-\alpha)\pi_1 + \alpha\pi_2]|^2 \, d\boldsymbol{\mu} \\ &= \alpha^2 \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |\pi_2 - \pi_3|^2 \, d\boldsymbol{\mu} = \alpha^2 \int_{\mathbb{R}^d\times\mathbb{R}^d} |\pi_2 - \pi_3|^2 \, d\boldsymbol{\mu}_{2,3} \\ &= \alpha^2 W_2^2(\mu_2, \mu_3) \\ &= (1-\alpha)W_2^2(\mu_1, \mu_1) + \alpha W_2^2(\mu_2, \mu_3) - \alpha(1-\alpha)W_2^2(\mu_2, \mu_3). \end{aligned}$$

□

We now use this convexity inequality for W_2^2 to prove Theorem 1.1.

Proof of Theorem 1.1. Since E is proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics for $\lambda \ge 0$, by Theorem 2.1, the proximal map $\mu \mapsto \mu_\tau$ is well-defined for $\mu \in \overline{D(E)}$ and $\tau > 0$. Let $\mu_\alpha^{\bar\mu\to\mu_\tau}$ be the generalized geodesic from $\bar\mu$ to μ_τ with base point μ on which E satisfies Equation (2-2). Defining $\mu_1 := \bar\mu$, $\mu_2 := \mu$, and $\mu_3 := \mu_\tau$, let $\mu_\alpha^{\bar\mu\to\mu}$ be the geodesic from $\bar\mu$ to μ described in Lemma 3.1. By Lemma 3.1,

$$W_2^2(\mu_\alpha^{\bar\mu\to\mu}, \mu_\alpha^{\bar\mu\to\mu_\tau}) \le (1-\alpha)W_2^2(\bar\mu, \bar\mu) + \alpha W_2^2(\mu, \mu_\tau) - \alpha(1-\alpha)W_2^2(\mu, \mu_\tau).$$

This allows us to bound $E_\tau(\mu_\alpha^{\bar\mu\to\mu})$ from above:

$$
\begin{aligned}
E_\tau(\mu_\alpha^{\bar\mu\to\mu}) &= \inf_{\nu\in\mathcal{P}_{\mu_0}(\mathbb{R}^d)}\left\{\frac{1}{2\tau}W_2^2(\mu_\alpha^{\bar\mu\to\mu},\nu)+E(\nu)\right\}\\
&\le \frac{1}{2\tau}W_2^2(\mu_\alpha^{\bar\mu\to\mu},\mu_\alpha^{\bar\mu\to\mu_\tau})+E(\mu_\alpha^{\bar\mu\to\mu_\tau})\\
&\le \frac{1}{2\tau}\left((1-\alpha)W_2^2(\bar\mu,\bar\mu)+\alpha W_2^2(\mu,\mu_\tau)-\alpha(1-\alpha)W_2^2(\mu,\mu_\tau)\right)\\
&\quad+(1-\alpha)E(\bar\mu)+\alpha E(\mu_\tau)-\alpha(1-\alpha)\frac{\lambda}{2}W_2^2(\bar\mu,\mu_\tau)\\
&\le (1-\alpha)E_\tau(\bar\mu)+\alpha E_\tau(\mu)-\alpha(1-\alpha)\left(\frac{1}{2\tau}W_2^2(\mu,\mu_\tau)+\frac{\lambda}{2}W_2^2(\bar\mu,\mu_\tau)\right).
\end{aligned}
$$

In the last step, we used that $(\bar\mu)_\tau=\bar\mu$, since E attains its minimum at $\bar\mu$. Now, we apply

$$
\alpha a^2+\beta b^2\ge\frac{\alpha\beta}{\alpha+\beta}(a+b)^2\quad\text{for }\alpha>0,\beta\ge0
$$

with $\alpha=1/\tau$ and $\beta=\lambda$:

$$
\begin{aligned}
E_\tau(\mu_\alpha^{\bar\mu\to\mu}) &\le (1-\alpha)E_\tau(\bar\mu)+\alpha E_\tau(\mu)-\alpha(1-\alpha)\frac{\lambda_\tau}{2}\left(W_2(\mu,\mu_\tau)+W_2(\bar\mu,\mu_\tau)\right)^2\\
&\le (1-\alpha)E_\tau(\bar\mu)+\alpha E_\tau(\mu)-\alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\mu,\bar\mu).
\end{aligned}
$$

□

We now use this convexity inequality to prove Theorem 1.2.

Proof of Theorem 1.2. We first prove the Talagrand inequality. Since E attains its minimum at $\bar\mu$, so does E_τ. Therefore, (1-16) implies that, for all $\mu\in\overline{D(E)}$,

$$
E_\tau(\bar\mu)\le E_\tau(\mu_\alpha^{\bar\mu\to\mu})\le(1-\alpha)E_\tau(\bar\mu)+\alpha E_\tau(\mu)-\alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\bar\mu,\mu).
$$

Rearranging gives $\alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\bar\mu,\mu)\le\alpha\left(E_\tau(\mu)-E_\tau(\bar\mu)\right)$. Thus, for all $\alpha\in(0,1)$,

$$
(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\bar\mu,\mu)\le E_\tau(\mu)-E_\tau(\bar\mu).
$$

Sending $\alpha\to0$ gives the Talagrand inequality (1-17).

We now prove the HWI inequality. Again by (1-16), for all $\mu\in\overline{D(E)}$,

$$
E_\tau(\mu_\alpha^{\bar\mu\to\mu})\le(1-\alpha)E_\tau(\bar\mu)+\alpha E_\tau(\mu)-\alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\mu,\bar\mu).
$$

Rearranging and using $\mu_\alpha^{\bar\mu\to\mu}=\mu_{1-\alpha}^{\mu\to\bar\mu}$ and $(1-\alpha)W_2(\mu,\bar\mu)=W_2(\mu,\mu_{1-\alpha}^{\mu\to\bar\mu})$

gives, for $\alpha \in (0,1)$,

$$(1-\alpha)E_\tau(\mu)-(1-\alpha)E_\tau(\bar\mu) \le E_\tau(\mu)-E_\tau(\mu_{1-\alpha}^{\mu\to\bar\mu})-\alpha(1-\alpha)\frac{\lambda_\tau}{2}W_2^2(\mu,\bar\mu)$$

$$E_\tau(\mu)-E_\tau(\bar\mu) \le \frac{E_\tau(\mu)-E_\tau(\mu_{1-\alpha}^{\mu\to\bar\mu})}{1-\alpha}-\alpha\frac{\lambda_\tau}{2}W_2^2(\mu,\bar\mu)$$

$$E_\tau(\mu)-E_\tau(\bar\mu) \le \frac{E_\tau(\mu)-E_\tau(\mu_{1-\alpha}^{\mu\to\bar\mu})}{W_2(\mu,\mu_{1-\alpha}^{\mu\to\bar\mu})}W_2(\mu,\bar\mu)-\alpha\frac{\lambda_\tau}{2}W_2^2(\mu,\bar\mu).$$

Sending $\alpha \to 1$ gives the HWI Inequality (1-18). □

Proof of Theorem 1.3. By Theorem 2.1, replacing ν with ν_τ,

$$\frac{1}{2\tau}\left(W_2^2(\mu_\tau,\nu_\tau)-W_2^2(\mu,\nu_\tau)\right)+\frac{\lambda}{2}W_2^2(\mu_\tau,\nu_\tau) \le E(\nu_\tau)-E(\mu_\tau)-\frac{1}{2\tau}W_2^2(\mu,\mu_\tau).$$

Similarly,

$$\frac{1}{2\tau}\left(W_2^2(\nu_\tau,\mu)-W_2^2(\nu,\mu)\right)+\frac{\lambda}{2}W_2^2(\nu_\tau,\mu) \le E(\mu)-E(\nu_\tau)-\frac{1}{2\tau}W_2^2(\nu,\nu_\tau).$$

Adding these and multiplying by 2τ gives

$$W_2^2(\mu_\tau,\nu_\tau)-W_2^2(\nu,\mu)+\lambda\tau\left[W_2^2(\mu_\tau,\nu_\tau)+W_2^2(\mu,\nu_\tau)\right] \le 2\tau\left[E(\mu)-E(\mu_\tau)\right]-W_2^2(\mu,\mu_\tau)-W_2^2(\nu,\nu_\tau).$$

Symmetrically, we also have

$$W_2^2(\mu_\tau,\nu_\tau)-W_2^2(\nu,\mu)+\lambda\tau\left[W_2^2(\mu_\tau,\nu_\tau)+W_2^2(\nu,\mu_\tau)\right] \le 2\tau\left[E(\nu)-E(\nu_\tau)\right]-W_2^2(\mu,\mu_\tau)-W_2^2(\nu,\nu_\tau).$$

Averaging gives

$$W_2^2(\mu_\tau,\nu_\tau)-W_2^2(\nu,\mu)+\frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau)+W_2^2(\mu,\nu_\tau)+W_2^2(\nu,\mu_\tau)\right] \le \tau\left[E(\nu)-E(\nu_\tau)+E(\mu)-E(\mu_\tau)\right]-W_2^2(\mu,\mu_\tau)-W_2^2(\nu,\nu_\tau).$$

This allows us to bound the change in $\Lambda_\tau(\mu,\nu)$ from above:

$$\Lambda_\tau(\mu_\tau,\nu_\tau)-\Lambda_\tau(\mu,\nu) = W_2^2(\mu_\tau,\nu_\tau)+\frac{\tau^2}{2}|\nabla_W E(\mu_\tau)|^2+\frac{\tau^2}{2}|\nabla_W E(\nu_\tau)|^2 - W_2^2(\mu,\nu)-\frac{\tau^2}{2}|\nabla_W E(\mu)|^2-\frac{\tau^2}{2}|\nabla_W E(\nu)|^2$$

$$\le \tau\,[E(\nu)-E(\nu_\tau)+E(\mu)-E(\mu_\tau)]-W_2^2(\mu,\mu_\tau)-W_2^2(\nu,\nu_\tau)$$
$$+\frac{\tau^2}{2}|\nabla_W E(\mu_\tau)|^2+\frac{\tau^2}{2}|\nabla_W E(\nu_\tau)|^2-\frac{\tau^2}{2}|\nabla_W E(\mu)|^2-\frac{\tau^2}{2}|\nabla_W E(\nu)|^2$$
$$-\frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau)+W_2^2(\mu,\nu_\tau)+W_2^2(\nu,\mu_\tau)\right].$$

By [Ambrosio et al. 2008, Equation 10.1.7, Lemma 10.1.5] and Hölder's inequality, the λ-convexity of E implies

$$E(\nu)-E(\nu_\tau)\le|\nabla_W E(\nu)|W_2(\nu,\nu_\tau)-\frac{\lambda}{2}W_2^2(\nu,\nu_\tau). \tag{3-3}$$

Combining this with the Euler–Lagrange equation (2-4),

$$\Lambda_\tau(\mu_\tau,\nu_\tau)-\Lambda_\tau(\mu,\nu)$$
$$\le\tau|\nabla_W E(\nu)|W_2(\nu,\nu_\tau)+\tau|\nabla_W E(\mu)|W_2(\mu,\mu_\tau)-W_2^2(\mu,\mu_\tau)-W_2^2(\nu,\nu_\tau)$$
$$+\frac{1}{2}W_2^2(\mu,\mu_\tau)+\frac{1}{2}W_2^2(\nu,\nu_\tau)-\frac{\tau^2}{2}|\nabla_W E(\mu)|^2-\frac{\tau^2}{2}|\nabla_W E(\nu)|^2$$
$$-\frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau)+W_2^2(\mu,\nu_\tau)+W_2^2(\nu,\mu_\tau)\right]-\frac{\lambda\tau}{2}\left[W_2^2(\nu,\nu_\tau)+W_2^2(\mu,\mu_\tau)\right].$$

Completing the square gives the result:

$$\Lambda_\tau(\mu_\tau,\nu_\tau)-\Lambda_\tau(\mu,\nu)$$
$$\le-\frac{1}{2}(\tau|\nabla_W E(\nu)|-W_2(\nu,\nu_\tau))^2-\frac{1}{2}(\tau|\nabla_W E(\mu)|-W_2(\mu,\mu_\tau))^2$$
$$-\frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau)+W_2^2(\mu,\nu_\tau)+W_2^2(\nu,\mu_\tau)+W_2^2(\nu,\nu_\tau)+W_2^2(\mu,\mu_\tau)\right].\quad\square$$

Proof of Corollary 1.6. First, we use $\lambda>0$ and the Euler–Lagrange equation (2-4) to rewrite (1-22):

$$\Lambda_\tau(\mu_\tau,\nu_\tau)-\Lambda_\tau(\mu,\nu)$$
$$\le-\frac{1}{2}(\tau|\nabla_W E(\nu)|-W_2(\nu,\nu_\tau))^2-\frac{1}{2}(\tau|\nabla_W E(\mu)|-W_2(\mu,\mu_\tau))^2$$
$$-\frac{\lambda\tau}{2}\left[2W_2^2(\mu_\tau,\nu_\tau)+W_2^2(\mu,\nu_\tau)+W_2^2(\nu,\mu_\tau)+\tau^2|\nabla_W E(\nu_\tau)|^2+\tau^2|\nabla_W E(\mu_\tau)|^2\right]$$
$$=-\frac{1}{2}(\tau|\nabla_W E(\nu)|-W_2(\nu,\nu_\tau))^2-\frac{1}{2}(\tau|\nabla_W E(\mu)|-W_2(\mu,\mu_\tau))^2$$
$$-\frac{\lambda\tau}{2}\left[2\Lambda_\tau(\mu_\tau,\nu_\tau)+W_2^2(\mu,\nu_\tau)+W_2^2(\nu,\mu_\tau)\right].$$

Rearranging terms, we have

$$(1+\lambda\tau)\Lambda_\tau(\mu_\tau,\nu_\tau)$$
$$\le \Lambda_\tau(\mu,\nu)-\frac{1}{2}(\tau|\nabla_W E(\nu)|-W_2(\nu,\nu_\tau))^2-\frac{1}{2}(\tau|\nabla_W E(\mu)|-W_2(\mu,\mu_\tau))^2$$
$$-\frac{\lambda\tau}{2}\big[W_2^2(\mu,\nu_\tau)+W_2^2(\nu,\mu_\tau)\big]. \tag{3-4}$$

By the triangle inequality,

$$\begin{aligned} W_2^2(\mu,\nu_\tau) &\ge W_2^2(\mu,\nu)+W_2^2(\nu,\nu_\tau)-2W_2(\mu,\nu)W_2(\nu,\nu_\tau)\\ &\ge W_2^2(\mu,\nu)-2W_2(\mu,\nu)W_2(\nu,\nu_\tau)\\ &\ge W_2^2(\mu,\nu)-2\Lambda_\tau^{1/2}(\mu,\nu)W_2(\nu,\nu_\tau), \end{aligned}$$

and we have a similar bound for $W_2^2(\mu_\tau,\nu)$.

Finally, for $\lambda\tau\le 1$,

$$\begin{aligned} \frac{1}{2}(\tau|\nabla_W E(\mu)|-W_2(\mu,\mu_\tau))^2 &\ge \lambda\tau\left(\frac{\tau^2}{2}|\nabla_W E(\mu)|^2-\tau|\nabla_W E(\mu)|W_2(\mu_\tau,\mu)\right)\\ &\ge \lambda\tau\left(\frac{\tau^2}{2}|\nabla_W E(\mu)|^2-\sqrt{2}\Lambda^{1/2}(\mu,\nu)W_2(\mu_\tau,\mu)\right), \end{aligned}$$

and again we have the same inequality with μ in place of ν. Using these inequalities in (3-4) we obtain the desired bound. □

4. Examples and applications

4A. ***Inequalities*** (1-16) ***and*** (1-22) ***are sharp.*** Our first example shows that the inequality (1-16) from Theorem 1.1 and the inequality (1-22) from Theorem 1.3 are both sharp. For $\lambda\in\mathbb{R}$, consider the functional $E:\mathcal{P}_2^a(\mathbb{R}^d)\to\mathbb{R}$ defined by

$$E(\mu)=\int\frac{\lambda x^2}{2}d\mu. \tag{4-1}$$

As shown in [Ambrosio et al. 2008, Example 9.3.1], E is proper, coercive, lower semicontinuous, and λ-convex along generalized geodesics.

Proposition 4.1. *For E given by* (4-1), $\lambda\ge 0$, *and* $\tau>0$, *define* $\lambda_\tau:=\dfrac{\lambda}{1+\lambda\tau}$. *Then E_τ is λ_τ-convex, and no more.*

Proposition 4.2. *For E given by* (4-1), $\mu,\nu\in D(E)$, *and* $\tau>0$ *small enough so that* $\lambda\tau>-1$, *there is equality in* (1-22).

We first prove the following lemma. For E given by (4-1), it is well-known that the proximal map is simply a scale transformation:

Lemma 4.3. *For E given by* (4-1), $\mu \in D(E)$, *and* $\tau > 0$ *small enough so that* $\lambda\tau > -1$, *the proximal map associated to E is the scale transformation*

$$\mu \mapsto (1+\lambda\tau)^{-1}\mathrm{id}\#\mu \tag{4-2}$$

where $\mathrm{id}(x) = x$ *is the identity transformation. Moreover, for any* $\mu, \nu \in D(E)$,

$$W_2^2(\mu_\tau, \nu_\tau) = \frac{1}{(1+\lambda\tau)^2} W_2^2(\mu, \nu) \tag{4-3}$$

and

$$W_2^2(\mu, \nu_\tau) = \frac{1}{1+\lambda\tau}\left[W_2^2(\mu, \nu) + 2\tau\left(E(\mu) - \frac{1}{1+\lambda\tau}E(\nu)\right)\right]. \tag{4-4}$$

Proof. At any $\mu \in D(E)$,

$$\nabla\frac{\delta E}{\delta\rho}(\mu) = \nabla\frac{\lambda x^2}{2} = \lambda x. \tag{4-5}$$

For $\tau > 0$ small enough so that $\lambda\tau > -1$, the Euler–Lagrange equation (2-5) becomes

$$\boldsymbol{t}_{\mu_\tau}^{\mu}(x) = x + \lambda\tau x = (1+\lambda\tau)x,$$

μ_τ-almost everywhere. This shows (4-2):

$$(1+\lambda\tau)^{-1}\mathrm{id}\#\mu = \mu_\tau.$$

Next, fix $\varphi : \mathbb{R}^d \to \mathbb{R}$ convex and define $\nu := \nabla\varphi\#\mu$. By uniqueness in the Brenier–McCann theorem, $\nabla\varphi$ is the optimal transport map from μ to ν. If ψ is defined by

$$\psi(x) = (1+\lambda\tau)^{-2}\varphi((1+\lambda\tau)x),$$

ψ is convex and $\nabla\psi\#\mu_\tau = \nu_\tau$. Again, by uniqueness in the Brenier–McCann Theorem, $\nabla\psi$ is the optimal transport map between μ_τ and ν_τ. Consequently,

$$\begin{aligned} W_2^2(\mu_\tau, \nu_\tau) &= \int_{\mathbb{R}^d} |\nabla\psi(x) - x|^2 d\mu_\tau \\ &= (1+\lambda\tau)^{-2}\int_{\mathbb{R}^d} |\nabla\varphi((1+\lambda\tau)x) - (1+\lambda\tau)x|^2 d\mu_\tau \\ &= (1+\lambda\tau)^{-2}\int_{\mathbb{R}^d} |\nabla\varphi(x) - x|^2 d\mu \\ &= (1+\lambda\tau)^{-2} W_2^2(\mu, \nu). \end{aligned}$$

This proves (4-3).

Finally, note that if φ is convex and $\nabla\varphi\#\mu = \nu$, by the definition of $W_2^2(\mu, \nu)$ and of E,

$$2\int_{\mathbb{R}^d} x \cdot \nabla\varphi(x) d\mu = \frac{2}{\lambda}(E(\mu) + E(\nu)) - W_2^2(\mu, \nu). \tag{4-6}$$

Using that

$$(1+\lambda\tau)^{-1}\nabla\varphi\#\mu = \nu_\tau,$$

we may argue as above to show

$$\begin{aligned} W_2^2(\mu, \nu_\tau) &= \int_{\mathbb{R}^d} |(1+\lambda\tau)^{-1}\nabla\varphi(x) - x|^2 d\mu \\ &= \frac{2}{\lambda}(1+\lambda\tau)^{-2}E(\nu) + \frac{2}{\lambda}E(\mu) - 2(1+\lambda\tau)^{-1}\int_{\mathbb{R}^d} x \cdot \nabla\varphi(x) d\mu. \end{aligned}$$

Combining this with (4-6) proves (4-4). □

Proof of Proposition 4.1. We first explicitly compute the Moreau–Yosida regularization of E. It follows from (4-2) and the definition of E that for all $\mu \in D(E)$ and $0 < \tau < \infty$,

$$W_2^2(\mu, \mu_\tau) = 2\lambda\tau^2 E(\mu_\tau). \tag{4-7}$$

Again by (4-2),

$$E(\mu_\tau) = (1+\lambda\tau)^{-2}E(\mu). \tag{4-8}$$

Hence,

$$E_\tau(\mu) = \frac{1}{2\tau}W_2^2(\mu, \mu_\tau) + E(\mu_\tau) = (1+\lambda\tau)E(\mu_\tau) = \frac{1}{1+\lambda\tau}E(\mu).$$

Thus, the Moreau–Yosida regularization of E in this (already very regular) case simply multiplies E by a constant.

It is a standard result (see [Ambrosio et al. 2008], for example) that E is λ-convex, and no more. (Its Hessian with respect to the W_2 Riemannian metric is λ times the identity.) It then follows immediately from $E_\tau(\mu) = \frac{1}{1+\lambda\tau}E(\mu)$ that E_τ is no more than λ_τ-convex. □

Proof of Proposition 4.2. We proceed by using Lemma 4.3 to express quantities appearing on either side of (1-22) in terms of $W_2^2(\mu, \nu)$, $E(\mu)$ and $E(\nu)$. By the symmetry of μ and ν, equations (4-3) and (4-4) allow us to express $W_2^2(\mu_\tau, \nu_\tau)$, $W_2^2(\mu, \nu_\tau)$ and $W_2^2(\nu, \mu_\tau)$ in these terms. By (2-6), (4-5), (4-7), and (4-8), we have

$$\begin{aligned} \tau^2|\nabla_W E(\mu)|^2 &= \tau^2 \int (\lambda x)^2 d\mu = 2\lambda\tau^2 E(\mu), \\ \tau^2|\nabla_W E(\mu_\tau)|^2 &= W_2^2(\mu, \mu_\tau) = 2\lambda\tau^2 E(\mu)/(1+\lambda\tau)^2. \end{aligned}$$

Symmetric identities hold with ν in place of μ.

Finally, direct calculation shows that both sides of (1-22) are equal to

$$-\frac{2\lambda\tau+\lambda^2\tau^2}{(1+\lambda\tau)^2}\left[W_2^2(\mu,\nu)+\lambda\tau^2(E(\mu)+E(\nu))\right]. \qquad \square$$

As we see from (4-3), the proximal map for E is always contracting in the W_2 metric for $\lambda > 0$. Thus, in this example, the additional terms in Λ_τ are not required to produce contraction. The point of this example is rather to show that (1-16) and (1-22) are sharp.

4B. ***The discrete gradient flow for the entropy and Rényi entropies.*** In our second example, we consider functionals E_p corresponding to the entropy and Rényi entropies. We apply Theorem 1.3 to obtain a sharp bound, *uniformly* in the steps of the discrete gradient flow sequence, on the rate at which rescaled solutions of the discrete gradient flow converge to certain limiting densities, known as *Barenblatt densities*. This result mirrors a well-known result obtained by Otto for the corresponding continuous gradient flow. In carrying out this analysis, we learn that the discrete gradient flow is surprisingly well-behaved, not only on average, but also uniformly in the steps. We also show that Otto's beautiful sharp results for the continuous gradient flow can be obtained very efficiently from the analysis of the discrete flow.

First, we define the functionals to be considered. For $p > 1 - 1/d$,[9] define $U_p : \mathbb{R}_+ \to \mathbb{R}$ by

$$U_p(s) := \begin{cases} \dfrac{s^p - s}{p-1} & \text{if } p \neq 1, \\ s\log s & \text{if } p = 1. \end{cases}$$

Let $\mathcal{P}_2^a(\mathbb{R}^d)$ be the set of probability measures with finite second moment that are absolutely continuous with respect to the Lebesgue measure. Define the functional $E_p : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ by

$$E_p(\mu) := \begin{cases} \int_{\mathbb{R}^d} U_p(f(x))dx & \text{if } \mu \in \mathcal{P}_2^a(\mathbb{R}^d),\ d\mu(x) = f(x)dx, \\ \infty & \text{otherwise.} \end{cases}$$

For $p = 1$, E_p is minus the entropy. For $p \neq 1$, E_p is minus the Rényi entropy. As shown in [Ambrosio et al. 2008, Example 9.3.6], E_p is proper, lower semicontinuous, and convex along generalized geodesics. As for coercivity, for $p > 1$, E_p is bounded below by $-1/(p-1)$, hence coercive. For $1 - \frac{1}{d} < p < 1$, E_p is not bounded below, since $\int_{\mathbb{R}^d} f^p(x)dx$ can be arbitrarily large. E_1 is neither bounded above nor below. Nevertheless, E_p is coercive for $1 > p > 1 - \frac{1}{d}$ when $d \geq 2$, and

[9] The borderline case $p = 1 - 1/d$ is more involved, and, for the sake of simplicity, we do not consider it in this paper. It may be possible to extend our approach to this case using the regularization techniques developed in [Blanchet et al. 2012].

for $1 > p > \frac{1}{3}$ when $d = 1$. Later, we shall need some of the estimates that imply this, so we now explain this case. The case $p = 1$ can be found in [Jordan et al. 1998].

By Hölder's inequality, with exponents $1/p$ and $1/(1-p)$, for all $\nu \in \mathcal{P}_2^a(\mathbb{R}^d)$ with $d\nu = f(x)dx$,

$$\int_{\mathbb{R}^d} f^p(x)dx = \int_{\mathbb{R}^d} f^p(x)(1+|x|^2)^p(1+|x|^2)^{-p}dx$$
$$\leq \left(\int_{\mathbb{R}^d} f(x)(1+|x|^2)dx\right)^p \left(\int_{\mathbb{R}^d} (1+|x|^2)^{-p/(1-p)}dx\right)^{1-p}.$$

Furthermore, $\int_{\mathbb{R}^d} f(x)|x|^2dx = \int_{\mathbb{R}^d} |x|^2 d\nu = W_2^2(\nu, \delta_0)$, where δ_0 is the Dirac mass at the origin. By the triangle inequality, for any $\mu \in \mathcal{P}_2^a(\mathbb{R}^d)$,

$$W_2(\nu, \delta_0) \leq W_2(\mu, \nu) + W_2(\mu, \delta_0),$$

so that

$$\int_{\mathbb{R}^d} f^p(x)dx \leq \left(\int_{\mathbb{R}^d} (1+|x|^2)^{-p/(1-p)}dx\right)^{1-p} (1+(W_2(\mu, \nu) + W_2(\mu, \delta_0))^2)^p.$$

Finally, defining

$$C_p := \frac{1}{1-p}\left(\int_{\mathbb{R}^d} (1+|x|^2)^{-p/(1-p)}dx\right)^{1-p},$$

we have for all $\mu, \nu \in \mathcal{P}_2^a(\mathbb{R}^d)$,

$$E_p(\nu) \geq -C_p\left(1 + 2\int_{\mathbb{R}^d} |x|^2 d\mu + 2W_2^2(\mu, \nu)\right)^p. \tag{4-9}$$

Thus, for all $\mu, \nu \in \mathcal{P}_2^a(\mathbb{R}^d)$,

$$\frac{1}{2\tau}W_2^2(\mu, \nu) + E_p(\nu)$$
$$\geq \frac{1}{2\tau}W_2^2(\mu, \nu) - C_p\left(1 + 2\int_{\mathbb{R}^d} |x|^2 d\mu + 2W_2^2(\mu, \nu)\right)^p. \tag{4-10}$$

For fixed μ, the right-hand side is bounded below for all $\tau > 0$ and $\nu \in \mathcal{P}_2^a(\mathbb{R}^d)$; hence E_p is coercive.

Note that the condition $p > 1 - \frac{1}{d}$ when $d \geq 2$, and $p > \frac{1}{3}$ when $d = 1$, is exactly the condition to ensure C_p is finite, and it is easy to see that coercivity fails when this is not the case. For a more general result, see [Ambrosio et al. 2008, Remark 9.3.7].

From this analysis, we may also extract an upper bound on $W_2^2(\mu, \mu_\tau)$ which will be useful later.

Lemma 4.4 (distance bound for the proximal map). *If $d \geq 2$, fix $p > 1 - 1/d$, and if $d = 1$, fix $p > \frac{1}{3}$. Let $\mu \in D(E_p)$ and*

$$M(\mu) := 1 + 2\int_{\mathbb{R}^d} |x|^2 d\mu.$$

Then for all τ small enough that $4pC_p\tau < 1$,

$$W_2^2(\mu, \mu_\tau) \leq 2\tau \frac{E_p(\mu) + C_p M(\mu)}{1 - 4pC_p\tau}.$$

A similar, but more complicated, bound in terms of the same quantities holds for all $\tau > 0$.

Proof. By the definition of the proximal map, taking $\nu = \mu$ in the variational problem (1-7), we obtain

$$E_p(\mu) \geq \frac{1}{2\tau} W_2^2(\mu, \mu_\tau) + E_p(\mu_\tau).$$

Then, by (4-10) with $\nu = \mu_\tau$ and Bernoulli's inequality, $(1+u)^p \leq 1 + pu$,

$$\begin{aligned} E_p(\mu) &\geq \frac{1}{2\tau} W_2^2(\mu, \mu_\tau) - C_p \left(M(\mu) + 2W_2^2(\mu, \mu_\tau)\right)^p \\ &= \frac{1}{2\tau} W_2^2(\mu, \mu_\tau) - C_p M^p(\mu) \left(1 + \frac{2W_2^2(\mu, \mu_\tau)}{M(\mu)}\right)^p \\ &\geq \frac{1}{2\tau} W_2^2(\mu, \mu_\tau) - C_p M^p(\mu) \left(1 + p\frac{2W_2^2(\mu, \mu_\tau)}{M(\mu)}\right). \\ &\geq \left[\frac{1}{2\tau} - 2pC_p\right] W_2^2(\mu, \mu_\tau) - C_p M(\mu). \end{aligned}$$

In the last line, we used that $M(\mu) \geq 1$.

The bound is simple due to the use of Bernoulli's inequality $(1+u)^p \leq 1 + pu$. Avoiding this, one obtains a bound without restriction on τ. Since we are mostly concerned with small τ, we leave the details to the reader. □

If $d \geq 2$, fix $p > 1 - 1/d$, and if $d = 1$, fix $p > \frac{1}{3}$. Then, E_p is proper, coercive, lower semicontinuous, and convex along generalized geodesics. Therefore, Theorem 2.1 guarantees that the proximal map and discrete gradient flow (1-10) are well-defined for $0 < \tau < \infty$, $\mu_0 \in \overline{D(E_p)}$. Before turning to the long-time asymptotics of the discrete gradient flow for E_p, we first investigate the contraction properties of $\Lambda_\tau(\mu, \nu)$ under the proximal map.

Unlike the functional considered in Section 4A, E_p is translation invariant. Specifically, for fixed $x_0 \in \mathbb{R}^d$, if T_{x_0} is the translation given by

$$T_{x_0}\mu := (\mathrm{id} - x_0)\#\mu,$$

then $E_p(T_{x_0}\mu) = E_p(\mu)$. The 2-Wasserstein distance is also translation invariant: for any $\mu, \nu \in \mathcal{P}_2^a(\mathbb{R}^d)$

$$W_2^2(\mu, \nu) = W_2^2(T_{x_0}\mu, T_{x_0}\nu).$$

Consequently, the proximal map associated to E_p commutes with translations:

$$(T_{x_0}\mu)_\tau = T_{x_0}(\mu_\tau).$$

On one hand, this implies that the proximal map does not contract strictly in W_2^2: for any $\nu \in \mathcal{P}_2^a(\mathbb{R}^d)$, $W_2^2(\nu, T_{x_0}\nu) = x_0^2$, so

$$W_2^2(\mu_\tau, (T_{x_0}\mu)_\tau) = W_2^2(\mu, T_{x_0}\mu).$$

On the other hand, because the functional E_p is strictly convex [Ambrosio et al. 2008; Otto 2001], strict inequality holds in (3-3) and hence in (1-21) of Theorem 1.3:

$$\Lambda_\tau(\mu_\tau, \nu_\tau) < \Lambda_\tau(\mu, \nu).$$

Therefore, $\Lambda_\tau(\mu, \nu)$ is strictly decreasing under the proximal map, even though $W_2^2(\mu, \nu)$ is not.

We now turn to the long-time asymptotics of the discrete gradient flow for E_p. As shown in [Otto 2001], the $\tau \to 0$ limit of the discrete gradient flow tends to the continuous gradient flow on $\mathcal{P}_2^a(\mathbb{R}^d)$, which corresponds to the porous medium equation or the fast diffusion equation:

$$\frac{\partial}{\partial t}\rho(t, x) = \Delta\rho(t, x)^p. \tag{4-11}$$

(For $p < 1$ this is the fast diffusion equation. For $p > 1$, it is the porous medium equation.) We show that for each $\tau > 0$, the discrete flow is a strikingly close analogue of the continuous flow.

A key feature of (4-11) is that it has *self-similar scaling solutions* known as *Barenblatt solutions*,

$$\sigma_p(t, x) := t^{-d\beta} h_p\left(\frac{x}{t^\beta}\right), \tag{4-12}$$

where

$$\beta := \frac{1}{2 + d(p-1)}, \tag{4-13}$$

and

$$h_p(x) := \begin{cases} \left(\lambda + \dfrac{1-p}{p}\dfrac{\beta}{2}|x|^2\right)^{1/(p-1)} & \text{if } 1 - \dfrac{1}{d} < p < 1, \\ \lambda\, e^{-\beta|x|^2/2} & \text{if } p = 1, \\ \left(\lambda + \dfrac{1-p}{p}\dfrac{\beta}{2}|x|^2\right)_+^{1/(p-1)} & \text{if } p > 1, \end{cases} \tag{4-14}$$

with normalizing constants $\lambda = \lambda(d, p)$ so that $\int_{\mathbb{R}}^{d} d\sigma_p(x) = \int_{\mathbb{R}}^{d} h_p(x)dx = 1$.

Definition 4.5 (Barenblatt density). If μ is a probability measure of the form $d\mu = \sigma_p(t, x)dx$, we call μ a *Barenblatt density*. Going forward, we will simply write $\mu = \sigma_p(t, x)dx$.

We now show that the Barenblatt densities are preserved under the discrete gradient flow. Before stating the next proposition, let us observe that $0 < \beta < 1$ for all values of $p > 1 - 1/d$. Thus, the function $s \mapsto s^\beta - \tau\beta s^{\beta-1}$ is strictly monotone increasing for $s \geq 0$ and yields the value 0 for $s = \tau\beta$. Consequently, for any $r > 0$, there is a unique $s > \tau\beta$ such that

$$r^\beta = s^\beta - \tau\beta s^{\beta-1}. \tag{4-15}$$

Definition 4.6 (proximal time-shift function). Define the proximal time-shift function $\theta_\tau : \mathbb{R}_+ \to \mathbb{R}_+$ so that, for any $r > 0$, $\theta_\tau(r)$ is the unique value of s that solves (4-15).

We have already observed that $\theta_\tau(r) > \tau\beta$ for all $r > 0$. Since $r^\beta - \tau\beta r^{\beta-1} < r^\beta$ for all $r > 0$, $\theta_\tau(r) > r$. The following lemma generalizes a result in [Carlen and Gangbo 2003] for the case $p = 1$, showing that the proximal map for the functional E_p takes $\sigma_p(r, x)dx$ to $\sigma_p(\theta_\tau(r), x)dx$. Thus the proximal map takes a Barenblatt density to a Barenblatt density with a larger "time parameter". Given that the class of Barenblatt densities is preserved at the discrete level, we would of course expect the time parameter to increase.

Proposition 4.7. *If $d \geq 2$, fix $p > 1 - 1/d$, and if $d = 1$, fix $p > \frac{1}{3}$. Let μ be a Barenblatt density, i.e. $\mu = \sigma_p(r, x)dx$ for some $r > 0$. Then, for $\tau > 0$, the image of μ under the proximal map for E_p is of the form*

$$\mu_\tau = \sigma_p(\theta_\tau(r), x)dx. \tag{4-16}$$

Proof. Given a Barenblatt density $\mu = \sigma_p(r, x)dx$ for some $r > 0$, let $s := \theta_\tau(r)$ and $\nu := \sigma_p(s, x)dx$. We compute

$$\nabla\frac{\delta E_p}{\delta\rho}(\nu) = U_p''(\sigma_p(s, x))\nabla\sigma_p(s, x) = p\sigma_p(s, x)^{p-2}\nabla\sigma_p(s, x)(x) = -\frac{\beta x}{s}$$
$$\nu\text{-almost everywhere,} \tag{4-17}$$

Next, note that since $s = \theta_\tau(r) > \tau\beta$,

$$\nabla\varphi(x) := x + \tau\nabla\frac{\delta E_p}{\delta\rho}(\nu) = \left(1 - \frac{\tau\beta}{s}\right)x$$

is the gradient of a convex function. Consequently, if we define

$$\rho := \nabla\varphi\#\nu,$$

uniqueness in the Brenier–McCann theorem guarantees that $\nabla\varphi$ is the optimal transport map between ν and ρ. Since

$$\nabla\varphi = \boldsymbol{t}_\nu^\rho = \mathrm{id} + \tau\nabla\frac{\delta E_p}{\delta\rho}(\nu)$$

is the Euler–Lagrange equation (2-5), $\nu = \rho_\tau$, the image of ρ under the proximal map. With the explicit form of $\nabla\varphi$ and $\sigma_p(s,x)$, we compute

$$\rho = \left(1-\frac{\tau\beta}{s}\right)^{-d}\sigma_p\left(s, \left(1-\frac{\tau\beta}{s}\right)^{-1}x\right)dx = \sigma_p\left(\left(1-\frac{\tau\beta}{s}\right)^{1/\beta}s, x\right)dx.$$

By the definition of $s = \theta_\tau(r)$,

$$r = \left(1-\frac{\tau\beta}{s}\right)^{1/\beta}s. \tag{4-18}$$

Therefore, $\rho = \sigma_p(r,x)dx = \mu$, so $\mu_\tau = \rho_\tau = \nu = \sigma_p(s,x)dx = \sigma_p(\theta_\tau(r),x)dx$. □

Note that when τ is very small compared to $t > 0$, and hence also compared to $s := \theta_\tau(t)$,

$$t = \left(1-\frac{\tau\beta}{s}\right)^{1/\beta}s \approx s - \frac{\tau\beta}{\beta} = s - \tau,$$

so $\theta_\tau(t) \approx t + \tau$. Thus, in this approximation, the proximal map shifts the time forward by τ, independent of t. To the extent this is accurate, it makes it very easy to understand the discrete gradient flow for E_p starting from a Barenblatt density: at the n-th step of size τ, one gets a Barenblatt density whose time parameter has been increased by approximately $n\tau$. The following lemma allows us to control this approximation in precise terms.

Lemma 4.8. *Fix $r > 0$. Then, for all $t \geq r$,*

$$\left(\frac{r}{r+\tau}\right)\tau \;\leq\; \theta_\tau(t) - t \;\leq\; \tau. \tag{4-19}$$

Proof. Let $s := \theta_\tau(t)$ for any $t \geq r$. We recall that $0 < \beta < 1$ for all $p > 1 - 1/d$. By the definition of θ_τ, we have

$$t^\beta = s^\beta - \tau\beta s^{\beta-1}.$$

Assume $s > t + \tau$. Then, by Bernoulli's inequality $(1+u)^{1-\beta} \leq (1+(1-\beta)u)$ with $u := \tau/t$,

$$\begin{aligned} t^\beta &= s^\beta - \tau\beta s^{\beta-1} > (t+\tau)^\beta - \tau\beta(t+\tau)^{\beta-1} \\ &= (t+\tau)^{\beta-1}(t+(1-\beta)\tau) = t^\beta(1+u)^{\beta-1}(1+(1-\beta)u) \geq t^\beta. \end{aligned}$$

This is a contradiction. Therefore, $\theta_\tau(t) = s \le t + \tau$, which proves the upper bound in (4-19).

To obtain the lower bound, we use the upper bound on s and (4-18) to obtain

$$s \ge t\left(1 - \frac{\tau\beta}{t+\tau}\right)^{-1/\beta}.$$

Then since $(1+u)^{-1/\beta} \ge 1 - u/\beta$ and $t \ge r$, we obtain

$$s \ge t\left(1 + \frac{1}{\beta}\frac{\tau\beta}{t+\tau}\right) \ge t + \tau\left(\frac{r}{r+\tau}\right). \qquad \square$$

We may now use Theorem 1.3 to control the rate at which rescaled solutions to the discrete gradient flow converge to a Barenblatt density. First, we define the rescaled discrete gradient flow. For any positive integer n, let θ_τ^n be the n-fold power of θ_τ. For $t > 0$, let S_t denote the scaling transformation given by

$$S_t \nu = \frac{\mathrm{id}}{t^\beta} \# \nu.$$

Since $t^{-\beta}x$ is the gradient of a convex function, uniqueness in the Brenier–McCann theorem implies that it is the optimal transport map from ν to $S_t\nu$.

Let μ be a Barenblatt density, i.e., $\mu = \sigma_p(r, x)dx$ for some $r > 0$. Then $S_r\mu = h_p(x)dx$. Let $\{\mu_n\}$ be the discrete gradient flow with initial data μ for fixed $\tau > 0$. By Proposition 4.7,

$$J_\tau^n \mu = \mu_n = \sigma_p(\theta_\tau^n(r), x)dx,$$

and by definition of the scaling transformation,

$$S_{\theta_\tau^n(r)} J_\tau^n \mu = S_{\theta_\tau^n(r)}\mu_n = h_p(x)dx \qquad \text{for all } n \in \mathbb{N}. \tag{4-20}$$

Thus, each step of the discrete gradient flow sequence is also a rescaling of $h_p(x)dx$.

In fact, something almost as good holds even when the initial data of the discrete gradient flow is not a Barenblatt density. We apply Theorem 1.3 to prove that if $\{\nu_n\}$ is a discrete gradient flow with initial data $\nu \in D(E_p)$ for fixed $\tau > 0$, then

$$\lim_{n\to\infty} S_{\theta_\tau^n(r)} J_\tau^n \nu = \lim_{n\to\infty} S_{\theta_\tau^n(r)}\nu_n = h_p(x)dx.$$

That is, if you wait a while and scale the solution to view it in a fixed length scale, what you see is (essentially) a Barenblatt density, no matter what the initial data $\nu \in D(E_p)$ looked like. Moreover, we show that $W_2(S_{\theta_\tau^n(r)}\nu_n, h_p(x)dx)$ essentially contracts at a precise polynomial rate.

Theorem 4.9 (discrete fast diffusion and porous medium flow). *If $d \ge 2$, fix $p > 1 - 1/d$, and if $d = 1$, fix $p > \frac{1}{3}$. Let $\nu \in D(E_p)$ and let $\mu = \sigma_p(r, x)dx$ for some*

$r > 0$. Given $0 < \tau \le 1$, let $\{\nu_n\}$ and $\{\mu_n\}$ be the discrete gradient flows (1-10) *with initial conditions ν and μ. Define the rescaled discrete gradient flow sequence*

$$\widetilde{\nu}_n := S_{\theta_\tau^n(r)}\nu_n.$$

Then, there is an explicitly computable constant K depending only on d, p, r, $E_p(\nu)$, and

$$M(\nu) := 1 + 2\int_{\mathbb{R}^d} |x|^2 d\nu,$$

such that

$$W_2^2(\widetilde{\nu}_n, h_p(x)dx) \le (\theta_\tau^n(r))^{-2\beta}[W_2(\nu,\mu)[W_2(\nu,\mu) + \tau^{1/2}K] + \tau K]. \quad (4\text{-}21)$$

From this, we readily recover Otto's contraction result for a continuous gradient flow as follows. For any $t > 0$, let $\mathrm{int}(t/\tau)$ denote the integer part of t/τ. By Lemma 4.8, $\theta_\tau(t) = t + \tau$, up to an error that vanishes uniformly in t as $\tau \to 0$. Thus, a simple iteration yields

$$\lim_{\tau\downarrow 0} \theta_\tau^{\mathrm{int}(t/\tau)}(r) = r + t. \quad (4\text{-}22)$$

Interpolating and taking the limit $\tau \to 0$ as in [Jordan et al. 1998], one obtains from $\{\nu_n\}$ a solution $\rho(t,x)$ to $(\partial/\partial t)\rho(t,x) = \Delta\rho(t,x)^p$ with $\rho(0,x)dx = \nu_0$. Define the rescaled solution

$$\widetilde{\rho}(t,x) := (r+t)^{d\beta}\rho(t,(r+t)^\beta x).$$

We then conclude that, for all $t > 0$,

$$W_2^2\big(\widetilde{\rho}(t,x)dx,\ h_p(x)dx\big) \le (r+t)^{-2\beta} W_2^2\big(\rho(0,x)dx,\ \sigma_p(r,x)dx\big).$$

One may choose r to minimize $W_2^2(\rho(0,x)dx,\ \sigma_p(r,x)dx)$. Otto has shown this contraction result is sharp. Hence the "near contraction" result we obtain in the discrete setting cannot be improved in any manner that is uniform in τ.

Other aspects of Otto's analysis that leverage this contraction into a bound on L^1 convergence may be applied at the discrete level without difficulty, and we do not go into the details here. On the other hand, while Otto proves a continuous gradient flow analogue of Theorem 1.3, his proof does not extend to the discrete case. Theorem 1.3 provides the means to carry out the discrete analysis and to show that the discrete gradient flow analogue of (4-11) is surprisingly complete.

Proof of Theorem 4.9. By Theorem 1.3, applied iteratively, we have

$$\Lambda_\tau(\nu_n,\mu_n) \le \Lambda_\tau(\nu_1,\mu_1) = \Lambda_\tau(\nu_\tau,\mu_\tau). \quad (4\text{-}23)$$

We make the comparison with $\Lambda_\tau(\nu_\tau,\mu_\tau)$, not $\Lambda_\tau(\nu,\mu)$, since $|\nabla_W E_p(\nu)|^2$ (and hence $\Lambda_\tau(\mu,\nu)$) may be infinite, but by [Ambrosio et al. 2008, Theorem 3.1.6],

the strict convexity of E implies

$$|\nabla_W E(\nu_\tau)|^2 < |\nabla_W E(\nu)|^2 \tag{4-24}$$

so $\Lambda_\tau(\nu_\tau, \mu_\tau) < \infty$. We shall show that $\Lambda_\tau(\nu_\tau, \mu_\tau)$ is very close to $W_2^2(\nu, \mu)$, differing by a term that is $\mathcal{O}(\tau^{1/2})$. Specifically, there exists a constant K depending only d, p, r, $E_p(\nu)$, and $M(\nu)$, such that

$$\Lambda_\tau(\nu_\tau, \mu_\tau) \le W_2(\nu, \mu)[W_2(\nu, \mu) + \tau^{1/2}K] + \tau K. \tag{4-25}$$

Using this in (4-23), we obtain

$$W_2^2(\nu_n, \mu_n) \le \Lambda_\tau(\nu_n, \mu_n) \le W_2(\nu, \mu)[W_2(\nu, \mu) + \tau^{1/2}K] + \tau K. \tag{4-26}$$

Next, by the scaling properties of the 2-Wasserstein metric and (4-20), for all $n \ge 1$,

$$(\theta_\tau^n(r))^{-2\beta} W_2^2(\nu_n, \mu_n) = W_2^2(S_{\theta_\tau^n(r)}\nu_n, S_{\theta_\tau^n(r)}\mu_n) = W_2^2(\widetilde{\nu}_n, h_p(x)dx).$$

Therefore,

$$W_2^2(\widetilde{\nu}_n, h_p(x)dx) \le (\theta_\tau^n(r))^{-2\beta}[W_2(\nu, \mu)[W_2(\nu, \mu) + \tau^{1/2}K] + \tau K],$$

which is (4-21).

It remains to prove (4-25). First, note that since $\mu = \sigma_p(r, x)dx$, (4-17) implies

$$\nabla \frac{\delta E_p}{\delta \rho}(\mu) = -\frac{\beta x}{r}.$$

Thus, by Lemma 2.2 and the definition of the length of the gradient (1-9),

$$\tau^2 \frac{\beta^2}{r^2} \int_{\mathbb{R}^d} |x|^2 \sigma_p(r, x)dx = \tau^2 |\nabla_W E_p(\mu_\tau)|^2 = W_2^2(\mu, \mu_\tau)\ . \tag{4-27}$$

We will consider the cases $p < 1$, $p = 1$, and $p > 1$ separately. For $1 - \frac{1}{d} < p < 1$, when $d \ge 2$, and $\frac{1}{3} < p < 1$, when $d = 1$, we may use the bound on $W_2(\nu, \nu_\tau)$ provided by Lemma 4.4 to show

$$\tau^2 |\nabla_W E_p(\nu_\tau)|^2 \le W_2^2(\nu, \nu_\tau) \le 2\tau \frac{E_p(\nu) + C_p M(\nu)}{1 - 4pC_p\tau}. \tag{4-28}$$

(This particular bound requires $4pC_p\tau < 1$, but one may prove a similar bound with a more complicated constant that holds for all $\tau > 0$.) By the triangle inequality,

$$\begin{aligned} W_2^2(\mu_\tau, \nu_\tau) &\le (W_2(\mu, \nu) + W_2(\mu, \mu_\tau) + W_2(\nu, \nu_\tau))^2 \\ &\le W_2^2(\mu, \nu) + 2W_2(\mu, \nu)[W_2(\mu, \mu_\tau) + W_2(\nu, \nu_\tau)] \\ &\qquad + 2W_2^2(\mu, \mu_\tau) + 2W_2^2(\nu, \nu_\tau). \end{aligned}$$

Combining this with (4-28) and (4-27) gives

$$\Lambda_\tau(\mu_\tau,\nu_\tau) \le W_2^2(\mu,\nu) + 2W_2(\mu,\nu)\left[\left(2\tau\frac{E_p(\nu)+C_pM(\nu)}{1-4\tau pC_p}\right)^{1/2}+\tau\frac{\beta}{r}\left(\int_{\mathbb{R}^d}|x|^2\sigma_p(r,x)dx\right)^{1/2}\right] + 5\tau\frac{E_p(\nu)+C_pM(\nu)}{1-4\tau pC_p}+\frac{5}{2}\tau^2\frac{\beta^2}{r^2}\int_{\mathbb{R}^d}|x|^2\sigma_p(r,x)dx.$$

This leads directly to (4-25) with an explicit constant.

For $p > 1$, by Lemma 2.2 and the definition of the proximal map,

$$\tau^2|\nabla_W E_p(\nu_\tau)|^2 \le W_2^2(\nu,\nu_\tau) \le 2\tau[E_p(\nu)-E_p(\nu_\tau)].$$

Since E_p is bounded below, an analogous argument leads to (4-25).

The case $p = 1$ is similar to the case $p < 1$; we leave the details to the reader. □

Acknowledgement

We thank Luigi Ambrosio for helpful comments on a draft of this paper. We thank Haim Brézis for an enlightening conversation. We thank the anonymous reviewers for many useful suggestions.

References

[Ambrosio et al. 2008] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*, 2nd ed., Birkhäuser, Basel, 2008.

[Baiocchi 1989] C. Baiocchi, "Discretization of evolution variational inequalities", pp. 59–92 in *Partial differential equations and the calculus of variations: Essays in honor of Ennio De Giorgi*, vol. 1, edited by F. Colombini et al., Progr. Nonlinear Differential Equations Appl. **1**, Birkhäuser, Boston, 1989.

[Blanchet et al. 2012] A. Blanchet, E. A. Carlen, and J. A. Carrillo, "Functional inequalities, thick tails and asymptotics for the critical mass Patlak–Keller–Segel model", *J. Funct. Anal.* **262**:5 (2012), 2142–2230.

[Brézis 1971] H. Brézis, "Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations", pp. 101–156 in *Contributions to nonlinear functional analysis* (Madison, WI, 1971), edited by E. H. Zarantonello, Academic Press, New York, 1971.

[Brézis 1973] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, Notas de Matemática **50**, North-Holland, Amsterdam, 1973.

[Carlen and Gangbo 2003] E. A. Carlen and W. Gangbo, "Constrained steepest descent in the 2-Wasserstein metric", *Ann. of Math.* (2) **157**:3 (2003), 807–846.

[De Giorgi 1993] E. De Giorgi, "New problems on minimizing movements", pp. 81–98 in *Boundary value problems for partial differential equations and applications*, edited by J.-L. Lions et al., RMA Res. Notes Appl. Math. **29**, Masson, Paris, 1993.

[Emmrich 1999] E. Emmrich, "Discrete versions of Gronwall's lemma and their application to the numerical analysis of parabolic problems", preprint 637, Technische Universität Berlin, 1999, Available at http://www3.math.tu-berlin.de/preprints/abstracts/Report-637-1999.rdf.html.

[Jordan et al. 1998] R. Jordan, D. Kinderlehrer, and F. Otto, "The variational formulation of the Fokker–Planck equation", *SIAM J. Math. Anal.* **29**:1 (1998), 1–17.

[Martinet 1970] B. Martinet, "Régularisation d'inéquations variationnelles par approximations successives", *Rev. Française Informat. Recherche Opérationnelle* **4**:Ser. R-3 (1970), 154–158.

[Martinet 1972] B. Martinet, *Algorithmes pour la résolution des problémes d'optimisation et de minmax*, thèse d'état, Université de Grenoble, 1972.

[McCann 1995] R. J. McCann, "Existence and uniqueness of monotone measure-preserving maps", *Duke Math. J.* **80**:2 (1995), 309–323.

[McCann 1997] R. J. McCann, "A convexity principle for interacting gases", *Adv. Math.* **128**:1 (1997), 153–179.

[Moreau 1965] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien", *Bull. Soc. Math. France* **93** (1965), 273–299.

[Moreau 1967] J. J. Moreau, "Fonctionnelles convexes", in *Séminaire sur les équations aux dérivées partielles* (Séminaire Jean Leray), vol. 2 (1966–1967), Collége de France, Paris, 1967.

[Otto 1996] F. Otto, "Doubly degenerate diffusion equations as steepest descent", preprint, 1996, Available at http://www-mathphys.iam.uni-bonn.de/~otto/publications/degenerate.ps.

[Otto 2001] F. Otto, "The geometry of dissipative evolution equations: the porous medium equation", *Comm. Partial Differential Equations* **26**:1-2 (2001), 101–174.

[Villani 2003] C. Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics **58**, American Mathematical Society, Providence, RI, 2003.

Received 25 May 2012. Revised 17 Oct 2012. Accepted 3 Nov 2012.

ERIC A. CARLEN: carlen@math.rutgers.edu
Department of Mathematics, Hill Center, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019, United States

KATY CRAIG: katycc@math.rutgers.edu
Department of Mathematics, Hill Center, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019, United States


MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 1, No. 1, 2013
dx.doi.org/10.2140/memocs.2013.1.67

PTOLEMY'S LONGITUDES AND ERATOSTHENES' MEASUREMENT OF THE EARTH'S CIRCUMFERENCE

LUCIO RUSSO

A statistical analysis of the longitudes reported in Ptolemy's *Geographia* shows that many of them were obtained by distorting in a linear way data which were known with good accuracy. As a consequence, a new estimate of the value of the stadion used by Eratosthenes is obtained, supporting the thesis that his measurement of the Earth's circumference was remarkably accurate. Some conjectures about possible simplifications introduced by Cleomedes in his account of Eratosthenes' method are also proposed.

1. The distortion of longitudes in Ptolemy's *Geographia*

The longitudes of 6345 localities[1] reported by Ptolemy in his *Geographia* [Stückelberger and Graßhoff 2006] are affected by an error which dilates their differences. While this error has been often remarked, it has not been so far analyzed in a quantitative way. The analysis of the distortion of the longitudes for all 6345 localities considered by Ptolemy is inconvenient for several reasons. First, many of the places are not identifiable with reasonable certainty. Furthermore for some regions the systematic error overlaps errors of different nature, due to the lack of knowledge of the country (this is the case, for example, for Indian localities). I have therefore preferred to consider a sample of eighty towns, chosen with the following criteria.

First, since it is plausible that Ptolemy's error stems from a wrong interpretation of hellenistic data, I have restricted the choice to the following regions, which were well known in the Greek world both in hellenistic and imperial times: Spain, Southern Gaul, Italy, Greece, Mediterranean coast of Africa west of Egypt, Egypt, regions of Asia that had belonged to the other hellenistic kingdoms.

Secondly, in order to minimize the influence of errors due to the lack of geographical knowledge and to enhance the effect of Ptolemy's systematic error, I have selected my (nonrandom) sample by trying to choose for each of the previous regions the most famous towns, as the ones whose coordinates were presumably

Communicated by Raffaele Esposito.
MSC2010: 01A20.
Keywords: Eratosthenes' measurement, Ptolemy's geography, stadion.

[1] The number of localities has been counted by A. Stückelberger and G. Graßhoff [2006, p. 23].

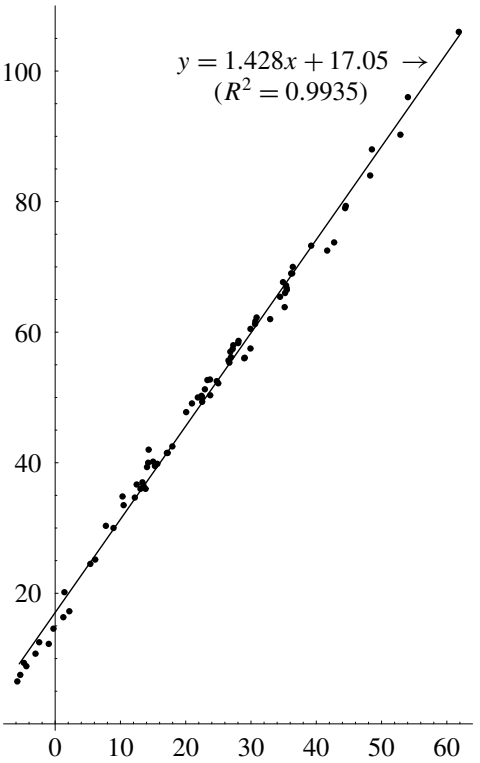


Figure 1. Ptolemy's longitudes (vertical axis) versus actual longitudes from Greenwich (horizonal axis) for the eighty towns in the chosen sample. See also Table 1.

best known, besides being identifiable with certainty. The towns of the sample are listed in Table 1 (page 77) with their longitudes, both actual and as reported by Ptolemy; the data are also plotted in Figure 1.

The regression line has equation

$$y = 1.428x + 17.05.$$

The term 17.05 is of course the longitude that Ptolemy would have assigned to Greenwich and is of no interest to us, while the regression coefficient, 1.428, gives a measure of the dilatation in longitude differences performed by Ptolemy.

We call x_i the actual longitudes from Greenwich of the towns of the sample, y_i their longitudes as reported by Ptolemy and z_i the corresponding values on the regression line ($z_i = 1.428x_i + 17.05$). The variances of the two last series of values are

$$\sigma^2(y_i) = 465.431, \qquad \sigma^2(z_i) = 462.406.$$

The coefficient of determination R^2, which is defined as the ratio $\sigma^2(z_i)/\sigma^2(y_i)$ and is considered a measure of how well empirical data are described by the regression line, is

$$R^2 = 0.9935.$$

A value of R^2 so close to 1 clearly shows that Ptolemy's numbers were obtained by distorting in a linear way data that were known with remarkable accuracy.

2. The value of the stadion used by Eratosthenes

Since antiquity Eratosthenes' measurement of the Earth's circumference is one of the most celebrated achievements of Alexandrian science. The principle of the method used by Eratosthenes in his measurement is well-known and it is not worth to be recalled here.[2] Since we know his result in "stadia", the accuracy of his measure cannot be evaluated without knowing the actual length of the "stadion" used by him. In the Greek world several different "stadia" had been in use and the value of the one used by Eratosthenes is a *vexata questio*. Hultsch, in 1882, had determined it as 157.5 meters [Hultsch 1882] and this measure was accepted by most of the scholars till the first half of the twentieth century. Among the many other values that have been proposed it seems that the most widely accepted nowadays is 185 meters, which is the length of the so-called "Attic stadion".[3] This value is documented in many sources, but not explicitly referring to Eratosthenes, while Hultsch's argument was based essentially only on a single statement by Pliny, which nevertheless refers explicitly to Eratosthenes.[4] If we accept Hultsch's value, the error of Eratosthenes' measure is less than 1%, while if we assume that his stadion was the Attic one the error is about 17%.

Whereas there is no general agreement on the length of the "stadion" used by Eratosthenes, all scholars agree that later geographers, like Hipparchus, Strabo, Marinus and Ptolemy, used his same stadion (as is shown by the fact that many distances in stadia have the same value for all of them). It is well known, on the other hand, that Ptolemy, like Marinus before him, did not accept Eratosthenes' measure of the meridian, corresponding to 700 stadia per degree, adopting instead the measure of 500 stadia per degree.[5] Since our regression coefficient is a fair approximation of the ratio 7/5 between the lengths of the Earth's circumference

[2]The reader is referred to [Russo 2004, pp. 68–69] for some considerations on Eratosthenes' method. While it may seem simple now, it was beyond the understanding of post-hellenistic antiquity.

[3]In [Rawlins 1982] the value of 185 meters for Eratosthenes' stadion is considered a well established fact. The same value is accepted by, among others, Dicks [1960] and by Berggren and Jones [2000].

[4]Pliny, *Naturalis historia* , XII, 53.

[5]The origin of this new measure is unknown. Ptolemy (*Geographia*, I, 11), without mentioning Eratosthenes' measure, simply states that there was a general consensus on the measure of 500 stadia

according Eratosthenes and Ptolemy, our computation shows that (as has often been suggested) the distortion operated by Ptolemy on the longitudes is not independent of the new value he had assumed for the length of the Earth's circumference. We may assume, as it is generally accepted,[6] that he had deduced his differences in longitude from known distances, measured in stadia, along a given circle of latitude, so that his distortion of longitudes compensates for the reduced dimensions of the Earth. We know, in fact, that, given the difficulty of determining longitudes by astronomical methods, hellenistic geographers like Eratosthenes preferred to use, instead of longitudes, distances along a given circle of latitude. Since we know that Ptolemy assumed that one degree of a great circle of the Earth had the length of 500 stadia, we can recover from his longitudes the original distances in stadia between a large number of localities, getting precious information on the actual value of the stadion used in geographical treatises.

We call Δl the difference in longitude between two arbitrary places, Δl_T their difference in longitude according to Ptolemy, d_m and d_s the measures, respectively in meters and in stadia, of the arc of equator comprised between their meridians.

Since $d_m \approx 111{,}100\Delta l$, $d_s \approx 500\Delta l_T$, we get that the value in meters of the stadion is

$$s = \frac{d_m}{d_s} \approx \frac{111{,}100}{500}\,\frac{\Delta l}{\Delta l_T} = 222.2\frac{\Delta l}{\Delta l_T}.$$

By replacing the ratio $\Delta l/\Delta l_T$ with its mean value given by the regression coefficient, 1/1.428, we obtain for the stadion the value of 155.6 meters. Since $155.6 \times 252{,}000 = 39{,}211{,}200$, this value would correspond to an error a little less than 2% on Eratosthenes' measurement of the great circle of the Earth.

A possible objection to this procedure is that we cannot exclude that the distances known to Ptolemy were affected by a significant systematic error (so that their accuracy was small, despite their remarkable precision). I can answer this objection in two ways. First, if all large distances were affected by the same systematic error, the value obtained for the stadion may be very different from the one understood by ancient geographers, but corresponds very well to its value *de facto*; in other words, we can use it to convert effectively to kilometers the large distances in stadia recorded by ancient geographers. Secondly, the circumstance that the value we obtained is remarkably close to the one determined by Hultsch on philological grounds (157.5 meters) makes the previous possibility unlikely, lending strong support to Hultsch's determination and allowing us to exclude, in my opinion, that Eratosthenes had used the Attic stadion of 185 meters or the even

per degree. We know from him that the same measure had been adopted by Marinus (Ptolemy, *Geographia*, I, 7; I, 11).

[6] See for example [Berggren and Jones 2000, p. 30].

larger stadia proposed by some scholars. We have to conclude that the relative measurement error was probably within a few percent.

3. Some conjectures on possible simplifications introduced in Cleomedes' account

According to Cleomedes' account of Eratosthenes' measurement,[7] the difference of latitude between Alexandria and Syene, supposed on the tropic (difference which is equal to the angle between sunbeams and the vertical in Alexandria at noon of the summer solstice), was measured as $\frac{1}{50}$ of a turn and the distance between the two cities (supposed on the same meridian) was estimated as 5000 stadia. The length of the great circle, measured in stadia, was then obtained as the result of the multiplication

$$50 \times 5{,}000 = 250{,}000.$$

The result of the previous section shows an accuracy of Eratosthenes' result which is hardly compatible with such round figures, which have been often considered a clear evidence of the crudeness of Eratosthenes' measure (this argument is used, for example, in [Goldstein 1984]).

On the other hand all sources other than Cleomedes unanimously give for the final result the value of 252,000 stadia.[8] The discrepancy is usually explained (see for example [Roller 2010, p. 143]) assuming that Eratosthenes had obtained the round figures reported by Cleomedes, but afterwards had added 2000 stadia to the final result in order to get a figure divisible by 60. Such a reconstruction is hardly acceptable. What number should have recorded Eratosthenes in his lost treatise "On the measurement of the Earth"?[9] If he had reported only the final figure 252,000, Cleomedes could not have recovered the original result of the measurement. Suppose, instead, that Eratosthenes had written that the measurement result had been 250,000 stadia, but that, in his opinion, it could have been convenient to replace it by 252,000. It would be hardly understandable, in this case, why no other source, except Cleomedes, should have recorded the value 250,000, which had the double advantage of being a round figure and the true result of the measurement.

It appears much more likely that the rounding of the figures was one of the simplifications introduced by Cleomedes in his short account (contained in about

[7] Cleomedes, *Caelestia*, I, 7, ll, 48–120 (pp. 35–37, ed. Todd).

[8] Strabo, *Geographia* (II, v, 7; II, v, 34); Geminus, *Introduction to the Phenomena*, XVI, 6; Macrobius, *Commentarii in Somnium Scipionis*, I, xx, 20; Vitruvius, *De architectura*, I, vi, 9; Plinius, *Naturalis Historia*, II, 247; Censorinus, *De die natali*, xiii, 5; Theon of Smyrna, *De utilitate mathematicae*, 124, 10–12 (ed. Hiller); Heron of Alexandria, *Dioptra*, xxxv, 302, 10–17 (ed. Schöne); Martianus Capella, *De nuptiis Philologiae et Mercurii*, VI, 596.

[9] The title of Eratosthenes' work is quoted by Heron of Alexandria (*Dioptra*, xxxv).

three pages in modern editions[10]) of the lost Eratosthenes' treatise in two books. Whereas all other sources, quoting the figure 252,000, intend to report Eratosthenes' result, Cleomedes clarifies in the beginning of his popularization that his only aim is to explain the "method (ἔφοδος)" used by Eratosthenes to readers unable to follow the geometric technicalities of the original work and the accuracy of the figures is clearly irrelevant for this purpose. By rounding the figures Cleomedes might better have achieved the goal to explain Eratosthenes' method without boring the reader with computations which are not immediately worked out mentally. On the other hand, since Cleomedes also writes that a circumference is three times its diameter[11] and it is not conceivable that Eratosthenes had used such a crude estimate of π we know that the rounding of the figures was actually part of the simplifications introduced by him.

Cleomedes could not round the final result 252,000 without altering at least one of the two factors whose product had given such result. On the other hand we have to exclude the possibility that the original multiplication was $50 \times 5{,}040 = 252{,}000$, because large distances are never recorded by ancient geographers with the accuracy of tens of stadia. Hence, if the product was 252,000, we must exclude the number 50 as first factor. Once excluded 50 itself, the only submultiple of 252,000 which can be reasonably rounded to 50 is 48.

We are thus led to conjecture that the original multiplication performed by Eratosthenes might have been

$$48 \times 5{,}250 = 252{,}000,$$

where 5,250 stadia was the measured distance between Alexandria and the northern tropic and $\frac{1}{48}$ of a turn was the measure of the angle between the vertical and the direction of the sunbeams at noon of the summer solstice in Alexandria.

I think that the conjecture above could be accepted, because it is strengthened by three independent elements:

(a) In Eratosthenes' time the angles $\frac{1}{12}$ of a turn (corresponding to one sign of the zodiac, or 30° in our notations), $\frac{1}{24}$ of a turn (half-sign or "step") and $\frac{1}{48}$ of a turn ("part"), as well as sixtieths of a turn, were privileged as units of measurement,[12] so that $\frac{1}{48}$ of a turn was a very natural result of an angular measurement, while the angle reported by Cleomedes ($\frac{1}{50}$ of a turn) is hard to express in the units then used.

[10] See note 8 above.

[11] Cleomedes, *Caelestia*, I, 7, 119–120.

[12] [Neugebauer 1975, pp. 671–672]; [Roller 2010, p. 151].

(b) 5,250 stadia is a plausible result of the measurement of Eratosthenes, because he used to express large distances as multiples of 250 stadia.[13]

(c) An important piece of evidence is provided by Strabo, who reports that the distance between Syene and the Mediterranean was estimated by Eratosthenes as 5,300 stadia.[14] Since Strabo always expresses large distances as multiples of 100 stadia,[15] his figure has the best possible agreement with the value of 5,250 stadia.

If the present conjecture is accepted, one of the consequences is that the error of the angular measure by Eratosthenes was much smaller than what has been so far supposed. The difference of latitude between Alexandria and the tropic was in fact at the time[16] $7°28'$, much nearer to $\frac{1}{48}$ of a turn ($7°30'$) than to Cleomedes' value, corresponding in our notations to $7°12'$.

Another consequence has to do with Eratosthenes' estimate of the error on the measure of the length of the arc of meridian between Alexandria and the tropic. If Eratosthenes assumed the value of 5,250 stadia, we have to think that he was confident to be able to choose the multiple of 250 stadia nearest to the true distance; in other words he may have thought that his error could be less than 125 stadia, or less than about 2.5%, in good agreement with the estimate obtained in Section 2.

In all expositions of Eratosthenes' measurement we read that he supposed that the town of Syene was exactly in the intersection of the tropic with the meridian through Alexandria.[17] Since, as is shown in Figure 2, Syene was actually not far from the tropic,[18] but its difference in longitude with Alexandria is not negligible at all, this assumption, too, seems hardly compatible with the estimate on the error of the result we have found in Section 2.

The universally shared belief that Eratosthenes supposed that Alexandria and Syene were on the same meridian is mainly drawn from Cleomedes' account. Actually, after having exposed Posidonius' method for measuring the Earth, Cleomedes introduces Eratosthenes' measurement with these words:

> ... Eratosthenes' method, being geometrical in nature, is considered more obscure. But what he says will become clear if we premise the

[13] For example the distance between Alexandria and Rhodes was estimated by Eratosthenes as 3,750 stadia (Strabo, *Geographia*, II, v, 24).

[14] Strabo, *Geographia*, XVII, i, 2. In [Rawlins 1982, p. 215], this passage is used, strangely enough, as a proof that Strabo's source was a map pre-dating Eratosthenes and that Eratosthenes had obtained his distance of 5000 stadia just by rounding the ancient value.

[15] [Shcheglov 2003–2007, p. 165].

[16] The latitude of the tropic, i.e., the obliquity of the ecliptic, is about $23°26'$ nowadays, but in hellenistic times it was about $23°44'$.

[17] See for example [Goldstein 1984; Dutka 1993].

[18] Since the latitude of Syene (today Aswan) is $24°05'$N, its distance from the tropic is now almost doubled, but in Eratosthenes' time it was about $21'$.

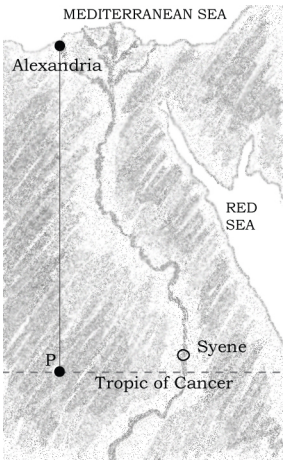


Figure 2. Alexandria, Syene and the tropic (shown in the position it had in Eratosthenes' time).

> following assumptions: suppose first that Alexandria and Syene are on the same meridian ...[19]

These sentences suggest that the assumptions on the coordinates of Syene might be one of the simplifications introduced by Cleomedes. Eratosthenes, for computing the distance between Alexandria and the tropic, had to identify the point P in our figure, i.e., the intersection of the meridian passing through Alexandria with the tropic. In other words he had to measure the component along the meridian of a segment joining Alexandria with any point on the tropic. The individuation of such a component was an operation usual not only in geometry,[20] but also in Eratosthenes' geography.[21] This is a mathematical operation which Cleomedes understandably might have preferred to avoid in exposing the method of Eratosthenes to readers unable to follow geometrical arguments, by replacing the abstract point P with a very concrete town.

[19] [...] ἡ δὲ τοῦ Ἐρατοσθένους γεωμετρικῆς ἐφόδου ἐχομένη καὶ δοκοῦσά τι ἀσαφέστερον ἔχειν. ποιήσει δὲ σαφῆ τὰ λεγόμενα ὑπ' αὐτοῦ τάδε προϋποθεμένων ἡμῶν. ὑποκείσθω ἡμῖν πρῶτον μὲν κἀνταῦθα ὑπὸ τῷ αὐτῷ μεσημβρινῷ κεῖσθαι Συήνην καὶ Ἀλεξάνδρειαν, [...] (Cleomedes, *Caelestia*, I, 7, 49–52).

[20] The orthogonal projection of a point on a line is one of the first procedures explained in a text which was certainly very familiar to hellenistic geographers: Euclid's *Elements* (it is the object of prop. 12 of the first book). Modern historians, who are usually less acquainted with this text, are more inclined to recognize Euclid's influence on hellenistic geographers in some geometrical shapes (see, for example, [Roller 2010, p. 26]) than in geometrical procedures.

[21] Strabo in his *Geographia*, often quoting Eratosthenes, reports several discussions concerning right triangles whose legs are aligned with a meridian and a circle of latitude. Although Strabo does not seem able to master the matter, it is clear that his source was considering orthogonal projections along the two directions.

It is true that also Strabo writes in a couple of passages that according to Eratosthenes the Nile flows along the meridian from Syene to Alexandria, but in the same passages the Nile is described as flowing along the same meridian even from Meroë to Syene,[22] while in the book devoted to Egypt, still quoting Eratosthenes, he describes the path of the Nile as far from being a north-south straight line[23] and in more than one instance Strabo appears to confuse distances with their orthogonal projections along a meridian.[24]

The accuracy of Eratosthenes' measurement resulting from our evaluation of the stadion lends strong support to the conjecture that it was based on land surveying. We know, in fact, that Egypt had a cadastre based on detailed surveying[25] and the use of "royal surveyors" even outside of Egypt is documented by Martianus Capella.[26] Furthermore, the fact that the title of Eratosthenes' treatise "On the measurement of the Earth" is transmitted by Heron's *Dioptra* — by a work devoted to the description of a surveying instrument and of its use — suggests the possibility that part of Eratosthenes' treatise had been devoted to surveying techniques.[27] That no measurement of the Earth's circumference was attempted in Europe until the seventeenth century is per se a strong indication that Eratosthenes planned and oversaw an enterprise requiring a degree of collective organization that cannot be taken for granted in other historical contexts.[28]

A possible objection to the reconstruction so far suggested is that it requires the drawing of an accurate map of Egypt and, whereas such a map was attributed to Eratosthenes in the past, in the last decades the appreciation of ancient cartography has been drastically reduced and the opinion has prevailed that Eratosthenes did not in fact prepare a map of Egypt.[29] We only have evidence of locally confined surveying in ancient Egypt and there is no direct evidence of a map of Egypt drawn

[22] Strabo, *Geographia*, I, iv, 2; II, v, 7.

[23] Strabo, *Geographia*, XVII, i, 2.

[24] For the scant reliability of Strabo in reporting his hellenistic sources, see [Shcheglov 2003–2007]. On our particular subject see also [Rawlins 1982], where some examples of orthogonal components of distances along the path of the Nile, considered by Eratosthenes and misunderstood by Strabo, are recovered.

[25] Valuable information on the Egyptian cadastre is contained in the Oxyrhynchus papyri; see in particular P.Oxy VI 0918 [Grenfell and Hunt 1908, p. 272]. Some useful references on surveying techniques in ancient Egypt are in [Dutka 1993].

[26] Martianus Capella, *De nuptiis Mercurii et Philologiae*, VI, 598.

[27] The use of dioptras by Eratosthenes is well attested (Theon of Alexandria, *Commentaria in Ptolemaei syntaxin mathematicam* i–iv (ed. Rome), 395, 1–2; Simplicius, *In Aristotelis de caelo commentaria*, 246a [CGA 1894, 550]).

[28] On this point see [Russo 2004, pp. 273–277].

[29] Good examples of this new trend are [Harley and Woodward 1987; Brodersen 1995; Rathmann 2007]. (I am indebted to a referee for suggesting these references in this context.) Scholars of ancient science know very well, however, that more recent and better do not always coincide.

by Eratosthenes (apart from the quantitative data reported by Strabo in *Geographia*, XVII, i, 2). The first Greek maps mentioned by our sources date back to the sixth century B.C., but surely did not incorporate quantitative data. On the other hand Ptolemy's *Geographia* is precisely a handbook for drawing maps of the whole *oikoumene* and each of its regions, and for this purpose it stores 12,690 numerical data. We do not know with certainty when the passage from purely symbolic maps to quantitative cartography was accomplished, but it seems reasonable that it was contemporary with the birth of mathematical geography and the introduction of geographical coordinates, i.e., in the time of or shortly before Eratosthenes. On the other hand the thesis that Ptolemy, in his handbook for drawing maps, drew heavily on data from hellenistic sources, in particular incorporating Eratosthenes' material expressed via his value of 700 stadia for a degree of the Earth's circumference, is not only proved by the results in the first two sections of the present paper, but has been shared by other authors on completely different grounds (see [Knobloch et al. 2003; Shcheglov 2004].[30]) Furthermore, the opinion that there were no quantitative maps in Eratosthenes' time is difficult to reconcile with Hipparchus' discussion, in the context of his criticism of Eratosthenes' geographical treatise, of particular directions reported in "ancient maps (ἀρχαῖοι πίνακες)".[31]

Finally, we have a linguistic clue suggesting that Eratosthenes might have extended on a different scale techniques used until then only in local surveying. We know that in hellenistic Egypt officials used a concept analogous to our cadastral sheet, i.e., a portion of land, containing several estates, which was numbered and whose extension and position were described in the cadastral register. Such a portion of land was called σφραγίς.[32] It was Eratosthenes who first introduced the same term σφραγίς in geography, to mean a vastly larger portion of land.[33]

Cleomedes reports an interesting remark made by Eratosthenes in his work. He had observed that at noon of the summer solstice the gnomons gave no shadow not only in the point where the sun was exactly at the zenith, but in a circle around it whose diameter was 300 stadia.[34] It was suggested in [Hultsch 1897] that Eratosthenes had gotten this information from people specifically sent for this purpose, but it is also possible that his estimate had a theoretical basis, being deduced from the knowledge of the angular size of the sun.[35] In either case the remark would

[30] I am indebted to the same anonymous referee for drawing my attention to these references.

[31] Hipparchus' fragment is reported in Strabo, *Geographia*, II, i, 11.

[32] See for example the Oxyrhynchus' papyrus quoted in note 26 above.

[33] See [Roller 2010, pp. 26–27] and Eratosthenes' fragments quoted therein.

[34] Cleomedes, *Caelestia*, I, 7, 101–102 (ed. Todd).

[35] Since the angular size of the sun is about half a degree, the width of the strip where the gnomons gave no true shadow (umbra), but only penumbra, is about half a degree in latitude, or about 350 stadia according to Eratosthenes' measure, but Eratosthenes may have considered that outside a strip 300 stadia wide most of sunlight was intercepted by the gnomons.

Table 1. Longitudes of the towns in the sample

	actual	Ptolemy's		actual	Ptolemy's
Calpe (Gibraltar)	5°21′ W	7°30′	Carthage	10°19′	34°50′
Malaca (Malaga)	4°25′ W	8°50′	Leptis Magna	14°19′	42°
Corduba	4°47′ W	9°20′	Berenice	20°04′	47°45′
Abdara (Adra)	3° 1′ W	10°45′	Ptolemais	20°57′	49°05′
Carthago nova (Cartagena)	0°59′ W	12°15′	Cyrene	21°51′	50°
Tarraco	1°15′	16°20′	Alexandria	29°55′	60°30′
Barcinon (Barcelona)	2°10′	17°15′	Naucratis	30°37′	61°15′
Numantia (Garray)	2°27′ W	12°30′	Oxyrynchus	30°40′	61°40′
Saguntum	0°16′ W	14°35′	Syene (Aswan)	32°56′	62°
Tolosa	1°25′	20°10′	Arsinoe in Eritrea (Assab)	42°44′	73°45′
Massalia (Marseille)	5°23′	24°30′	Chalcedon	29°02′	56°05′
Olbia (Hyères)	6°08′	25°10′	Nicomedia	29°55′	57°30′
Genua	8°56′	30°	Lampsacus	26°41′	55°20′
Populonium	10°29′	33°30′	Pitane	26°56′	56°10′
Roma	12°29′	36°40′	Miletus	27°17′	58°
Cumae (Arco Felice)	14°04′	39°20′	Pergamus	27°11′	57°25′
Paestum	15°00′	40°10′	Sardes	28°02′	58°20′
Croton	17°07′	41°30′	Mytilene	26°33′	55°40′
Rhegium Julium	15°39′	39°50′	Rhodes (Lindos)	28°05′	58°40′
Tarentum	17°14′	41°30′	Samos	26°50′	57°
Brundisium	17°57′	42°30′	Sinope	35°09′	63°50′
Ravenna	12°12′	34°40′	Perga	30°51′	62°15′
Ancona	13°31′	36°30′	Caesarea in Cappadocia	35°29′	66°30′
Camerinum	13°04′	36°	Tarsus	34°54′	67°40′
Capua (Santa Maria C. V.)	14°15′	40°	Phasis in Colchis (Poti)	41°40′	72°30′
Panormus	13°22′	37°	Sidon	35°22′	67°10′
Syracuse	15°17′	39°30′	Antiochia on the Orontes	36°09′	69°
Pola	13°51′	36°	Apamea	36°24′	70°
Abdera	24°59′	52°10′	Carrae	39°13′	73°15′
Byzantium	28°58′	56°	Damascus	36°18′	69°
Philippopolis	24°45′	52°30′	Hierosolyma (Jerusalem)	35°13′	66°
Pella	22°31′	49°20′	Gaza	34°27′	65°25′
Stagira	23°45′	50°20′	Petra	35°27′	66°45′
Athens	23°43′	52°45′	Seleucia on the Tigris	44°31′	79°20′
Thebes (in Boeotia)	23°19′	52°40′	Babylonia (al-Hilla)	44°25′	79°
Delphi	22°30′	50°	Susa	48°15′	84°
Corinth	22°56′	51°15′	Ecbatana (Hamadan)	48°31′	88°
Lacedaemon	22°25′	50°15′	Persepolis	52°53′	90°15′
Tingis Caesarea (Tangier)	5°48′ W	6°30′	Hecatompylon	54°02′	96°
Hippo Regius	7°46′	30°20′	Antiochia Margiana (Merv)	61°50′	106°

Actual longitudes are from Greenwich. Given in parentheses are the current names (or names of nearby modern towns) when they differ from the classical ones.

make no sense if the distance between Alexandria and the tropic was roughly estimated in thousands of stadia. Furthermore, since Syene was 245 stadia away from the tropic, Eratosthenes had determined the distance between Alexandria and the tropic as a multiple of 250 stadia and the central line of a strip 300 stadia wide can certainly be identified with the precision of some tens of stadia, it seems possible that the idea of considering Syene to be on the tropic was another simplification (which, however, Cleomedes shares with many other authors).

I conclude with a remark on a method for measuring large distances which is often recalled in the context of Eratosthenes' measurement. In most of the popular accounts we read that the distance between Alexandria and Syene was reported to Eratosthenes by a "bematist", a man trained to keep a regular pace when marching and to record the number of steps between places. The use of bematists is often presented as the usual method for measuring large distances in the Greek world. But a search of the Thesaurus Linguae Graecae has yielded the result that the word bematist (βηματιστὴς) is attested only once in the entire corpus of Greek literature,[36] in a passage concerning the method used for measuring the distances traveled by the army during the campaign of Alexander the Great, i.e., in circumstances in which usual surveying was hardly practicable.

References

[Berggren and Jones 2000] J. L. Berggren and A. Jones, *Ptolemy's Geography: An annotated translation of the theoretical chapters*, Princeton University Press, 2000.

[Brodersen 1995] K. Brodersen, *Terra cognita: Studien zur römischen Raumerfassung*, Olms, Hildesheim, 1995.

[CGA 1894] J. L. Heiberg (editor), *Commentaria in Aristotelem graeca*, vol. 7, Reimer, Berlin, 1894.

[Dicks 1960] D. R. Dicks (editor), *The geographical fragments of Hipparchus*, Athlone Press, London, 1960.

[Dutka 1993] J. Dutka, "Eratosthenes' measurement of the Earth reconsidered", *Arch. Hist. Exact Sci.* **46**:1 (1993), 55–66.

[Goldstein 1984] B. R. Goldstein, "Eratosthenes on the "measurement" of the earth", *Historia Math.* **11**:4 (1984), 411–416.

[Grenfell and Hunt 1908] B. P. Grenfell and A. S. Hunt (editors), *The Oxyrhynchus papyri*, vol. VI, Egypt Exploration Fund, London, 1908. P.Oxy VI 0918 can be viewed at http://archive.org/stream/oxyrhynchuspappt06grenuoft#page/272/mode/2up.

[Harley and Woodward 1987] J. B. Harley and D. Woodward (editors), *The history of cartography*, vol. 1: *Cartography in prehistoric, ancient, and medieval Europe and the mediterranean*, University of Chicago Press, 1987.

[Hultsch 1882] F. Hultsch, *Griechische und römische Metrologie*, Weidmann, Berlin, 1882. Reprinted 1971.

[36] Athenaeus, *Deipnosophistae*, X, 59, 2.

[Hultsch 1897] F. Hultsch, *Poseidonios über die Grösse und Entfernung der Sonne*, vol. 1, Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen / Philosophisch-historische Klasse. N.F **5**, 1897.

[Knobloch et al. 2003] E. Knobloch, D. Lelgemann, and A. Fuls, "Zur hellenistischen Methode der Bestimmung des Erdumfanges und zur Asienkarte des Klaudios Ptolemaios", *Zeitschrift für Geodäsie, Geoinformation und Landmanagment* **128**:3 (2003), 211–217.

[Neugebauer 1975] O. Neugebauer, *A history of ancient mathematical astronomy*, Springer, Berlin, 1975.

[Rathmann 2007] M. Rathmann (editor), *Wahrnehmung und Erfassung geographischer Räume in der Antike*, Philipp von Zabern, Mainz, 2007.

[Rawlins 1982] D. Rawlins, "The Eratosthenes–Strabo Nile map: Is it the earliest surviving instance of spherical cartography? Did it supply the 5000 stades arc for Eratosthenes' experiment?", *Arch. Hist. Exact Sci.* **26**:3 (1982), 211–219.

[Roller 2010] D. W. Roller (editor), *Eratosthenes' geography: Fragments collected and translated, with commentary and additional material*, Princeton University Press, 2010.

[Russo 2004] L. Russo, *The forgotten revolution: How science was born in 300 BC and why it had to be reborn*, Springer, Berlin, 2004.

[Shcheglov 2003–2007] D. Shcheglov, "Hipparchus' table of climata and Ptolemy's *Geography*", *Orbis Terrarum* **9** (2003–2007), 159–192.

[Shcheglov 2004] D. A. Shcheglov, "Ptolemy's system of seven climata and Eratosthenes' *Geography*", *Geographia Antiqua* **13** (2004), 21–37.

[Stückelberger and Graßhoff 2006] A. Stückelberger and G. Graßhoff (editors), *Klaudios Ptolemaios Handbuch der Geographie: griechisch-deutsch*, Schwabe, Basel, 2006.

Received 21 Sep 2012. Revised 31 Oct 2012. Accepted 2 Dec 2012.

LUCIO RUSSO: lucio.russo@tiscali.it
Department of Mathematics, Università di Roma Tor Vergata, via Keplero 10, I-00142 Roma, Italy

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 1, No. 1, 2013
dx.doi.org/10.2140/memocs.2013.1.81

TV-MIN AND GREEDY PURSUIT FOR CONSTRAINED JOINT SPARSITY AND APPLICATION TO INVERSE SCATTERING

ALBERT FANNJIANG

This paper proposes a general framework for compressed sensing of constrained joint sparsity (CJS) which includes total variation minimization (TV-min) as an example. The gradient- and 2-norm error bounds, independent of the ambient dimension, are derived for the CJS version of basis pursuit and orthogonal matching pursuit. As an application the results extend Candès, Romberg, and Tao's proof of exact recovery of piecewise constant objects with noiseless incomplete Fourier data to the case of noisy data.

1. Introduction

One of the most significant developments of the last decade in imaging and signal processing is compressive sensing (CS) which promises reconstruction with fewer data than the ambient dimension. CS capability [Candès and Tao 2005; Donoho 2006] hinges on favorable sensing matrices and enforcing a key piece of prior knowledge, that is, sparse objects.

Consider the linear inverse problem $Y = \mathbf{\Phi} X + E$ where $X \in \mathbb{C}^m$ is the *sparse* object vector to be recovered, $Y \in \mathbb{C}^n$ is the measurement data vector, and $E \in \mathbb{C}^n$ represents the (model or external) errors. The great insight of CS is that the sparseness of X, as measured by the sparsity $\|X\|_0 \equiv \#$ of nonzero elements in X, can be effectively enforced by ℓ_1-minimization (ℓ_1-min) [Chen et al. 2001; Donoho and Huo 2001]:

$$\min \|Z\|_1 \quad \text{subject to (s.t.)} \quad \|\mathbf{\Phi} Z - Y\|_2 \le \|E\|_2, \tag{1}$$

with favorable sensing matrices $\mathbf{\Phi}$.

The ℓ_1-min idea dates back to geophysics research in the 1970s [Claerbout and Muir 1973; Taylor et al. 1979]. The ℓ_1-minimizer is often a much better approximation of the sparse object than the traditional minimum energy solution

Communicated by Micol Amar.
This research is partially supported by the U.S. National Science Foundation under grant DMS-0908535.
MSC2010: 15A29.
Keywords: total variation, joint sparsity, multiple measurement vectors, compressive sensing.

via ℓ_2-minimization because the 1-norm is closer to $\|\cdot\|_0$ than the 2-norm. Moreover, the ℓ_1-min principle is a convex optimization problem and can be efficiently computed. The ℓ_1-min principle is effective in recovering the sparse object with a number of data points n much smaller than m if the sensing matrix $\mathbf{\Phi}$ satisfies some favorable conditions such as the restricted isometry property (RIP) [Candès and Tao 2005]: $\mathbf{\Phi}$ is said to satisfy the RIP of order k if

$$(1-\delta_k)\|Z\|_{2,2}^2 \le \|\mathbf{\Phi}Z\|_2^2 \le (1+\delta_k)\|Z\|_2^2 \tag{2}$$

for any k-sparse vector Z where the minimum of such a constant δ_k is the restricted isometry constant (RIC) of order k.

The drawback of the RIP is that only a few special types of matrices are known to satisfy it, including independently and identically distributed (i.i.d.) random matrices and random partial Fourier matrices formed by random row selections of the discrete Fourier transform.

A more practical alternative CS criterion is furnished by the incoherence property as measured by one minus the mutual coherence [Donoho and Elad 2003; Tropp 2004]:

$$\mu(\mathbf{\Phi}) = \max_{i\neq j} \frac{\left|\sum_k \Phi_{ik}^* \Phi_{kj}\right|}{\sqrt{\sum_k |\Phi_{ki}|^2}\sqrt{\sum_k |\Phi_{kj}|^2}}. \tag{3}$$

A parallel development in image denoising pioneered by Osher and coworkers [Rudin et al. 1992; Rudin and Osher 1994] seeks to enforce edge detection by total variation minimization (TV-min):

$$\min \int |\nabla g| \quad \text{s.t.} \quad \int |g-f|^2 \le \varepsilon^2, \tag{4}$$

where f is the noisy image and ε is the noise level. The idea is that for the class of piecewise constant functions, the gradient is sparse and can be effectively enforced by TV-minimization.

For digital images, the TV-min approach to deblurring can be formulated as follows. Let $f \in \mathbb{C}^{p\times q}$ be a noisy complex-valued data set of $p \times q$ pixels. Let T be the transformation from the true object to the ideal sensors, modeling the imaging process. Replacing the total variation in (4) by the discrete total variation

$$\|g\|_{\mathrm{TV}} \equiv \sum_{i,j} \sqrt{|\Delta_1 g(i,j)|^2 + |\Delta_2 g(i,j)|^2},$$

$$\Delta g = (\Delta_1 g, \Delta_2 g)(i,j) \equiv \big(g(i+1,j)-g(i,j), g(i,j+1)-g(i,j)\big),$$

we obtain

$$\min \|g\|_{\mathrm{TV}} \quad \text{s.t.} \quad \|Tg-f\|_2 \le \varepsilon \tag{5}$$

(see [Chambolle and Lions 1997; Chan and Shen 2005]).

In a breakthrough paper, Candès et al. [2006] show the equivalence of (5) to (1) for a random partial Fourier matrix with noiseless data ($\varepsilon = 0$) and obtain a performance guarantee of exact reconstruction of piecewise constant objects from (5).

A main application of this present work is to extend the result of Candès et al. to inverse scattering with noisy data. In this context it is natural to work with the continuum setting in which the object is a vector in an infinite-dimensional function space, for example, $L^2(\mathbb{R}^d)$. To fit into CS's discrete framework, we discretize the object function by pixelating the ambient space with a regular grid of equal spacing ℓ.

The grid spacing ℓ can be thought of as the resolution length, the fundamental parameter of the discrete model from which all other parameters are derived. For example, the total number of resolution cells is proportional to ℓ^{-d}, that is, $m = \mathbb{O}(\ell^{-d})$. As we will assume that the original object is well approximated by the discrete model in the limit $\ell \to 0$, the sparsity s of the edges of a piecewise constant object is proportional to ℓ^{1-d}, that is, the object is nonfractal. It is important to keep in mind the continuum origin of the discrete model in order to avoid confusion about the small ℓ limit throughout the paper.

First we introduce the notation for multivectors $\boldsymbol{Y} \in \mathbb{C}^{n\times d}$:

$$\|\boldsymbol{Y}\|_{b,a} = \left(\sum_{j=1}^{n} \|\mathrm{row}_j(\boldsymbol{Y})\|_a^b\right)^{1/b}, \quad a, b \geq 1, \tag{6}$$

where $\mathrm{row}_j(\boldsymbol{Y})$ is the j-th row of $\boldsymbol{Y}$. The 2, 2-norm is exactly the Frobenius norm. To avoid confusion with the *subordinate* matrix norm [Golub and Van Loan 1996], it is more convenient to view $\boldsymbol{Y}$ as multivectors rather than a matrix.

We aim at the following error bounds. Let V be the discretized object and $\hat{V}$ an estimate of V. We will propose a compressive sampling scheme that leads to the error bound for the TV-minimizer $\hat{V}$:

$$\|\Delta V - \Delta \hat{V}\|_{2,2} = \mathbb{O}(\varepsilon), \quad \ell \to 0, \tag{7}$$

implying via the discrete Poincaré inequality that

$$\|V - \hat{V}\|_2 = \mathbb{O}(\varepsilon/\ell) \tag{8}$$

independent of the ambient dimension d.

If $\hat{V}$ is the reconstruction by using a version of the greedy algorithm, orthogonal matching pursuit (OMP) [Pati et al. 1993; Davis et al. 1997], for multivectors then in addition to (7) we also have

$$\|V - \hat{V}\|_2 = \mathbb{O}(\varepsilon/\sqrt{\ell}) \tag{9}$$

independent of the ambient dimension d (see Section 3). We do not know if the bound (9) applies to the TV-minimizer.

A key advantage of the greedy algorithm used to prove (9) is the exact recovery of the gradient support (that is, the edge location) under proper conditions (Theorem 2 in Section 3). On the one hand, TV-min requires fewer data points for recovery: $\mathcal{O}(s)$ for TV-min under the RIP versus $\mathcal{O}(s^2)$ for the greedy algorithm under incoherence where the sparsity $s = \mathcal{O}(\ell^{1-d})$ is as already mentioned. On the other hand, the greedy algorithm is computationally more efficient and incoherent measurements are much easier to design and verify than the RIP.

At heart our theory is based on reformulation of TV-min as CS of joint sparsity with linear constraints (such as the curl-free constraint in the case of TV-min): basis pursuit denoising (BPDN) for constrained joint sparsity (CJS) is formulated as

$$\min\|\boldsymbol{Z}\|_{1,2}, \qquad \text{s.t.} \quad \|\boldsymbol{Y} - \varphi(\boldsymbol{Z})\|_{2,2} \le \varepsilon, \quad \mathscr{L}\boldsymbol{Z} = 0, \tag{10}$$

where

$$\varphi(\boldsymbol{Z}) = [\boldsymbol{\Phi}_1 Z_1, \dots, \boldsymbol{\Phi}_d Z_d], \quad Z_j = \text{the } j\text{-th column of } \boldsymbol{Z},$$

and $\mathscr{L}$ represents a linear constraint. Without loss of generality, we assume the matrices $\{\boldsymbol{\Phi}_j\} \subset \mathbb{C}^{n\times m}$ all have *unit 2-norm* columns.

In connection to TV-min, Z_j is the j-th directional gradient of the discrete object V. And from the definition of discrete gradients, it is clear that every measurement of Z_j can be deduced from two measurements of the object V, slightly shifted in the j-th direction with respect to each other. As shown below, for inverse scattering we have $\boldsymbol{\Phi}_j = \boldsymbol{\Phi}$ for all j, and $\mathscr{L}$ is the curl-free constraint which takes the form

$$\Delta_1 Z_2 = \Delta_2 Z_1$$

for $d = 2$ (see (53)). Our main results, Theorems 1 and 2, constitute performance guarantees for CJS based, respectively, on the RIP and incoherence of the measurement matrices $\boldsymbol{\Phi}_j$.

1.1. ***Comparison of existing theories.*** The gradient-based method of [Patel et al. 2012] modifies the original Fourier measurements to obtain Fourier measurements of the corresponding vertical and horizontal edge images which then are separately reconstructed by the standard CS algorithms. This approach attempts to take advantage of typically lower *separate* sparsity and is different from TV-min. Nevertheless, a similar 2-norm error bound [Patel et al. 2012, Proposition V.2] to (8) is obtained.

Needell and Ward [2012] obtain interesting results for *anisotropic* total variation (ATV) minimization in terms of the objective function

$$\|g\|_{\text{ATV}} \equiv \sum_{i,j} |\Delta_1 g(i, j)| + |\Delta_2 g(i, j)|.$$

While for *real*-valued objects in two dimensions, the isotropic TV seminorm is

equivalent to the anisotropic version, the two seminorms are, however, not the same in dimension greater than 3 and/or for complex-valued objects. A rather remarkable result of Needell and Ward is the bound $\|V - \hat{V}\|_2 = \mathbb{O}(\varepsilon)$, modulo a logarithmic factor, for $d = 2$. This is achieved by proving a strong Sobolev inequality for two dimensions under the additional assumption of the RIP with respect to the bivariate Haar transform. Unfortunately, this latter assumption prevents the results in [Needell and Ward 2012] from being directly applicable to structured measurement matrices such as Fourier-like matrices which typically have high mutual coherence with any compactly supported wavelet basis when adjacent subbands are present. Their approach also does not guarantee exact recovery of the gradient support.

It is worthwhile to further consider these existing approaches from the perspective of the CJS framework for arbitrary d. The approach of [Patel et al. 2012] can be reformulated as solving d standard BPDNs

$$\min\|Z_\tau\|_1, \quad \text{s.t.} \quad \|Y_\tau - \boldsymbol{\Phi} Z_\tau\|_2 \le \varepsilon, \quad \tau = 1, \dots, d,$$

separately without the curl-free constraint $\mathscr{L}$ where Z_τ and Y_τ are, respectively, the τ-th columns of $\boldsymbol{Z}$ and $\boldsymbol{Y}$. To recover the original image from the directional gradients, an additional step of consistent integration becomes an important part of the approach in [Patel et al. 2012].

From the CJS perspective, the ATV-min considered in [Needell and Ward 2012] can be reformulated as follows. Let $\tilde{Z} \in \mathbb{C}^{dm}$ be the image gradient vector by stacking the d directional gradients and let $\tilde{Y} \in \mathbb{C}^{dn}$ be the similarly concatenated data vector. Likewise let $\tilde{\boldsymbol{\Phi}} = \operatorname{diag}(\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_d) \in \mathbb{C}^{dn \times dm}$ be the block-diagonal matrix with blocks $\boldsymbol{\Phi}_j \in \mathbb{C}^{n \times m}$. Then ATV-min is equivalent to BPDN for a *single* constrained and concatenated vector:

$$\min\|\tilde{Z}\|_1, \quad \text{s.t.} \quad \|\tilde{Y} - \tilde{\boldsymbol{\Phi}} \tilde{Z}\|_2 \le \varepsilon, \quad \tilde{\mathscr{L}} \tilde{Z} = 0, \tag{11}$$

where $\tilde{\mathscr{L}}$ is the same constraint $\mathscr{L}$ reformulated for concatenated vectors. Repeating verbatim the proofs of Theorems 1 and 2 we obtain the same error bounds as (7)–(9) for ATV-min as formulated in (11) under the same conditions for $\boldsymbol{\Phi}_j$ separately.

In [Needell and Ward 2012], ATV-min is formulated in terms of the image, instead of the image gradient, to get rid of the curl-free constraint. To proceed the differently concatenated matrix $[\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_d]$ is then assumed to satisfy the RIP of higher order demanding $2dn$ measurement data points. For $d = 2$, Needell and Ward assume the RIP of order $5s$ with $\delta_{5s} < \frac{1}{3}$ for $[\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2]$, which is much more stringent than the RIP of order $2s$ with $\delta_{2s} < \sqrt{2} - 1$ for $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ *separately* in (11). In particular, $\boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2$ is allowed for (11) but not for Needell and Ward. To get the aforementioned favorable $\mathbb{O}(\varepsilon)$ 2-norm error bound for $d = 2$, an additional measurement matrix satisfying the RIP with respect to the bivariate Haar basis is needed, which, as mentioned above, excludes partial Fourier measurements.

1.2. ***Organization.*** The rest of the paper is organized as follows. In Section 2, we present a performance guarantee for BPDN for CJS and obtain error bounds. In Section 3, we analyze the greedy approach to sparse recovery of CJS and derive error bounds, including an improved 2-norm error bound. In Section 4, we review the scattering problem starting from the continuum setting and introduce the discrete model. In Section 5, we discuss various sampling schemes including the forward and backward sampling schemes for inverse scattering for point objects. In Section 6 we formulate TV-min for piecewise constant objects as BPDN for CJS. We present numerical examples and conclude in Section 7. We present the proofs in the Appendices.

2. BPDN for CJS

Consider the linear inversion problem

$$\boldsymbol{Y} = \varphi(\boldsymbol{X}) + \boldsymbol{E}, \quad \mathscr{L}\boldsymbol{X} = 0, \tag{12}$$

where

$$\varphi(\boldsymbol{X}) = [\boldsymbol{\Phi}_1 X_1, \boldsymbol{\Phi}_2 X_2, \ldots, \boldsymbol{\Phi}_d X_d], \quad \boldsymbol{\Phi}_j \in \mathbb{C}^{n\times m},$$

and the corresponding BPDN

$$\min\|\boldsymbol{Z}\|_{1,2}, \quad \text{s.t.} \quad \|\boldsymbol{Y} - \varphi(\boldsymbol{Z})\|_{2,2} \le \varepsilon = \|\boldsymbol{E}\|_{2,2}, \quad \mathscr{L}\boldsymbol{Z} = 0. \tag{13}$$

For TV-min in d dimensions, $\boldsymbol{\Phi}_j = \boldsymbol{\Phi}$ for all j, the vector $\boldsymbol{X}$ represents the discrete gradient of the unknown object V, and $\mathscr{L}$ is the curl-free constraint. Without loss of generality, we assume the matrices $\{\boldsymbol{\Phi}_j\}$ all have unit 2-norm columns.

We say that $\boldsymbol{X}$ is s-row sparse if the number of nonzero rows in $\boldsymbol{X}$ is at most s. With a slight abuse of terminology we call $\boldsymbol{X}$ the object (of CJS).

In the following theorems, we let the object $\boldsymbol{X}$ be general, not necessarily s-row sparse. Let $\boldsymbol{X}^{(s)}$ consist of the s largest rows in the 2-norm of $\boldsymbol{X}$. Then $\boldsymbol{X}^{(s)}$ is the best s-row sparse approximation of $\boldsymbol{X}$.

Theorem 1. *Suppose that the linear map φ satisfies the RIP of order $2s$*

$$(1-\delta_{2s})\|\boldsymbol{Z}\|_{2,2}^2 \le \|\varphi(\boldsymbol{Z})\|_{2,2}^2 \le (1+\delta_{2s})\|\boldsymbol{Z}\|_{2,2}^2 \tag{14}$$

for any $2s$-row sparse $\boldsymbol{Z}$ with $\delta_{2s} < \sqrt{2}-1$. Let $\hat{\boldsymbol{X}}$ be the minimizer of (13)*. Then*

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_{2,2} \le C_1 s^{-1/2}\|\boldsymbol{X} - \boldsymbol{X}^{(s)}\|_{1,2} + C_2\varepsilon \tag{15}$$

for absolute constants C_1 and C_2 depending only on δ_{2s}.

Remark 1. Note that the RIP for joint sparsity (14) follows straightforwardly from the assumption of the separate RIP:

$$(1-\delta_{2s})\|Z\|_2^2 \le \|\boldsymbol{\Phi}_j Z\|_2^2 \le (1+\delta_{2s})\|Z\|_2^2 \quad \text{for all } j,$$

with a common RIC.

Remark 2. For the standard Lasso with a particular choice of regularization parameter, [Candès and Plan 2009, Theorem 1.3] guarantees exact support recovery under a favorable sparsity constraint. In our setting and notation, their TV-min principle suggests

$$\min_{\mathcal{L}\boldsymbol{Z}=0} \lambda\sigma\|\boldsymbol{Z}\|_{1,2} + \tfrac{1}{2}\|\boldsymbol{Y} - \varphi(\boldsymbol{Z})\|_{2,2}^2, \quad \lambda = 2\sqrt{2\log m}, \tag{16}$$

where $\sigma^2 = \varepsilon^2/(2n)$ is the variance of the assumed Gaussian noise in each entry of $\boldsymbol{Y}$. Unfortunately, even if the result of Candès and Plan can be extended to (16), it is inadequate for our purpose because they assume independently selected support and signs, which is clearly not satisfied by the gradient of a piecewise constant object.

The proof of Theorem 1 is given in Appendix A.

The error bound (15) implies (7) for s-row sparse $\boldsymbol{X}$. For the 2-norm bound (8), we apply the discrete Poincaré inequality [Cheung 1998]

$$\|f\|_2^2 \le \frac{m^{2/d}}{4d}\|\Delta f\|_2^2$$

to get

$$\|V - \hat{V}\|_2 \le \frac{m^{1/d}}{2d^{1/2}} C_2\varepsilon = \mathbb{O}\Big(\frac{\varepsilon}{\ell}\Big), \tag{17}$$

since $\ell \sim m^{-1/d}$.

3. Greedy pursuit for CJS

One idea to improve the error bound is through exact recovery of the support. This can be achieved by greedy algorithms. As before, we consider the general linear inversion with CJS (12) with $\|\boldsymbol{E}\|_{2,2} = \varepsilon$.

Algorithm 1 on the next page is an extension of the joint-sparsity greedy algorithms of [Cotter et al. 2005; Chen and Huo 2006; Tropp et al. 2006] to a setting with multiple sensing matrices.

Note that the linear constraint is not enforced in Algorithm 1.

A natural indicator of the performance of OMP is the mutual coherence (3) [Tropp 2004; Donoho et al. 2006]. Let

$$\mu_{\max} = \max_j \mu(\boldsymbol{\Phi}_j).$$

Then, analogous to [Donoho et al. 2006, Theorem 5.1], we have the following performance guarantee.

Algorithm 1. OMP for joint sparsity

Input: $\{\boldsymbol{\Phi}_j\}$, $\boldsymbol{Y}$, $\varepsilon > 0$.
Initialization: $\boldsymbol{X}^0 = 0$, $\boldsymbol{R}^0 = \boldsymbol{Y}$, and $\mathcal{S}^0 = \varnothing$.
Iteration:
1) $i_{\max} = \arg\max_i \sum_{j=1}^d |\Phi_{j,i}^* R_j^{k-1}|$, where $\Phi_{j,i}^*$ is the conjugate transpose of the i-th column of $\boldsymbol{\Phi}_j$.
2) $\mathcal{S}^k = \mathcal{S}^{k-1} \cup \{i_{\max}\}$.
3) $\boldsymbol{X}^k = \arg\min \|\boldsymbol{\Phi Z} - \boldsymbol{Y}\|_{2,2}$ s.t. $\operatorname{supp}(\boldsymbol{Z}) \subseteq S^k$.
4) $\boldsymbol{R}^k = \boldsymbol{Y} - \varphi(\boldsymbol{X}^k)$.
5) Stop if $\sum_j \|R_j^k\|_2 \le \varepsilon$.
Output: $\boldsymbol{X}^k$.

Theorem 2. *Suppose the sparsity s satisfies*

$$s < \frac{1}{2}\Big(1 + \frac{1}{\mu_{\max}}\Big) - \frac{\sqrt{d}\varepsilon}{\mu_{\max} X_{\min}}, \qquad X_{\min} = \min_k \|\operatorname{row}_k(\boldsymbol{X})\|_1. \tag{18}$$

Let $\mathbf{Z}$ be the output of Algorithm 1, *with the stopping rule that the residual drops to the level ε or below. Then* $\operatorname{supp}(\boldsymbol{Z}) = \operatorname{supp}(\boldsymbol{X})$.

Let $\hat{\boldsymbol{X}}$ solve the least-squares problem

$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{B}} \|\boldsymbol{Y} - \boldsymbol{\Phi B}\|_{2,2}, \qquad \textit{s.t.} \qquad \operatorname{supp}(\boldsymbol{B}) \subseteq \operatorname{supp}(\boldsymbol{X}), \quad \mathcal{L}\boldsymbol{B} = 0. \tag{19}$$

Then

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_{2,2} \le \frac{2\varepsilon}{\sqrt{1 - \mu_{\max}(s-1)}}. \tag{20}$$

The proof of Theorem 2 is given in Appendix B.

The main advantage of Theorem 2 over Theorem 1 is the guarantee of exact recovery of the support of $\boldsymbol{X}$. Moreover, a better 2-norm error bound follows because now the gradient error is guaranteed to vanish outside a set of cardinality $\mathcal{O}(\ell^{1-d})$: Let $\mathbb{L} \subset \mathbb{Z}^d$ be a finite lattice of $\mathcal{O}(\ell^{-d})$ cardinality and $\{\mathbb{L}_l : l = 1, \dots, L\}$ a partition of $\mathbb{L}$, i.e., $\mathbb{L} = \bigcup_l \mathbb{L}_l$ and $\mathbb{L}_l \cap \mathbb{L}_k = \varnothing$ for $l \ne k$. Let the scaled sets $\ell\mathbb{L}_l$, $l = 1, \dots, L$, be the level sets of the object V such that

$$V = \sum_{l=1}^{L} v_l \, \mathrm{I}_{\ell\mathbb{L}_l},$$

where $\mathrm{I}_{\ell\mathbb{L}_l}$ is the indicator function of $\ell\mathbb{L}_l$. The reconstructed object $\hat{V}$ from $\hat{\boldsymbol{X}}$ given in (19) also takes the same form:

$$\hat{V} = \sum_{l=1}^{L} \hat{v}_l \, \mathrm{I}_{\ell\mathbb{L}_l}.$$

To fix the undetermined constant, we may assume that $v_1 = \hat{v}_1$. Since

$$\|\Delta(V - \hat{V})\|_{2,2} = \mathbb{O}(\varepsilon)$$

by (20) and the gradient error occurs only on the boundaries of $\ell\mathbb{L}_l$ of cardinality $\mathbb{O}(\ell^{1-d})$, we have

$$|v_l - \hat{v}_l| = \mathbb{O}(\varepsilon \ell^{(d-1)/2}) \quad \text{for all } \ l.$$

Namely

$$\|V - \hat{V}\|_\infty = \mathbb{O}(\varepsilon \ell^{(d-1)/2})$$

and thus

$$\|V - \hat{V}\|_2 = \mathbb{O}(\varepsilon/\sqrt{\ell}).$$

4. Application: inverse scattering

In this section, we discuss the main application of the CJS formulation, that is, the TV-min for the inverse scattering problem.

A monochromatic wave u propagating in a heterogeneous medium characterized by a variable refractive index $n^2(\boldsymbol{r}) = 1 + v(\boldsymbol{r})$ is governed by the Helmholtz equation

$$\nabla^2 u(\boldsymbol{r}) + \omega^2(1 + v(\boldsymbol{r}))u(\boldsymbol{r}) = 0, \tag{21}$$

where v describes the medium inhomogeneities. For simplicity, the wave velocity is assumed to be unity and hence the wavenumber ω equals the frequency.

Consider the scattering of the incident plane wave

$$u^{\mathrm{i}}(\boldsymbol{r}) = e^{i\omega \boldsymbol{r}\cdot\hat{\boldsymbol{d}}}, \tag{22}$$

where $\hat{\boldsymbol{d}}$ is the incident direction. The scattered field $u^{\mathrm{s}} = u - u^{\mathrm{i}}$ then satisfies

$$\nabla^2 u^{\mathrm{s}} + \omega^2 u^{\mathrm{s}} = -\omega^2 v u \tag{23}$$

which can be written as the Lippmann–Schwinger equation:

$$u^{\mathrm{s}}(\boldsymbol{r}) = \omega^2 \int_{\mathbb{R}^d} v(\boldsymbol{r}')(u^{\mathrm{i}}(\boldsymbol{r}') + u^{\mathrm{s}}(\boldsymbol{r}'))G(\boldsymbol{r}, \boldsymbol{r}')\, d\boldsymbol{r}', \tag{24}$$

where G is the Green function for the operator $-(\nabla^2 + \omega^2)$.

The scattered field necessarily satisfies Sommerfeld's radiation condition

$$\lim_{r\to\infty} r^{(d-1)/2}\Big(\frac{\partial}{\partial r} - i\omega\Big)u^{\mathrm{s}} = 0,$$

reflecting the fact that the energy which is radiated from the sources represented by the right-hand side of (23) must scatter to infinity.

Thus the scattered field has the far-field asymptotic

$$u^{\mathrm{s}}(\boldsymbol{r}) = \frac{e^{i\omega|\boldsymbol{r}|}}{|\boldsymbol{r}|^{(d-1)/2}}\big(A(\hat{\boldsymbol{r}}, \hat{\boldsymbol{d}}, \omega) + \mathbb{O}(|\boldsymbol{r}|^{-1})\big), \quad \hat{\boldsymbol{r}} = \boldsymbol{r}/|\boldsymbol{r}|, \tag{25}$$

where A is the scattering amplitude and d the spatial dimension. In inverse scattering theory, the scattering amplitude is the measurement data determined by the formula [Colton and Kress 1998]

$$A(\hat{\boldsymbol{r}}, \hat{\boldsymbol{d}}, \omega) = \frac{\omega^2}{4\pi}\int d\boldsymbol{r}' v(\boldsymbol{r}')u(\boldsymbol{r}')e^{-i\omega\boldsymbol{r}'\cdot\hat{\boldsymbol{r}}},$$

which under the Born approximation becomes

$$A(\hat{\boldsymbol{r}}, \hat{\boldsymbol{d}}, \omega) = \frac{\omega^2}{4\pi}\int d\boldsymbol{r}' v(\boldsymbol{r}')e^{i\omega\boldsymbol{r}'\cdot(\hat{\boldsymbol{d}}-\hat{\boldsymbol{r}})}. \tag{26}$$

For simplicity of notation we consider the two-dimensional case in detail. Let $\mathbb{L} \subset \mathbb{Z}^2$ be a square sublattice of m integral points. Suppose that s point scatterers are located in a square lattice of spacing ℓ:

$$\ell\,\mathbb{L} = \{\boldsymbol{r}_j = \ell(p_1, p_2) : j = (p_1 - 1)\sqrt{m} + p_2,\ \boldsymbol{p} = (p_1, p_2) \in \mathbb{L}\}. \tag{27}$$

In the context of inverse scattering, it is natural to treat the size of the discrete ambient domain $\ell\,\mathbb{L}$ as being fixed independent of the resolution length ℓ. In particular, $m \sim \ell^{-2}$ in two dimensions.

First let us motivate the inverse scattering sampling scheme in the case of *point* scatterers and let v_j, $j = 1, \ldots, m$ be the strength of the scatterers. In other words, the total object is a sum of δ-functions:

$$v(\boldsymbol{r}) = \sum_j v_j \delta(\boldsymbol{r} - \boldsymbol{r}_j). \tag{28}$$

Let $\mathcal{S} = \{\boldsymbol{r}_{i_j} : j = 1, \ldots, s\}$ be the locations of the scatterers. Hence $v_j = 0$ for all $\boldsymbol{r}_j \notin \mathcal{S}$.

For point objects the scattering amplitude becomes a finite sum:

$$A(\hat{\boldsymbol{r}}, \hat{\boldsymbol{d}}, \omega) = \frac{\omega^2}{4\pi}\sum_{j=1}^{m} v_j e^{i\omega\boldsymbol{r}_j\cdot(\hat{\boldsymbol{d}}-\hat{\boldsymbol{r}})}. \tag{29}$$

In the Born approximation the exciting field $u(\boldsymbol{r}_j)$ is replaced by the incident field $u^{\mathrm{i}}(\boldsymbol{r}_j)$.

5. Sampling schemes

Next we review the sampling schemes introduced in [Fannjiang 2010] for *point* objects (28).

Let $\hat{\boldsymbol{d}}_l$ and $\hat{\boldsymbol{r}}_l$, $l = 1, \ldots, n$, be various incident and sampling directions for the frequencies ω_l, $l = 1, \ldots, n$, to be determined later. Define the measurement vector $Y = (y_l) \in \mathbb{C}^n$ with

$$y_l = \frac{4\pi}{\omega^2\sqrt{n}} A(\hat{\boldsymbol{r}}_l, \hat{\boldsymbol{d}}_l, \omega_l), \quad l = 1, \ldots, n. \tag{30}$$

The measurement vector is related to the *point* object vector $X = (v_j) \in \mathbb{C}^m$ by the sensing matrix $\boldsymbol{\Phi}$ as

$$Y = \boldsymbol{\Phi} X + E, \tag{31}$$

where E is the measurement error. Let θ_l and $\tilde{\theta}_l$ be the polar angles of $\hat{\boldsymbol{d}}_l$ and $\hat{\boldsymbol{r}}_l$, respectively. The (l, j)-entry of $\boldsymbol{\Phi} \in \mathbb{C}^{n\times m}$ is

$$n^{-1/2} e^{-i\omega_l \hat{\boldsymbol{r}}_l \cdot \boldsymbol{r}_j} e^{i\omega_l \hat{\boldsymbol{d}}_l \cdot \boldsymbol{r}_j} = n^{-1/2} e^{i\omega_l \ell (p_2(\sin\theta_l - \sin\tilde{\theta}_l) + p_1(\cos\theta_l - \cos\tilde{\theta}_l))}, \tag{32}$$

with $j = (p_1 - 1) + p_2$. Note that $\boldsymbol{\Phi}$ has unit 2-norm columns.

Let (ξ_l, ζ_l) be i.i.d. uniform random variables on $[-1, 1]^2$ and let ρ_l and ϕ_l be the polar coordinates as in

$$(\xi_l, \zeta_l) = \rho_l(\cos\phi_l, \sin\phi_l), \quad \rho_l = \sqrt{\xi_l^2 + \zeta_l^2} \le \sqrt{2}. \tag{33}$$

Let the sampling angle $\tilde{\theta}_l$ be related to the incident angle θ_l via

$$\theta_l + \tilde{\theta}_l = 2\phi_l + \pi, \tag{34}$$

and set the frequency ω_l to be

$$\omega_l = \frac{\Omega\rho_l}{\sqrt{2}\sin(\theta_l - \tilde{\theta}_l/2)}, \tag{35}$$

where Ω is a control parameter. Then the entries (32) of the sensing matrix $\boldsymbol{\Phi}$ under the condition

$$\Omega\ell = \pi/\sqrt{2} \tag{36}$$

are those of the random partial Fourier matrix

$$e^{i\pi(p_1\xi_l + p_2\zeta_l)}, \quad l = 1, \ldots, n, \quad p_1, p_2 = 1, \ldots, \sqrt{m}. \tag{37}$$

We consider two particular sampling schemes: The first employs multiple frequencies with the sampling angle always in the back-scattering direction, resembling the imaging geometry of synthetic aperture radar; the second employs only a single high frequency with the sampling angle in the forward direction, resembling the imaging geometry of X-ray tomography.

I. *Backward Sampling.* This scheme employs Ω-band limited probes, that is, $\omega_l \in [-\Omega, \Omega]$. This and (35) lead to the constraint

$$\left|\sin\frac{\theta_l-\tilde\theta_l}{2}\right| \ge \frac{\rho_l}{\sqrt{2}}. \tag{38}$$

A simple way to satisfy (34) and (38) is to set

$$\phi_l = \tilde\theta_l = \theta_l - \pi, \tag{39}$$

$$\omega_l = \frac{\Omega\rho_l}{\sqrt{2}}, \tag{40}$$

where $l = 1, \dots, n$. In this case the scattering amplitude is sampled exactly in the backward direction, resembling synthetic-aperture radar imaging. In contrast, exact forward sampling with $\tilde\theta_l = \theta_l$ almost surely violates the constraint (38).

II. *Forward Sampling.* This scheme employs single-frequency probes no less than Ω:

$$\omega_l = \gamma\Omega, \quad \gamma \ge 1, \quad l = 1, \dots, n. \tag{41}$$

We set

$$\theta_l = \phi_l + \arcsin\frac{\rho_l}{\gamma\sqrt{2}}, \tag{42}$$

$$\tilde\theta_l = \phi_l - \arcsin\frac{\rho_l}{\gamma\sqrt{2}}. \tag{43}$$

The difference between the incident angle and the sampling angle is

$$\theta_l - \tilde\theta_l = 2\arcsin\frac{\rho_l}{\gamma\sqrt{2}}, \tag{44}$$

which diminishes as $\gamma \to \infty$. In other words, in the high-frequency limit, the sampling angle approaches the incident angle, resembling X-ray tomography [Natterer 1986].

6. Piecewise constant objects

Next let us consider the following class of piecewise constant objects:

$$v(\boldsymbol{r}) = \sum_{\boldsymbol{p}\in\mathbb{L}} v_{\boldsymbol{p}}\, \mathrm{I}_{\square}\Big(\frac{\boldsymbol{r}}{\ell} - \boldsymbol{p}\Big), \quad \square = \big[-\tfrac{1}{2}, \tfrac{1}{2}\big]^2, \tag{45}$$

where $\mathrm{I}_{\square}$ is the indicator function of the unit square $\square$. As remarked in Section 1, we think of the pixelated v as a discrete approximation of some compactly supported function on $\mathbb{R}^2$ and having a well-defined limit as $\ell \to 0$. Set $V = (v_j) \in \mathbb{C}^m$ and $j = (p_1 - 1)\sqrt{m} + p_2$.

The discrete version of (26) is, however, not exactly the same as (29) since extended objects have different scattering properties from those of point objects.

The integral on the right-hand side of (26), modulo the discretization error, is

$$\int d\boldsymbol{r}' v(\boldsymbol{r}') e^{i\omega \boldsymbol{r}'\cdot(\hat{\boldsymbol{d}}-\hat{\boldsymbol{r}})} = \sum_{\boldsymbol{p}\in\mathbb{L}} v_{\boldsymbol{p}} e^{\mathrm{i}\omega\ell \boldsymbol{p}\cdot(\hat{\boldsymbol{d}}-\hat{\boldsymbol{r}})} \int e^{\mathrm{i}\omega \boldsymbol{r}'\cdot(\hat{\boldsymbol{d}}-\hat{\boldsymbol{r}})} \mathrm{I}_{\square}\Big(\frac{\boldsymbol{r}'}{\ell}\Big)\, \mathrm{d}\boldsymbol{r}'.$$

Now letting $\hat{\boldsymbol{d}}_l$, $\hat{\boldsymbol{r}}_l$, and ω_l, $l = 1, \ldots, n$, be selected according to Scheme I or II and substituting them into the above equation, we obtain

$$\begin{aligned}\int d\boldsymbol{r}' v(\boldsymbol{r}') e^{i\omega_l \boldsymbol{r}'\cdot(\hat{\boldsymbol{d}}_l-\hat{\boldsymbol{r}}_l)} &= \ell^2 \sum_{\boldsymbol{p}\in\mathbb{L}} v_{\boldsymbol{p}} e^{i\pi(p_1\xi_l+p_2\eta_l)} \int_{\square} e^{i\pi(x\xi_l+y\eta_l)}\, dx\, dy \\ &= \ell^2 \sum_{\boldsymbol{p}\in\mathbb{L}} v_{\boldsymbol{p}} e^{i\pi(p_1\xi_l+p_2\eta_l)} \frac{2\sin(\pi\xi_l/2)}{\pi\xi_l} \frac{2\sin(\pi\eta_l/2)}{\pi\eta_l}.\end{aligned}$$

Let

$$x_j = \ell^2 v_{\boldsymbol{p}}, \quad j = (p_1-1)\sqrt{m} + p_2,$$

and

$$y_l = \frac{4\pi}{\omega_l^2 \tilde{g}_l \sqrt{n}} A(\hat{\boldsymbol{r}}_l, \hat{\boldsymbol{d}}_l, \omega_l) + E_l, \quad l = 1, \ldots, n,$$

where

$$\tilde{g}_l = \frac{2\sin(\pi\xi_l/2)}{\pi\xi_l} \frac{2\sin(\pi\eta_l/2)}{\pi\eta_l}$$

and $E = (e_l)$ is the noise vector.

Define the sensing matrix $\boldsymbol{\Phi} = [\phi_{kp}]$ as

$$\phi_{kp} = \frac{1}{\sqrt{n}} e^{\mathrm{i}\pi(p_1\xi_k+p_2\eta_k)}, \quad p = (p_1-1)\sqrt{m} + p_2, \quad p_1, p_2 = 1, \ldots, \sqrt{m}. \tag{46}$$

Then the system above can be written in the same form as (31):

$$Y = \boldsymbol{\Phi} X + E, \quad X = (x_j), \tag{47}$$

where the data and error vectors have been modified as above to account for the differences between extended and point objects.

Our goal is to establish the performance guarantee for TV-min

$$\min \|Z\|_{\mathrm{TV}}, \quad \text{s.t.} \quad \|Y - \boldsymbol{\Phi} Z\|_2 \le \|E\|_2. \tag{48}$$

We accomplish this by transforming (48) into BPDN for CJS (13).

Define $\boldsymbol{X} = (X_1, X_2)$ with

$$(X_1, X_2) = \ell^2(\Delta_1 V, \Delta_2 V) \in \mathbb{C}^{m\times 2}.$$

Suppose the support of $\{v_{\boldsymbol{p}+\boldsymbol{e}_1}, v_{\boldsymbol{p}+\boldsymbol{e}_2}\}$ is contained in $\mathbb{L}$. Simple calculation yields

$$y_l = \frac{\ell^2}{\sqrt{n}} e^{i\pi\xi_l} \sum_{\boldsymbol{p}\in\mathbb{L}} v_{\boldsymbol{p}+\boldsymbol{e}_1} e^{\mathrm{i}\pi(p_1\xi_l+p_2\eta_l)} = \frac{\ell^2}{\sqrt{n}} e^{i\pi\eta_l} \sum_{\boldsymbol{p}\in\mathbb{L}} v_{\boldsymbol{p}+\boldsymbol{e}_2} e^{\mathrm{i}\pi(p_1\xi_l+p_2\eta_l)}$$

and thus

$$(e^{-\mathrm{i}\pi\xi_l} - 1) y_l = \frac{\ell^2}{\sqrt{n}} \sum_{\boldsymbol{p}\in\mathbb{L}} (v_{\boldsymbol{p}+\boldsymbol{e}_1} - v_{\boldsymbol{p}}) e^{\mathrm{i}\pi(p_1\xi_l+p_2\eta_l)}, \tag{49}$$

$$(e^{-\mathrm{i}\pi\eta_l} - 1) y_l = \frac{\ell^2}{\sqrt{n}} \sum_{\boldsymbol{p}\in\mathbb{L}} (v_{\boldsymbol{p}+\boldsymbol{e}_2} - v_{\boldsymbol{p}}) e^{\mathrm{i}\pi(p_1\xi_l+p_2\eta_l)}. \tag{50}$$

Define $\boldsymbol{Y} = (Y_1, Y_2)$ with

$$Y_1 = ((e^{-\mathrm{i}\pi\xi_l} - 1) y_l), \quad Y_2 = ((e^{-\mathrm{i}\pi\eta_l} - 1) y_l) \in \mathbb{C}^n,$$

and $\boldsymbol{E} = (E_1, E_2)$ with

$$E_1 = ((e^{-\mathrm{i}\pi\xi_l} - 1) e_l), \quad E_2 = ((e^{-\mathrm{i}\pi\eta_l} - 1) e_l) \in \mathbb{C}^n. \tag{51}$$

We rewrite (47) in the form

$$\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{X} + \boldsymbol{E}, \tag{52}$$

subject to the constraint

$$\Delta_1 X_2 = \Delta_2 X_1 \tag{53}$$

which is the discrete version of curl-free condition. This ensures that the reconstruction by line integration of $(v_{\boldsymbol{p}})$ from $\boldsymbol{X}$ is consistent (that is, path-independent).

To see that (53) is necessary and sufficient for the recovery of $(v_{\boldsymbol{p}})$, consider, for example, the notations in Figure 1 and suppose $v_{0,0}$ is known. By definition of the difference operators Δ_1 and Δ_2 we have

$$v_{1,0} = v_{0,0} + (\Delta_1 V)_{0,0}, \quad v_{0,1} = v_{0,0} + (\Delta_2 V)_{0,0}.$$

In general, we can determine $v_{\boldsymbol{p}}$, $\boldsymbol{p} \in \mathbb{L}$, iteratively from the relationship

$$v_{\boldsymbol{p}+\boldsymbol{e}_1} = v_{\boldsymbol{p}} + (\Delta_1 V)_{\boldsymbol{p}}, \quad v_{\boldsymbol{p}+\boldsymbol{e}_2} = v_{\boldsymbol{p}} + (\Delta_2 V)_{\boldsymbol{p}},$$

and the knowledge of V at any grid point. The path-independence in evaluating v_{p_1+1,p_2+1},

$$\begin{aligned} v_{p_1+1,p_2+1} &= v_{p_1,p_2} + (\Delta_1 V)_{p_1,p_2} + (\Delta_2 V)_{p_1+1,p_2} \\ &= v_{p_1,p_2} + (\Delta_2 V)_{p_1,p_2} + (\Delta_1 V)_{p_1,p_2+1}, \end{aligned}$$

implies that

$$(\Delta_2 V)_{p_1+1,p_2} - (\Delta_2 V)_{p_1,p_2} = (\Delta_1 V)_{p_1,p_2+1} - (\Delta_1 V)_{p_1,p_2},$$

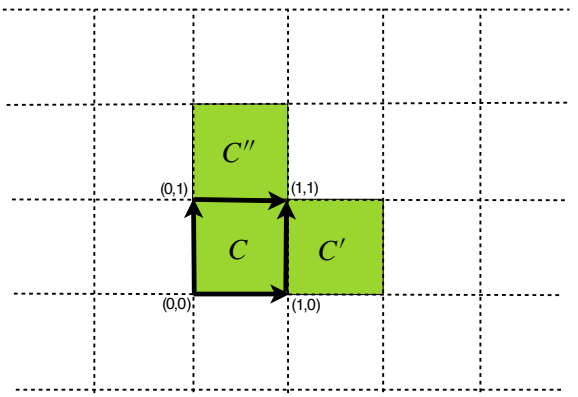


Figure 1. Consistency among cells C, C', and C''.

which is equivalent to (53).

Now (47) is equivalent to (52) with the constraint (53) provided that the value of V at (any) one grid point is known. The equivalence between the original TV-min (48) and the CJS formulation (13) with $\mathbf{\Phi}_j = \mathbf{\Phi}$ for all j, then hinges on the equivalence of their respective feasible sets which can be established under the assumption of Gaussian noise. When E in (47) is Gaussian noise, then so is $\boldsymbol{E}$, and vice versa, with variances precisely related to each other.

The random partial Fourier measurement matrix satisfies the RIP with $n = \mathbb{O}(s)$, up to a logarithmic factor [Candès et al. 2006], while its mutual coherence μ behaves like $\mathbb{O}(n^{-1/2})$ [Fannjiang et al. 2010]. Therefore (18) implies the sparsity constraint $s = \mathbb{O}(\sqrt{n})$ for the greedy approach which is more stringent than $s = \mathbb{O}(n)$ for the BPDN approach.

7. Conclusion

We have developed a general compressive sensing (CS) theory (Theorems 1 and 2) for constrained joint sparsity with multiple sensing matrices and obtained performance guarantees parallel to those for CS theory for a single measurement vector and matrix.

From the general theory we have derived 2-norm error bounds for the object and the gradient, independent of the ambient dimension, for total variation minimization (TV-min) and greedy estimates of piecewise constant objects.

In addition, the constrained joint sparsity (CJS) greedy algorithm can recover exactly the gradient support (that is, the edges of the object) leading to an improved 2-norm error bound. Although the CJS greedy algorithm needs a higher number of measurement data points than TV-min for Fourier measurements the incoherence property required is much easier to check, and is often the only practical way to

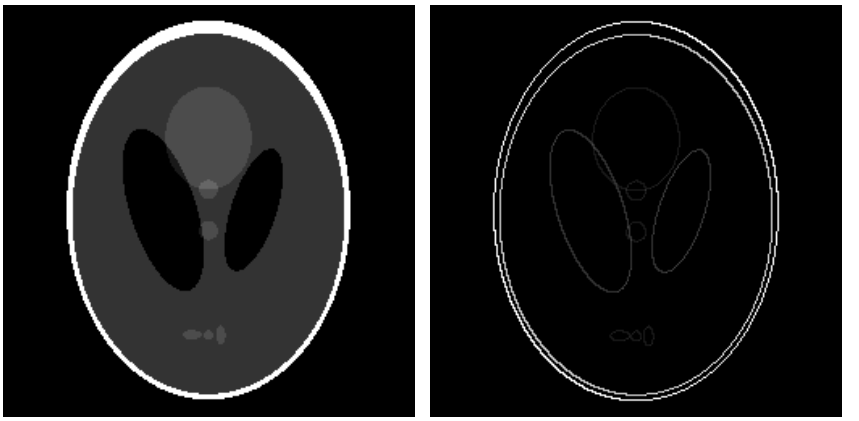


Figure 2. The original 256×256 Shepp–Logan phantom (left) and the Shepp–Logan phantom and the magnitudes of its gradient (right) with sparsity $s = 2{,}184$.

verify the restricted isometry property when the measurement matrix is not independently and identically distributed or Fourier.

We end by presenting a numerical example demonstrating the noise stability of TV-min. Efficient algorithms for TV-min denoising/deblurring exist [Beck and Teboulle 2009; Weiss et al. 2009]. We use the open source code ℓ_1-MAGIC for our simulation.

Figure 2 shows the 256×256 image of the Shepp–Logan phantom and the modulus of its gradient. Clearly the sparsity ($s = 2{,}184$) of the gradient is much smaller than that of the original image. We take 10,000 Fourier measurement data points for the ℓ_1-min (1) and TV-min (5) reconstructions.

Because the image is not sparse, ℓ_1-min reconstruction produces a poor result even in the absence of noise, shown in Figure 3. The relative error is 66.8% in

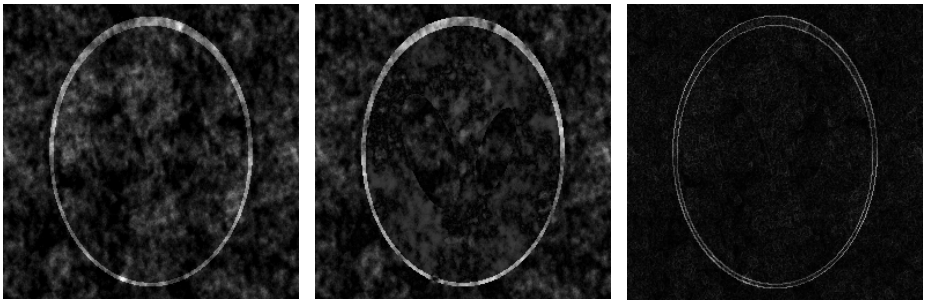


Figure 3. Noiseless ℓ_1-min reconstructed image (left) and the differences (middle) from the original image. The plot on the right is the gradient of the reconstructed image.

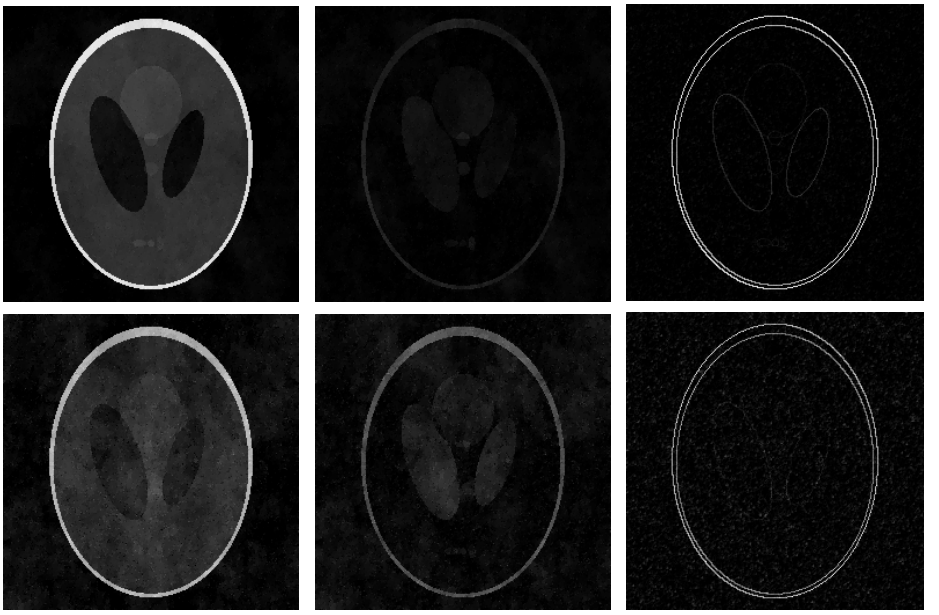


Figure 4. TV-reconstructed image with 5% (top left) and 10% (bottom left) and the respective differences (middle) from the original image. The plots on the right column are the magnitudes of the reconstructed image gradients.

the ℓ_2 norm and 72.8% in the TV norm. Only the outer boundary, which has the largest pixel values, is reasonably recovered.

Figure 4 shows the results of TV-min reconstruction in the presence of 5% (top) or 10% (bottom) noise. Evidently, the performance is greatly improved.

Appendix A: Proof of Theorem 1

The argument is patterned after [Candès 2008] with adaptation to the CJS setting.

Proposition 1. *We have*

$$|\Re\langle\varphi(\boldsymbol{Z}), \varphi(\boldsymbol{Z}')\rangle| \le \delta_{s+s'}\|\boldsymbol{Z}\|_{2,2}\|\boldsymbol{Z}'\|_{2,2}$$

for all $\boldsymbol{Z}, \boldsymbol{Z}'$ *supported on disjoint subsets* $T, T' \subset \{1, \dots, m\}$ *with* $|S| \le s$ *and* $|S'| \le s'$.

Proof. Without loss of generality, suppose that $\|\boldsymbol{Z}\|_{2,2} = \|\boldsymbol{Z}'\|_{2,2} = 1$. Since $\boldsymbol{Z} \perp \boldsymbol{Z}'$, $\|\boldsymbol{Z} \pm \boldsymbol{Z}'\|_{2,2}^2 = 2$. Hence we have from the RIP (14)

$$2(1 - \delta_{s+s'}) \le \|\varphi(\boldsymbol{Z} \pm \boldsymbol{Z}')\|_{2,2}^2 \le 2(1 + \delta_{s+s'}). \tag{54}$$

By the parallelogram identity and (54)

$$|\Re\langle\varphi(\boldsymbol{Z}),\varphi(\boldsymbol{Z}')\rangle| = \tfrac{1}{4}\left|\|\varphi(\boldsymbol{Z})+\varphi(\boldsymbol{Z}')\|_{2,2}^2 - \|\varphi(\boldsymbol{Z})-\varphi(\boldsymbol{Z}')\|_{2,2}^2\right| \le \delta_{s+s'},$$

which proves the proposition. □

By the triangle inequality and the fact that $\boldsymbol{X}$ is in the feasible set we have

$$\|\varphi(\hat{\boldsymbol{X}}-\boldsymbol{X})\|_{2,2} \le \|\varphi(\hat{\boldsymbol{X}})-\boldsymbol{Y}\|_{2,2} + \|\boldsymbol{Y}-\varphi(\boldsymbol{X})\|_{2,2} \le 2\varepsilon. \tag{55}$$

Set $\hat{\boldsymbol{X}} = \boldsymbol{X}+\boldsymbol{D}$ and decompose $\boldsymbol{D}$ into a sum of $\boldsymbol{D}_{S_0}, \boldsymbol{D}_{S_1}, \boldsymbol{D}_{S_2}, \ldots$, each of row sparsity at most s. Here S_0 corresponds to the locations of the s largest rows of $\boldsymbol{X}$; S_1 the locations of the s largest rows of $\boldsymbol{D}_{S_0^c}$; S_2 the locations of the next s largest rows of $\boldsymbol{D}_{S_0^c}$; and so on.

Step (i). Define the norm

$$\|\boldsymbol{Z}\|_{\infty,2} = \max_j \|\mathrm{row}_j(\boldsymbol{Z})\|_2.$$

For $j \ge 2$,

$$\|\boldsymbol{D}_{S_j}\|_{2,2} \le s^{1/2}\|\boldsymbol{D}_{S_j}\|_{\infty,2} \le s^{-1/2}\|\boldsymbol{D}_{S_{j-1}}\|_{2,2}$$

and hence

$$\sum_{j\ge2}\|\boldsymbol{D}_{S_j}\|_{2,2} \le s^{-1/2}\sum_{j\ge1}\|\boldsymbol{D}_{S_j}\|_{1,2} \le s^{-1/2}\|\boldsymbol{D}_{S_0^c}\|_{1,2}. \tag{56}$$

This yields, by the Cauchy–Schwarz inequality,

$$\|\boldsymbol{D}_{(S_0\cup S_1)^c}\|_{2,2} = \left\|\sum_{j\ge2}\boldsymbol{D}_{S_j}\right\|_{2,2} \le \sum_{j\ge2}\|\boldsymbol{D}_{S_j}\|_{2,2} \le s^{-1/2}\|\boldsymbol{D}_{S_0^c}\|_{1,2}. \tag{57}$$

Also we have

$$\begin{aligned}\|\boldsymbol{X}\|_{1,2} \ge \|\hat{\boldsymbol{X}}\|_{1,2} &= \|\boldsymbol{X}_{S_0}+\boldsymbol{D}_{S_0}\|_{1,2} + \|\boldsymbol{X}_{S_0^c}+\boldsymbol{D}_{S_0^c}\|_{1,2}\\ &\ge \|\boldsymbol{X}_{S_0}\|_{1,2} - \|\boldsymbol{D}_{S_0}\|_{1,2} - \|\boldsymbol{X}_{S_0^c}\|_{1,2} + \|\boldsymbol{D}_{S_0^c}\|_{1,2},\end{aligned}$$

which implies

$$\|\boldsymbol{D}_{S_0^c}\|_{1,2} \le 2\|\boldsymbol{X}_{S_0^c}\|_{1,2} + \|\boldsymbol{D}_{S_0}\|_{1,2}. \tag{58}$$

Note that $\|\boldsymbol{X}_{S_0^c}\|_{1,2} = \|\boldsymbol{X}-\boldsymbol{X}^{(s)}\|_{1,2}$ by definition. Applying (57), (58), and the Cauchy–Schwartz inequality to $\|\boldsymbol{D}_{S_0}\|_{1,2}$ gives

$$\|\boldsymbol{D}_{(S_0\cup S_1)^c}\|_{2,2} \le \|\boldsymbol{D}_{S_0}\|_{2,2} + 2e_0, \tag{59}$$

where $e_0 \equiv s^{-1/2}\|\boldsymbol{X}-\boldsymbol{X}^{(s)}\|_{1,2}$.

Step (ii). Define the inner product

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_{i,j} A_{ij}^* B_{ij}.$$

Observe that

$$\begin{aligned}
&\|\varphi(\boldsymbol{D}_{S_0 \cup S_1})\|_{2,2}^2 \\
&\quad = \langle \varphi(\boldsymbol{D}_{S_0 \cup S_1}), \varphi(\boldsymbol{D}) \rangle - \Big\langle \varphi(\boldsymbol{D}_{S_0 \cup S_1}), \sum_{j \geq 2} \varphi(\boldsymbol{D}_{S_j}) \Big\rangle \\
&\quad = \Re \langle \varphi(\boldsymbol{D}_{S_0 \cup S_1}), \varphi(\boldsymbol{D}) \rangle - \sum_{j \geq 2} \Re \langle \varphi(\boldsymbol{D}_{S_0 \cup S_1}), \varphi(\boldsymbol{D}_{S_j}) \rangle \\
&\quad = \Re \langle \varphi(\boldsymbol{D}_{S_0 \cup S_1}), \varphi(\boldsymbol{D}) \rangle - \sum_{j \geq 2} [\Re \langle \varphi(\boldsymbol{D}_{S_0}), \varphi(\boldsymbol{D}_{S_j}) \rangle + \Re \langle \varphi(\boldsymbol{D}_{S_1}), \varphi(\boldsymbol{D}_{S_j}) \rangle]. \quad (60)
\end{aligned}$$

From (55) and the RIP (14) it follows that

$$|\langle \varphi(\boldsymbol{D}_{S_0 \cup S_1}), \varphi(\boldsymbol{D}) \rangle| \leq \|\varphi(\boldsymbol{D}_{S_0 \cup S_1})\|_{2,2} \|\varphi(\boldsymbol{D})\|_{2,2} \leq 2\varepsilon \sqrt{1 + \delta_{2s}} \|\boldsymbol{D}_{S_0 \cup S_1}\|_{2,2}.$$

Moreover, it follows from Proposition 1 that

$$|\Re \langle \varphi(\boldsymbol{D}_{S_0}), \varphi(\boldsymbol{D}_{S_j}) \rangle| \leq \delta_{2s} \|\boldsymbol{D}_{S_0}\|_{2,2} \|\boldsymbol{D}_{S_j}\|_{2,2}, \quad (61)$$

$$|\Re \langle \varphi(\boldsymbol{D}_{S_1}), \varphi(\boldsymbol{D}_{S_j}) \rangle| \leq \delta_{2s} \|\boldsymbol{D}_{S_0}\|_{2,2} \|\boldsymbol{D}_{S_j}\|_{2,2}, \quad (62)$$

for $j \geq 2$. Since S_0 and S_1 are disjoint:

$$\|\boldsymbol{D}_{S_0}\|_{2,2} + \|\boldsymbol{D}_{S_1}\|_{2,2} \leq \sqrt{2} \sqrt{\|\boldsymbol{D}_{S_0}\|_{2,2}^2 + \|\boldsymbol{D}_{S_1}\|_{2,2}^2} = \sqrt{2} \|\boldsymbol{D}_{S_0 \cup S_1}\|_{2,2}.$$

Also by (60)–(62) and the RIP,

$$\begin{aligned}
(1 - \delta_{2s}) \|\boldsymbol{D}_{S_0 \cup S_1}\|_{2,2}^2 &\leq \|\varphi(\boldsymbol{D}_{S_0 \cup S_1})\|_{2,2}^2 \\
&\leq \|\boldsymbol{D}_{S_0 \cup S_1}\|_{2,2} \Big(2\varepsilon \sqrt{1 + \delta_{2s}} + \delta_{2s} \sum_{j \geq 2} \|\boldsymbol{D}_{S_j}\|_{2,2} \Big).
\end{aligned}$$

Therefore from (56) we obtain

$$\|\boldsymbol{D}_{S_0 \cup S_1}\|_{2,2} \leq \alpha \varepsilon + \rho s^{-1/2} \|\boldsymbol{D}_{S_0^c}\|_{1,2}, \quad \alpha = \frac{2\sqrt{1 + \delta_{2s}}}{1 - \delta_{2s}}, \quad \rho = \frac{\sqrt{2} \delta_{2s}}{1 - \delta_{2s}},$$

and moreover by (58) and the definition of e_0

$$\|\boldsymbol{D}_{S_0 \cup S_1}\|_{2,2} \leq \alpha \varepsilon + \rho \|\boldsymbol{D}_{S_0}\|_{2,2} + 2\rho e_0$$

after applying the Cauchy–Schwartz inequality to bound $\|\boldsymbol{D}_{S_0}\|_{1,2}$ by $s^{1/2} \|\boldsymbol{D}_{S_0}\|_{2,2}$. Thus

$$\|\boldsymbol{D}_{S_0 \cup S_1}\|_{2,2} \leq (1 - \rho)^{-1} (\alpha \varepsilon + 2\rho e_0)$$

if (14) holds.

Finally,

$$\begin{aligned}\|\boldsymbol{D}\|_{2,2} &\le \|\boldsymbol{D}_{S_0\cup S_1}\|_{2,2} + \|\boldsymbol{D}_{(S_0\cup S_1)^c}\|_{2,2} \le 2\|\boldsymbol{D}_{S_0\cup S_1}\|_{2,2} + 2e_0\\ &\le 2(1-\rho)^{-1}(\alpha\varepsilon + (1+\rho)e_0),\end{aligned}$$

which is the desired result.

Appendix B: Proof of Theorem 2

We prove the theorem by induction.

Let $\operatorname{supp}(\boldsymbol{X}) = \mathcal{S} = \{J_1, \dots, J_s\}$ and

$$X_{\max} = \|\operatorname{row}_{J_1}(\boldsymbol{X})\|_1 \ge \|\operatorname{row}_{J_2}(\boldsymbol{X})\|_1 \ge \cdots \ge \|\operatorname{row}_{J_s}(\boldsymbol{X})\|_1 = X_{\min}.$$

In the first step,

$$\begin{aligned}\sum_{j=1}^{d} |\Phi_{j,J_1}^* Y_j| &= \sum_{j=1}^{d} \left| X_{J_1 j} + X_{J_2 j}\Phi_{j,J_1}^*\Phi_{j,J_2} + \cdots + X_{J_s j}\Phi_{j,J_1}^*\Phi_{j,J_s} + \Phi_{j,J_1}^* E_j \right|\\ &\ge X_{\max} - X_{\max}(s-1)\mu_{\max} - \sum_j \|E_j\|_2. \qquad \text{(B.1)}\end{aligned}$$

On the other hand, for any $l \notin \operatorname{supp}(\boldsymbol{X})$,

$$\begin{aligned}\sum_{j=1}^{d} |\Phi_{j,l}^* Y_j| &= \sum_{j=1}^{d} \left| X_{J_1 j}\Phi_{j,l}^*\Phi_{j,J_1} + X_{J_2 j}\Phi_{j,l}^*\Phi_{j,J_2} + \cdots X_{J_s j}\Phi_{j,l}^*\Phi_{j,J_s} + \Phi_{j,l}^* E_j \right|\\ &\le X_{\max} s\mu_{\max} + \sum_j \|E_j\|_2. \qquad \text{(B.2)}\end{aligned}$$

Hence, if

$$(2s-1)\mu_{\max} + \frac{2\sum_j \|E_j\|_2}{X_{\max}} < 1,$$

then the right-hand side of (B.1) is greater than the right-hand side of (B.2) which implies that the first index selected by OMP must belong to $\operatorname{supp}(\boldsymbol{X})$.

To continue the induction process, we state the straightforward generalization of a standard uniqueness result for sparse recovery to the joint sparsity setting [Donoho et al. 2006, Lemma 5.3].

Proposition 2. *Let $\boldsymbol{Z} = \varphi(\boldsymbol{X})$ and $\boldsymbol{Y} = \boldsymbol{Z} + \boldsymbol{E}$. Let $\mathcal{S}^k$ be a set of k indices and let $\mathbf{A} \in \mathbb{C}^{n\times d}$ with $\operatorname{supp}(\mathbf{A}) = \mathcal{S}^k$. Define*

$$\boldsymbol{Y}' = \boldsymbol{Y} - \varphi(\mathbf{A}) \qquad \text{(B.3)}$$

and

$$\boldsymbol{Z}' = \boldsymbol{Z} - \varphi(\mathbf{A}).$$

Clearly, $\boldsymbol{Y}' = \boldsymbol{Z}' + \boldsymbol{E}$. *If* $\mathcal{S}^k \subsetneq \operatorname{supp}(\boldsymbol{X})$ *and the sparsity* s *of* $\boldsymbol{X}$ *satisfies* $2s < 1 + \mu_{\max}^{-1}$, *then* $\boldsymbol{Z}'$ *has a unique sparsest representation* $\boldsymbol{Z}' = \varphi(\boldsymbol{X}')$ *with the sparsity of* $\boldsymbol{X}'$ *at most* s.

Proposition 2 says that the selection of a column, followed by the formation of the residual signal, leads to a situation like before, where the ideal noiseless signal has no more representing columns than before, and the noise level is the same.

Suppose that the set $\mathcal{S}^k \subseteq \operatorname{supp}(\boldsymbol{X})$ of k distinct indices has been selected and that $\mathbf{A}$ in Proposition 2 solves the following least-squares problem:

$$\mathbf{A} = \arg\min \|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{B}\|_{2,2}, \qquad \text{s.t.} \qquad \operatorname{supp}(\boldsymbol{B}) \subseteq \mathcal{S}^k, \tag{B.4}$$

without imposing the constraint $\mathcal{L}$. This is equivalent to the concatenation $\mathbf{A} = [A_j]$ of d separate least-squares solutions

$$A_j = \arg\min_{B_j} \|Y_j - \boldsymbol{\Phi}_j B_j\|_2, \qquad \text{s.t.} \qquad \operatorname{supp}(B_j) \subseteq \mathcal{S}^k. \tag{B.5}$$

Let $\boldsymbol{\Phi}_{j,\mathcal{S}^k}$ be the column submatrix of $\boldsymbol{\Phi}_j$ indexed by the set $\mathcal{S}^k$. By (B.3) and (B.5), we have $\boldsymbol{\Phi}^*_{j,\mathcal{S}^k} Y'_j = 0$ for all j, which implies that no element of $\mathcal{S}^k$ gets selected at the $(k+1)$-st step.

In order to ensure that some element in $\operatorname{supp}(\boldsymbol{X}) \setminus \mathcal{S}^k$ gets selected at the $(k+1)$-st step we only need to repeat the calculation (B.1)–(B.2) to obtain the condition

$$(2s-1)\mu_{\max} + \frac{2\sum_j \|E_j\|_2}{\|X_{J_{k+1}}\|_1} < 1. \tag{B.6}$$

Since $\sum_j \|E_j\|_2 \le \sqrt{d}\|\boldsymbol{E}\|_{2,2} = \sqrt{d}\varepsilon$, (B.6) follows from

$$(2s-1)\mu_{\max} + \frac{2\sqrt{d}\varepsilon}{X_{\min}} < 1, \tag{B.7}$$

which is the same as (18) and allows us to apply Proposition 2 repeatedly.

By the s-th step, all elements of the support set are selected and by the nature of the least-squares solution the 2-norm of the residual is at most ε. Thus the stopping criterion is met and the iteration stops after s steps.

On the other hand, it follows from the calculation

$$\begin{aligned}
\sum_j \|Y'_j\|_2 &\ge \sum_{j=1}^{d} \left|\Phi^*_{j,J_{k+1}} Y'_j\right| = \sum_j \left|X_{J_{k+1}j} + \sum_{i=k+2}^{s} X_{J_i i}\Phi^*_{j,J_{k+1}}\Phi_{i,J_i} + \Phi^*_{j,J_{k+1}} E_j\right| \\
&\ge \|\operatorname{row}_{J_{k+1}}(\boldsymbol{X})\|_1 - \mu_{\max}(s-k-1)\|\operatorname{row}_{J_{k+2}}(\boldsymbol{X})\|_1 - \sum_j \|E_j\|_2 \\
&\ge (1 - \mu_{\max}(s-k-1))\|\operatorname{row}_{J_{k+1}}(\boldsymbol{X})\|_1 - \sum_j \|E_j\|_2,
\end{aligned}$$

and (B.7) (equivalently, $X_{\min}(1-\mu_{\max}(2s-1)) > 2\sqrt{d}\varepsilon$) that $\|\boldsymbol{Y}\|_{1,2} > \sqrt{d}\varepsilon$ for $k = 0, 1, \ldots, s-1$. Thus the iteration does not stop until $k = s$.

Since $\hat{\boldsymbol{X}}$ is the solution of the least-squares problem (19), we have

$$\|\boldsymbol{Y} - \boldsymbol{\Phi}\hat{\boldsymbol{X}}\|_{2,2} \le \|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{X}\|_{2,2} \le \varepsilon$$

and

$$\|\boldsymbol{\Phi}(\boldsymbol{X} - \hat{\boldsymbol{X}})\|_{2,2}^2 \le 2\|\boldsymbol{Y} - \boldsymbol{\Phi}\boldsymbol{X}\|_{2,2}^2 + 2\|\boldsymbol{Y} - \boldsymbol{\Phi}\hat{\boldsymbol{X}}\|_{2,2}^2 \le 2\varepsilon^2,$$

which implies

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_{2,2} \le \sqrt{2}\varepsilon/\lambda_{\min},$$

where

$$\lambda_{\min} = \min_j\{\text{the } s\text{-th singular value of the column submatrix of } \boldsymbol{\Phi}_j \text{ indexed by } \mathcal{S}\}.$$

The desired error bound (20) can now be obtained from the following result [Donoho et al. 2006, Lemma 2.2].

Proposition 3. *Suppose $s < 1 + \mu(\boldsymbol{\Phi}_j)^{-1}$. Every $m \times s$ column submatrix of $\boldsymbol{\Phi}_j$ has the s-th singular value bounded below by $\sqrt{1-\mu(\boldsymbol{\Phi}_j)(s-1)}$.*

By Proposition 3, $\lambda_{\min} \ge \sqrt{1-\mu_{\max}(s-1)}$ and thus

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_{2,2} \le \frac{\sqrt{2}\varepsilon}{\sqrt{1-\mu_{\max}(s-1)}}.$$

Acknowledgements

I thank Stan Osher and Justin Romberg for the suggestion of publishing this note at the Institute for Pure and Applied Mathematics workshop "Challenges in Synthetic Aperture Radar," February 6–10, 2012. I thank the anonymous referees and Deanna Needell for pointing out the reference [Needell and Ward 2012] which helps me appreciate more deeply the strengths and weaknesses of my approach. I am grateful to Wenjing Liao for preparing Figures 2–4.

References

[Beck and Teboulle 2009] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems", *IEEE Trans. Image Process.* **18**:11 (2009), 2419–2434.

[Candès 2008] E. J. Candès, "The restricted isometry property and its implications for compressed sensing", *C. R. Acad. Sci. Paris Sér. I Math.* **346**:9-10 (2008), 589–592.

[Candès and Plan 2009] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization", *Ann. Stat.* **37**:5A (2009), 2145–2177.

[Candès and Tao 2005] E. J. Candès and T. Tao, "Decoding by linear programming", *IEEE Trans. Inf. Theory* **51**:12 (2005), 4203–4215.

[Candès et al. 2006] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information", *IEEE Trans. Inf. Theory* **52**:2 (2006), 489–509.

[Chambolle and Lions 1997] A. Chambolle and P.-L. Lions, "Image recovery via total variation minimization and related problems", *Numer. Math.* **76**:2 (1997), 167–188.

[Chan and Shen 2005] T. F. Chan and J. Shen, *Image processing and analysis: variational, PDE, wavelet, and stochastic methods*, Society for Industrial and Applied Mathematics, Philadelphia, 2005.

[Chen and Huo 2006] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors", *IEEE Trans. Signal Process.* **54**:12 (2006), 4634–4643.

[Chen et al. 2001] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit", *SIAM Rev.* **43**:1 (2001), 129–159.

[Cheung 1998] W.-S. Cheung, "Discrete Poincaré-type inequalities", *Tamkang J. Math.* **29**:2 (1998), 145–153.

[Claerbout and Muir 1973] J. F. Claerbout and F. Muir, "Robust modeling with erratic data", *Geophys.* **38**:5 (1973), 826–844.

[Colton and Kress 1998] D. Colton and R. Kress, *Inverse acoustic and electromagnetic scattering theory*, 2nd ed., Applied Mathematical Sciences **93**, Springer, Berlin, 1998. 3rd ed. in 2013.

[Cotter et al. 2005] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors", *IEEE Trans. Signal Process.* **53**:7 (2005), 2477–2488.

[Davis et al. 1997] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations", *Constr. Approx.* **13**:1 (1997), 57–98.

[Donoho 2006] D. L. Donoho, "Compressed sensing", *IEEE Trans. Inf. Theory* **52**:4 (2006), 1289–1306.

[Donoho and Elad 2003] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization", *Proc. Nat. Acad. Sci. USA* **100**:5 (2003), 2197–2202.

[Donoho and Huo 2001] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition", *IEEE Trans. Inf. Theory* **47**:7 (2001), 2845–2862.

[Donoho et al. 2006] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise", *IEEE Trans. Inf. Theory* **52**:1 (2006), 6–18.

[Fannjiang 2010] A. C. Fannjiang, "Compressive inverse scattering, II: Multi-shot SISO measurements with born scatterers", *Inverse Probl.* **26**:3 (2010), Article ID #035009.

[Fannjiang et al. 2010] A. C. Fannjiang, T. Strohmer, and P. Yan, "Compressed remote sensing of sparse objects", *SIAM J. Imaging Sci.* **3**:3 (2010), 595–618.

[Golub and Van Loan 1996] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[Natterer 1986] F. Natterer, *The mathematics of computerized tomography*, B. G. Teubner, Stuttgart, 1986.

[Needell and Ward 2012] D. Needell and R. Ward, "Stable image reconstruction using total variation minimization", preprint, 2012. arXiv 1202.6429v6

[Patel et al. 2012] V. M. Patel, R. Maleh, A. C. Gilbert, and R. Chellappa, "Gradient-based image recovery methods from incomplete Fourier measurements", *IEEE Trans. Image Process.* **21**:1 (2012), 94–105.

[Pati et al. 1993] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition", pp. 40–44 in *Conference record of the Twenty-Seventh Asilomar Conference on Signals, Systems and Computers* (Pacific Grove, CA, 1993), vol. 1, edited by A. Singh, IEEE Computer Society Press, Los Alamitos, CA, 1993.

[Rudin and Osher 1994] L. I. Rudin and S. Osher, "Total variation based image restoration with free local constraints", *Proc. IEEE ICIP* **1** (1994), 31–35.

[Rudin et al. 1992] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms", *Physica D* **60**:1–4 (1992), 259–268.

[Taylor et al. 1979] H. L. Taylor, S. C. Banks, and J. F. McCoy, "Deconvolution with the $\ell - 1$ norm", *Geophys.* **44**:1 (1979), 39–52.

[Tropp 2004] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation", *IEEE Trans. Inf. Theory* **50**:10 (2004), 2231–2242.

[Tropp et al. 2006] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation, I: Greedy pursuit", *Signal Process.* **86**:3 (2006), 572–588.

[Weiss et al. 2009] P. Weiss, L. Blanc-Féraud, and G. Aubert, "Efficient schemes for total variation minimization under constraints in image processing", *SIAM J. Sci. Comput.* **31**:3 (2009), 2047–2080.

Received 11 May 2012. Revised 17 Sep 2012. Accepted 12 Nov 2012.

ALBERT FANNJIANG: fannjiang@math.ucdavis.edu
Department of Mathematics, University of California, Davis, One Shields Ave., Davis, CA 95616-8633, United States

MATHEMATICS AND MECHANICS OF COMPLEX SYSTEMS
Vol. 1, No. 1, 2013
dx.doi.org/10.2140/memocs.2013.1.105

ON THE THEORY OF DIFFUSION AND SWELLING IN FINITELY DEFORMING ELASTOMERS

GARY J. TEMPLET AND DAVID J. STEIGMANN

The role of a relaxed local intermediate configuration associated with free swelling is examined in the context of diffusion of a liquid in an isotropic elastomer. It is found that this configuration is energetically optimal if the free-energy function of the polymer-liquid gel is polyconvex. Further aspects of the general theory of diffusion in elastomers are also discussed.

1. Introduction

We study the modern continuum theory for diffusion of an incompressible liquid in an incompressible elastomer [Treloar 1975; Truesdell 1962; Adkins 1964; Weitsman 1987; Shi et al. 1981; Rajagopal 2003; Baek and Srinivasa 2004; Prasad and Rajagopal 2006]. In addition to examining the structure of constitutive equations and initial-boundary-value problems, we study the role of a local intermediate configuration induced by *free swelling*, defined as dilation of the polymer network in the absence of stress due to the presence of an infused liquid. We show, for isotropic elastomers, that the local free-swelling deformation minimizes energy if the free-energy function for the polymer network satisfies the condition of polyconvexity. In fact we are able to weaken this requirement to rank-one convexity. Accordingly the decomposition of the deformation into elastic and swelling deformations, assumed *a priori* in current theories of diffusion in polymers [Pence and Tsai 2005; Hong et al. 2008; Duda et al. 2010; Chester and Anand 2011; Duda et al. 2011], is meaningful for free-energy functions commonly used to model elastomers. Moreover, because of the basic role played by polyconvexity in Ball's landmark existence theory [1977] for conventional elasticity, our results indicate that an extension of that theory to accommodate diffusion should be feasible.

We use standard notation such as $\boldsymbol{A}^t$, $\boldsymbol{A}^{-1}$, $\boldsymbol{A}^*$, $\operatorname{tr}\boldsymbol{A}$ and J_A. These are respectively the transpose, the inverse, the cofactor, the trace and the determinant of a tensor $\boldsymbol{A}$, regarded as a linear transformation from a three-dimensional vector space to itself, the latter being identified with the translation space of the usual three-dimensional Euclidean point space. We also use Sym^+ to denote the linear

Communicated by Pierre Seppecher.
MSC2010: 74FXX.
Keywords: diffusion, polymers, polyconvexity.

space of positive symmetric tensors and Orth^+ to denote the group of rotation tensors. We denote the standard tensor product of vectors by interposing the symbol $\otimes$. The Euclidean inner product of tensors $\boldsymbol{A}$ and $\boldsymbol{B}$ is $\boldsymbol{A}\cdot\boldsymbol{B}=\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^t)$, and the associated norm is $|\boldsymbol{A}|=\sqrt{\boldsymbol{A}\cdot\boldsymbol{A}}$. Lastly, D is used to denote the gradient with respect to position $\boldsymbol{x}$ in a reference configuration and $F_{\boldsymbol{A}}$ stands for the tensor-valued derivative of a scalar-valued function $F(\boldsymbol{A})$.

Section 2 contains an outline of the general theory, including the diffusive balance law, the swelling constraint and the definition of dissipation. The latter is shown to lead to a constitutive structure in which the stress and chemical potential are determined, modulo a Lagrange multiplier, by a free-energy function that depends on the deformation gradient and the concentration of diffusant. The role played by polyconvexity in this theory is examined and used to analyze the free-swelling problem, which forms the basis of a decomposition of the deformation into elastic and swelling components. The general constitutive equation for the mobility tensor is also examined and shown to satisfy all invariance and symmetry requirements. The relatively simple theory for two-dimensional problems is summarized in Section 3, and an Appendix is included in which necessary and sufficient conditions for polyconvexity in respect of isotropic materials are discussed.

2. Three-dimensional theory

2.1. ***Basic equations and inequalities.*** We outline the basic theory for diffusion in elastomers undergoing finite deformations. For the most part our model may be viewed as a specialization of a thermodynamical theory presented in [Gurtin et al. 2010]. Here, however, thermal effects are suppressed, and the model reverts to conventional finite-elasticity theory in the absence of diffusion. The effects of inertia are also suppressed, in deference to the fact that the associated time scales typically differ markedly from those associated with the effects of diffusion. Further, we do not take chemical reactions into account, although it is possible to do so in a manner that is compatible with the present framework.

Let $c(\boldsymbol{x},t)$ be the concentration of diffusant, where t is the time and $\boldsymbol{x}$ is position in a fixed reference placement κ. For example, it is common to take the concentration to be the number of molecules of diffusant per unit volume of the reference configuration adopted for the dry elastomer [Chester and Anand 2011], the latter then being identified with κ. Alternatively, in [Gurtin et al. 2010] the concentration is defined to be the mass of diffusant per unit reference volume. The ratio of the second of these definitions to the first is the mass of a molecule of diffusant, a constant. Accordingly, these definitions are equivalent. A third definition of c, equivalent to these and more convenient for our purposes, is introduced in the next subsection.

For any subvolume π of κ we have the diffusive balance

$$\frac{d}{dt}\int_\pi c\,dv = -\int_{\partial\pi} \boldsymbol{m}\cdot\boldsymbol{n}\,da \quad \text{for all } \pi\subset\kappa, \tag{1}$$

where $\boldsymbol{m}(\boldsymbol{x},t)$ is the flux of diffusant and $\partial\pi$ is the piecewise smooth boundary, with exterior unit normal $\boldsymbol{n}$, of the region $\pi\subset\kappa$. The inequality $\boldsymbol{m}\cdot\boldsymbol{n}>0$ (resp., <0) corresponds to local transport of diffusant out of (resp., into) π; transport vanishes when $\boldsymbol{m}\cdot\boldsymbol{n}$ vanishes. The flux $\boldsymbol{m}$ arises entirely from diffusant transport, resulting in a change of mass of the gel (polymer-diffusant mixture) associated with κ. Consistent with this interpretation, we suppress bulk production of diffusant, which would otherwise require the addition of a volumetric source term to (1).

The local form of (1) is

$$\dot{c}+\operatorname{Div}\boldsymbol{m}=0 \quad \text{in } \kappa, \tag{2}$$

where the superposed dot stands for $\partial/\partial t$ at fixed $\boldsymbol{x}$ and Div is the divergence with respect to $\boldsymbol{x}$.

We assume the existence of a free-energy function $\Psi(\boldsymbol{F},c)$, yielding the energy of the gel per unit volume of κ, where $\boldsymbol{F}$, with $J_F>0$, is the gradient of a deformation function $\boldsymbol{y}=\boldsymbol{\chi}(\boldsymbol{x},t)$ yielding the position at time t of a point associated with position $\boldsymbol{x}\in\kappa$. Thus $\boldsymbol{\chi}$ represents the motion of the infused elastomer. For convenience we suppress reference to a possible explicit dependence on $\boldsymbol{x}$ in the notation for the free-energy function.

The power supplied to the arbitrary part π of the body, assuming no body forces and no volumetric sources of diffusant, is

$$P(\pi,t)=\int_{\partial\pi}(\boldsymbol{p}\cdot\dot{\boldsymbol{y}}-\boldsymbol{q}\cdot\boldsymbol{n})\,da, \tag{3}$$

where

$$\boldsymbol{p}=\boldsymbol{P}\boldsymbol{n}, \tag{4}$$

in which $\boldsymbol{P}$ is the Piola stress, is the traction acting on the boundary, and $\boldsymbol{q}$ is the rate of energy supply due to the flux of diffusant; that is, the rate at which the energy content of the gel changes due to the transport of diffusant across $\partial\pi$. The sign appearing before $\boldsymbol{q}$ conforms to the convention adopted in (1). With this interpretation of $\boldsymbol{q}$, it follows on logical grounds alone that

$$\boldsymbol{q}=\mu\boldsymbol{m}, \tag{5}$$

where μ, the *chemical potential*, is the energy conducted by the diffusant into the gel; ie., the energy per unit of diffusant concentration. The same conclusion follows from a rigorous thermodynamic treatment of diffusion in which the chemical potential is admitted as an independent variable [Gurtin et al. 2010].

Following [Gurtin et al. 2010] we suppose that $\boldsymbol{m}$ depends constitutively on the list $\{\boldsymbol{F}, c, D\mu\}$, while $\boldsymbol{P}$ depends constitutively on $\{\boldsymbol{F}, c\}$. Further, if the net force on an arbitrary subvolume $\pi \subset \kappa$ vanishes, as we assume, then the stress satisfies

$$\operatorname{Div} \boldsymbol{P} = \boldsymbol{0} \quad \text{in } \kappa. \tag{6}$$

The dissipation is denoted by $\mathcal{D}(\pi, t)$ and defined as the difference between the power supplied to π and the rate of change of the total energy in π. Thus,

$$\mathcal{D}(\pi, t) = P(\pi, t) - \frac{d}{dt}\int_\pi \Psi \, dv. \tag{7}$$

It is assumed to be non-negative for every subregion; i.e.,

$$\mathcal{D}(\pi, t) \geq 0. \tag{8}$$

Using (4), applying the divergence theorem and supposing all fields to be smooth in π yields

$$\int_{\partial\pi} \boldsymbol{p} \cdot \dot{\boldsymbol{y}} \, da = \int_\pi \operatorname{Div}(\boldsymbol{P}^t \dot{\boldsymbol{y}}) \, dv, \tag{9}$$

in which

$$\operatorname{Div}(\boldsymbol{P}^t \dot{\boldsymbol{y}}) = \boldsymbol{P} \cdot \dot{\boldsymbol{F}}, \tag{10}$$

by virtue of (4) and (6). Accordingly, (2), (3), (7) and (10), together with $\operatorname{Div}(\mu \boldsymbol{m}) = \mu \operatorname{Div} \boldsymbol{m} + \boldsymbol{m} \cdot D\mu$, furnish

$$\mathcal{D}(\pi, t) = \int_\pi (\boldsymbol{P} \cdot \dot{\boldsymbol{F}} + \mu \dot{c} - \dot{\Psi} - \boldsymbol{m} \cdot D\mu) \, dv, \tag{11}$$

and (8) then yields the local restriction

$$(\Psi_{\boldsymbol{F}} - \boldsymbol{P}) \cdot \dot{\boldsymbol{F}} + (\Psi_c - \mu)\dot{c} + \boldsymbol{m} \cdot D\mu \leq 0. \tag{12}$$

2.2. ***Volume decomposition, the swelling constraint and the basic constitutive structure.*** Most workers adopt the assumption that the volume of the gel is simply the sum of the volumes of the elastomer and the liquid diffusant. We invoke this assumption here and thus conclude that for any $\pi \subset \kappa$,

$$\int_\pi J_F \, dv = \int_\pi (J_{Fe} + J_{Fd}) \, dv, \tag{13}$$

where J_{Fe} and J_{Fd} respectively are the current volume of elastomer and diffusant per unit reference volume. Accordingly,

$$J_F = J_{Fe} + J_{Fd} \quad \text{in} \quad \kappa. \tag{14}$$

Here and henceforth we assume that sufficient liquid is available to support dilation. The alternative, corresponding to an unsaturated condition [Deng and Pence 2010;

Rivlin 1977], entails the global constraint that the total dilation reduce to the sum of the elastomer and liquid volumes. This carries a *uniform* reaction pressure that must be added to the expression for the (Cauchy) stress (see previous references).

Following [Chester and Anand 2011], here we suppose the contribution of the diffusant to arise solely from its transport into κ. Thus,

$$J_{Fd} = c_d v_d, \tag{15}$$

where c_d is the number of molecules of diffusant per unit volume of κ and v_d is the volume of a molecule of diffusant. The latter is presumed to be constant if the diffusant is incompressible. Accordingly, J_{Fd} and c_d furnish equivalent definitions of concentration in this case. The assumption that the elastomer is essentially incompressible is also virtually ubiquitous. This implies that $J_{Fe} = 1$ and hence the *swelling constraint* [Chester and Anand 2011]

$$J_F = 1 + c, \quad \text{where } c = c_d v_d, \tag{16}$$

applicable to incompressible elastomers infused with an incompressible diffusant. Because c is defined as the product of non-negative scalars, it follows from (16) that

$$J_F \geq 1. \tag{17}$$

The simplicity of (16) justifies our choice of c — the volume of diffusant present in the gel per unit volume of dry elastomer — as the measure of concentration. Accordingly, for a gel consisting of an incompressible elastomer infused with an incompressible diffusant, the deformation and concentration are not independent.

To account for this interdependence in (12), we use (16) in the form $\dot{c} = \dot{J}_F$, together with $\dot{J}_F = \boldsymbol{F}^* \cdot \dot{\boldsymbol{F}}$, obtaining

$$[\Psi_{\boldsymbol{F}} - \boldsymbol{P} + (\Psi_c - \mu)\boldsymbol{F}^*] \cdot \dot{\boldsymbol{F}} + \boldsymbol{m} \cdot D\mu \leq 0, \tag{18}$$

in which $\dot{\boldsymbol{F}}$ is unrestricted and, importantly, $\boldsymbol{F}$ and c are to be regarded as being independent when computing the derivatives of Ψ. Thus Ψ is regarded as a smooth extension of the free-energy function from the nine-dimensional manifold in $\mathbb{R}^{10}$ defined by the constraint (16). The extended function is thus defined and differentiable for all $\{\boldsymbol{F}, c\} \in \mathbb{R}^{10}$. Accordingly, our constitutive hypotheses may be combined with a standard argument (e.g. [Liu 2002]) to conclude from (18) that

$$\boldsymbol{P} = \Psi_{\boldsymbol{F}} - q\boldsymbol{F}^*, \quad \text{where } q = \mu - \Psi_c, \tag{19}$$

together with

$$\boldsymbol{m} \cdot D\mu \leq 0. \tag{20}$$

Because the argument is purely local in space and time, q may vary with $\boldsymbol{x}$ and t. Accordingly, it is an additional field to be determined in the course of the analysis

of the initial-boundary-value problem, which includes the constraint (16). The dissipation in any subvolume is then given by

$$\mathcal{D}(\pi, t) = -\int_{\pi} \boldsymbol{m} \cdot D\mu \, dv \tag{21}$$

and satisfies inequality (8). The connection $\boldsymbol{P} = \boldsymbol{T}\boldsymbol{F}^*$, where $\boldsymbol{T}$ is the Cauchy stress, implies that q is mechanically indistinguishable from a pressure. Equivalent results are derived in [Chester and Anand 2011], albeit by a rather circuitous procedure.

Our hypotheses, together with an important result due to Gurtin [2000], yield the general constitutive structure compatible with (20) in the form

$$\boldsymbol{m} = \boldsymbol{M}(\boldsymbol{F}, c, D\mu) D\mu, \tag{22}$$

in which the (2nd order) *mobility tensor* $\boldsymbol{M}$ satisfies

$$D\mu \cdot \boldsymbol{M}(\boldsymbol{F}, c, D\mu) D\mu \leq 0. \tag{23}$$

In this work we restrict attention to the practically important case in which the mobility tensor is insensitive to $D\mu$. Its symmetric part is therefore non-positive, and is negative definite if (23) holds as a strict inequality for all non-zero $D\mu$, so that $\boldsymbol{m} \cdot D\mu = 0$ if and only if $D\mu = \boldsymbol{0}$. Henceforth we assume this to be the case.

Thus the problem is to solve the five equations consisting of the diffusive balance (2), the equilibrium equation (6), and the swelling constraint (16) for the five variables in the list $\{\boldsymbol{\chi}, c, q\}$. Standard boundary data entail the specification of $\boldsymbol{p}$ or $\boldsymbol{y}$, and μ or $\boldsymbol{m} \cdot \boldsymbol{n}$, on (possibly different) complementary parts of $\partial\kappa$. A distribution of chemical potential may be specified in the domain κ at an initial time, and the constitutive equation $(19)_2$ may be used to generate an initial distribution of concentration in terms of the specified chemical potential field and the (unknown) initial deformation gradient and Lagrange multiplier fields; the swelling constraint, the equilibrium equation and associated boundary data may then be used to determine initial deformation and Lagrange multiplier fields, and the diffusive balance (2) then used to obtain the initial distribution of the time derivative of concentration. This information is used to advance the concentration in time, and the procedure then repeated for a specified time interval to generate the evolving spatial distribution of the list $\{\boldsymbol{\chi}, c, q\}$.

Frame invariance of material response requires that the free-energy function and the mobility tensor depend on $\boldsymbol{F}$ via the right Cauchy–Green deformation tensor, or equivalently via the right stretch tensor [Gurtin et al. 2010]. The restriction on the free-energy function is equivalent to the symmetry of the Cauchy stress [Rajagopal and Srinivasa 2012] and so renders the moment-of-momentum balance redundant whenever the linear momentum balance is satisfied.

We apply (20) for $\pi = \kappa$ and suppose traction $\boldsymbol{p}$ to be assigned on a part of the boundary and position to be assigned on the complementary part. Suppose the diffusant flux $\boldsymbol{m} \cdot \boldsymbol{n}$ and chemical potential μ vanish on complementary parts of $\partial\kappa$. If the traction field is conservative in the sense that the power is expressible as

$$P(\kappa, t) = \frac{d}{dt} L, \tag{24}$$

where L is a suitable load potential, then

$$\frac{d}{dt} E(\kappa, t) + \mathcal{D}(\kappa, t) = 0, \tag{25}$$

where $\mathcal{D}(\kappa, t)$ is given by (21) with $\pi = \kappa$, and

$$E = \int_\kappa \Psi \, dv - L \tag{26}$$

is the total potential energy of the gel. Inequality (8) then implies that the potential energy is dissipated.

Specifically, the energy associated with the state $\{\boldsymbol{\chi}(\boldsymbol{x}, t), c(\boldsymbol{x}, t)\}$ is no larger than that associated with state $\{\boldsymbol{\chi}(\boldsymbol{x}, t-\tau), c(\boldsymbol{x}, t-\tau)\}$ for any $\tau > 0$. Therefore the energy of a state at fixed time t is optimal relative to any trajectory tending to that state. Consequently, if $\{\boldsymbol{\chi}(\boldsymbol{x}, t), c(\boldsymbol{x}, t)\}$ is a trajectory starting at $\{\boldsymbol{\chi}(\boldsymbol{x}, t_0), c(\boldsymbol{x}, t_0)\}$ and tending to $\{\boldsymbol{\chi}_\infty(\boldsymbol{x}), c_\infty(\boldsymbol{x})\}$ a.e. as $t \to \infty$, then $E_\infty \leq E_0$, where E_0 and E_∞ are the values of E at times t_0 and infinity; asymptotically stable states thus minimize the potential energy. Moreover, if the decay of energy is *gradual*, in the sense that $E(t-\tau) \to E(t) + o(\tau)$ as $t \to \infty$, then $dE/dt \to 0$ as $t \to \infty$. In this case we have from (21) that $D\mu_\infty(\boldsymbol{x}) = \boldsymbol{0}$, where μ_∞ is the large-time limit of the chemical potential. Indeed, the alternative implies, via (23), that there is some $\bar{\boldsymbol{x}} \in \kappa$ where $D\mu_\infty \cdot \boldsymbol{M}(\boldsymbol{F}_\infty, \sigma_\infty) D\mu_\infty < 0$. By continuity, the latter inequality obtains in a subvolume of non-zero measure, and a contradiction then follows from the fact that $\mathcal{D} \to 0$ as $t \to \infty$.

In practice the diffusive balance (2) is integrated in time using Euler forward differencing, for example, to generate the concentration $c(\boldsymbol{x}, t_n)$ at time t_n, say. One then seeks a deformation $\boldsymbol{\chi}(\boldsymbol{x}, t_n)$ that minimizes the energy under the constraint $J_F(\boldsymbol{x}, t_n) = 1 + c(\boldsymbol{x}, t_n)$. This procedure motivates a definition, in Section 2.4, of a kinematic decomposition of the deformation gradient into elastic and swelling parts in which the latter is associated with an unstressed state at an assigned value of c. It also suggests a framework whereby an extension of Ball's theory [1977], in which polyconvexity figures prominently, may be explored to establish the existence of an energy minimizing deformation $\boldsymbol{\chi}(\boldsymbol{x}, t_n)$.

2.3. ***Polyconvex energies for isotropic materials.*** If the free-energy function is frame invariant and isotropic relative to κ then it is expressible as a function of the

list $\{i_1, i_2, i_3, c\}$, where

$$i_1 = \operatorname{tr} \boldsymbol{U}, \quad i_2 = \operatorname{tr} \boldsymbol{U}^*, \quad i_3 = J_F \tag{27}$$

are the principal invariants of the right-stretch tensor $\boldsymbol{U} \in \mathrm{Sym}^+$ in the polar decomposition

$$\boldsymbol{F} = \boldsymbol{R}\boldsymbol{U} \tag{28}$$

of the deformation gradient, with $\boldsymbol{R} \in \mathrm{Orth}^+$. These are related to the more frequently used invariants

$$I_1 = \operatorname{tr} \boldsymbol{C}, \quad I_2 = \operatorname{tr} \boldsymbol{C}^*, \quad I_3 = J_C \tag{29}$$

of the Cauchy–Green deformation tensor $\boldsymbol{C} = \boldsymbol{F}^t \boldsymbol{F}$ by the invertible transformation [Steigmann 2002]

$$I_1 = i_1^2 - 2i_2, \quad I_2 = i_2^2 - 2i_1 i_3, \quad I_3 = i_3^2. \tag{30}$$

However, grounds for using the i_k in the formulation of constitutive equations, rather than I_k, are given below.

Because of the swelling constraint, any constitutive dependence of the energy on i_3 is induced by its dependence on c, provided that c is strictly positive. The invariant i_3 is therefore redundant whenever the swelling constraint is operative, and the free-energy function is then expressible in the form

$$\Psi(\boldsymbol{F}, c) = \psi(i_1, i_2, c). \tag{31}$$

Technically this function is defined on the manifold $\mathcal{M}$ in $\mathbb{R}^{10}$ defined by the swelling constraint $J_F - c = 1$. However, the same function is well defined for states $(\boldsymbol{F}, c) \in \mathbb{R}^{10}$ that do not satisfy the constraint. Thus (31) may also serve as an extension of the free-energy function into $\mathbb{R}^{10}$, for purposes of differentiation in (19). The associated Piola stress is given by [Steigmann 2002]

$$\boldsymbol{P} = \boldsymbol{R}\boldsymbol{\sigma}, \tag{32}$$

where

$$\boldsymbol{\sigma} = \Big(\frac{\partial \psi}{\partial i_1} + i_1 \frac{\partial \psi}{\partial i_2}\Big)\boldsymbol{I} - \frac{\partial \psi}{\partial i_2}\boldsymbol{U} - q\boldsymbol{U}^* \tag{33}$$

is the (symmetric) Biot stress, and the chemical potential is

$$\mu = \frac{\partial \psi}{\partial c} + q. \tag{34}$$

We remark that any alternative extension must, of course, reduce to (31) on the constraint manifold. Accordingly, because any trajectory $(\boldsymbol{F}(t), c(t)) \in \mathcal{M}$ has a derivative with respect to t that lies in the tangent space to $\mathcal{M}$ at the considered instant, it follows that the part of the derivative $(\Psi_{\boldsymbol{F}}, \Psi_c)$ of the extended function,

in the direction $(\boldsymbol{F}^*, -1)$ orthogonal to $\mathcal{M}$, cannot be determined by inequality (18). Therefore, the use of an alternative extension affects only the as-yet-undetermined scalar field q. The extension (31) may therefore be used without loss of generality.

We may also use the swelling constraint to eliminate the concentration in favor of i_3 and thus express the energy as a function of $\{i_1, i_2, i_3\}$, in the manner of a conventional compressible solid [Treloar 1975; Pence and Tsai 2005; Pence and Tsai 2006; Deng and Pence 2010], albeit subject to the requirement $i_3 > 1$. In this interpretation we write (31) in the form

$$\Psi'(\boldsymbol{F}) = \varphi(i_1, i_2, i_3), \tag{35}$$

where $\Psi'(\boldsymbol{F}) = \Psi(\boldsymbol{F}, J_F - 1)$ and, of course,

$$\varphi(i_1, i_2, i_3) = \psi(i_1, i_2, i_3 - 1). \tag{36}$$

The minimum-energy argument of Section 2.2 applies to this surrogate compressible material.

In view of this it is natural to investigate constitutive functions that satisfy the well-known polyconvexity condition of Ball's existence theory for energy minimizers [1977]. Polyconvexity of Ψ' is the requirement that there exists a function

$$\Phi(\boldsymbol{F}, \boldsymbol{F}^*, J_F) = \Psi'(\boldsymbol{F}), \tag{37}$$

in general non-unique [Podio-Guidugli 1991; Podio-Guidugli and Vergara-Caffarelli 1991], which is jointly convex in its arguments; we make the obvious choice

$$\Phi(\boldsymbol{F}, \boldsymbol{F}^*, J_F) = \varphi\big\{\mathrm{tr}(\sqrt{\boldsymbol{F}^t\boldsymbol{F}}),\ \mathrm{tr}(\sqrt{(\boldsymbol{F}^*)^t\boldsymbol{F}^*}),\ J_F\big\}. \tag{38}$$

If Φ is differentiable, then [Ball 1977]

$$\Psi'(\bar{\boldsymbol{F}}) - \Psi'(\boldsymbol{F}) \geq \Phi_{\boldsymbol{F}}(\boldsymbol{F}, \boldsymbol{F}^*, J_F)\cdot(\bar{\boldsymbol{F}} - \boldsymbol{F}) + \Phi_{\boldsymbol{F}^*}(\boldsymbol{F}, \boldsymbol{F}^*, J_F)\cdot(\bar{\boldsymbol{F}}^* - \boldsymbol{F}^*) + \Phi_{J_F}(\boldsymbol{F}, \boldsymbol{F}^*, J_F)(J_{\bar{F}} - J_F) \tag{39}$$

for all deformation gradients $\boldsymbol{F}$ and $\bar{\boldsymbol{F}}$. Here we use

$$(i_1)_{\boldsymbol{F}} = (i_2)_{\boldsymbol{F}^*} = \boldsymbol{R} \tag{40}$$

[Steigmann 2003] to conclude that

$$\Phi_{\boldsymbol{F}} = \frac{\partial\varphi}{\partial i_1}\boldsymbol{R} \quad \text{and} \quad \Phi_{\boldsymbol{F}^*} = \frac{\partial\varphi}{\partial i_2}\boldsymbol{R}. \tag{41}$$

Polyconvexity ensures that the pointwise values of the free energy fulfill the (non-local) quasiconvexity condition, which is always satisfied by energy minimizers [Ball 1977]. Further, it guarantees the sequential weak lower semi-continuity of the potential energy functional which, together with coercivity of the free-energy function, ensures the existence of energy-minimizing deformations belonging to

an appropriate Sobolev space. Here, coercivity refers to the restriction [Ball 1977; Steigmann 2003]

$$\varphi(i_1, i_2, i_3) \geq k_1(i_1^p + i_2^q + i_3^r) + k_2, \tag{42}$$

where $p \geq 2$, $q \geq p/(p-1)$, $r > 1$ and k_1, k_2 are constants with $k_1 > 0$. Moreover, polyconvexity is free from the well-known objections [Ball 1977] raised against ordinary convexity of the free energy with respect to the deformation gradient, a condition which also implies quasiconvexity. However, Ball's existence theory is not immediately applicable here because of inequality (17); whereas that theory relies on a hypothesis about the behavior of the free-energy function in the limit $J_F \to 0$.

Thus our restriction to polyconvex energies is motivated by the fact that quasiconvexity, which is necessary for energy minimizers, is thereby assured. Further, the coercivity condition, while an integral part of Ball's theory, is evidently restrictive as a number of explicit solutions to equilibrium boundary-value problems have been obtained using non-coercive free energies [Carroll 1988]. Therefore we do not impose coercivity.

By [Steigmann 2003], the function $\Psi'(\boldsymbol{F})$ defined by (37) and (38) is polyconvex if and only if

$$\begin{aligned} &\varphi \text{ is a convex function of all three arguments jointly, and} \\ &\varphi \text{ is a nondecreasing function of } i_1 \text{ and } i_2; \end{aligned} \tag{43}$$

that is, if and only if

$$\varphi(\bar{\imath}_1, \bar{\imath}_2, \bar{\imath}_3) - \varphi(i_1, i_2, i_3) \geq (\bar{\imath}_1 - i_1)\frac{\partial \varphi}{\partial i_1} + (\bar{\imath}_2 - i_2)\frac{\partial \varphi}{\partial i_2} + (\bar{\imath}_3 - i_3)\frac{\partial \varphi}{\partial i_3} \tag{44}$$

together with

$$\frac{\partial \varphi}{\partial i_1} \geq 0 \quad \text{and} \quad \frac{\partial \varphi}{\partial i_2} \geq 0, \tag{45}$$

in which i_k and $\bar{\imath}_k$, respectively, are the invariants associated with $\boldsymbol{F}$ and $\bar{\boldsymbol{F}}$ and the derivatives are evaluated at i_k. Sufficiency follows from (41) and the fact that i_1 and i_2 are convex functions of $\boldsymbol{F}$ and $\boldsymbol{F}^*$, respectively [Steigmann 2003], i.e.,

$$i_1(\bar{\boldsymbol{F}}) - i_1(\boldsymbol{F}) \geq \boldsymbol{R}(\boldsymbol{F}) \cdot (\bar{\boldsymbol{F}} - \boldsymbol{F}) \quad \text{and} \quad i_2(\bar{\boldsymbol{F}}^*) - i_2(\boldsymbol{F}^*) \geq \boldsymbol{R}(\boldsymbol{F}^*) \cdot (\bar{\boldsymbol{F}}^* - \boldsymbol{F}^*). \tag{46}$$

A proof of the necessity of (44) and (45) for polyconvexity is given here in the Appendix. In this regard it is important to note that (43) is equivalent to polyconvexity in respect of the function defined by (38), but may not be if alternative choices are adopted. This is due to the fact that the function Φ defined by (37) is not unique.

The simplicity of the polyconvexity criteria (43) supports our preference for a constitutive formulation based on the stretch tensor.

Of particular relevance to the present work is the free-energy function

$$\Psi_{(c)}(\boldsymbol{F}; \boldsymbol{x}) = \Psi(\boldsymbol{F}, c(\boldsymbol{x})), \tag{47}$$

obtained by fixing the concentration field at $c(\boldsymbol{x})$, say, which may be specified as an arbitrary function taking non-negative values. We expect that Ball's theory for incompressible materials, adapted to accommodate the assignment of $J_F(\boldsymbol{x})$ in accordance with the swelling constraint, may yield the existence of minimizers in this case.

Remark. This free-energy function pertains to a gel in which diffusion has ceased and the volumetric deformation $J_F(\boldsymbol{x})$ is fixed. Equation (2), with $(19)_2$, (22) and $\dot{c} = 0$, then becomes a restriction on the function $q(\boldsymbol{x})$. The associated flux $\boldsymbol{m}$ is divergence-free and the net (integrated) flux through the boundary $\partial\kappa$ vanishes. The argument in Section 2.2 about minimum-energy states is applicable under the slightly stronger condition that $\boldsymbol{m} \cdot \boldsymbol{n}$ and μ vanish pointwise on complementary parts of the boundary; these yield data for the determination of $q(\boldsymbol{x})$, if desired.

The function $\Psi_{(c)}$ is polyconvex at $\boldsymbol{x}^* \in \kappa$ if and only if there exists a convex function

$$\Phi_{(c^*)}(\boldsymbol{F}, \boldsymbol{F}^*) = \Psi(\boldsymbol{F}, c^*), \tag{48}$$

where $c^* = c(\boldsymbol{x}^*)$; that is, if and only if

$$\begin{aligned}\Psi_{(c)}(\overline{\boldsymbol{F}}; \boldsymbol{x}^*) - \Psi_{(c)}(\boldsymbol{F}; \boldsymbol{x}^*) \\ \geq \Phi_{(c^*)\boldsymbol{F}}(\boldsymbol{F}, \boldsymbol{F}^*) \cdot (\overline{\boldsymbol{F}} - \boldsymbol{F}) + \Phi_{(c^*)\boldsymbol{F}^*}(\boldsymbol{F}, \boldsymbol{F}^*) \cdot (\overline{\boldsymbol{F}}^* - \boldsymbol{F}^*).\end{aligned} \tag{49}$$

In the present circumstances we take this function to be

$$\Phi_{(c)}(\boldsymbol{F}, \boldsymbol{F}^*) = \psi\left\{\mathrm{tr}(\sqrt{\boldsymbol{F}^t \boldsymbol{F}}),\ \mathrm{tr}(\sqrt{(\boldsymbol{F}^*)^t \boldsymbol{F}^*}),\ c\right\}. \tag{50}$$

The inequality does not involve the determinant of the deformation gradient; this is fixed at the value $1 + c^*$ by the swelling constraint. Necessary and sufficient conditions in this case are [Steigmann 2003]:

$$\begin{aligned}&\psi \text{ is a convex function of } i_1 \text{ and } i_2 \text{ jointly, and} \\ &\psi \text{ is a nondecreasing function of } i_1 \text{ and } i_2;\end{aligned} \tag{51}$$

that is,

$$\psi(\bar{\imath}_1, \bar{\imath}_2, c^*) - \psi(i_1, i_2, c^*) \geq (\bar{\imath}_1 - i_1)\frac{\partial \psi}{\partial i_1} + (\bar{\imath}_2 - i_2)\frac{\partial \psi}{\partial i_2}, \tag{52}$$

together with

$$\frac{\partial \psi}{\partial i_1} \geq 0 \quad \text{and} \quad \frac{\partial \psi}{\partial i_2} \geq 0, \tag{53}$$

in which i_k and $\bar{i}_k$, respectively, are the invariants associated with $\boldsymbol{F}$ and $\overline{\boldsymbol{F}}$ and the derivatives are again evaluated at i_k. Sufficiency is proved in [Steigmann 2003] whereas the proof of necessity given in the Appendix applies here as well.

2.4. ***Local free swelling.*** Consider a subvolume $\pi \subset \kappa$ with $\boldsymbol{p}$ and $\mu\boldsymbol{m}\cdot\boldsymbol{n}$ vanishing everywhere on $\partial\pi$. The foregoing argument about the decay of energy is thus applicable to π with the load potential equal to a constant. This is the *free-swelling problem*, and plays a central role in [Treloar 1975; Pence and Tsai 2005; Hong et al. 2008; Duda et al. 2010; Chester and Anand 2011], where it is used to specify constitutive information; i.e., restrictions on the response of the gel at points in κ. To make contact with these ideas, we make repeated use of the following simple result: If $F(\boldsymbol{x})$ is continuous and $MV(F)$ is the mean value of F in π, then there is a point $\boldsymbol{x}^* \in \pi$ such that $F(\boldsymbol{x}^*) = MV(F)$. Further, if $d(\pi) \to 0$, where $d(\pi) = \sup_{\boldsymbol{x}_1,\boldsymbol{x}_2\in\pi} |\boldsymbol{x}_1 - \boldsymbol{x}_2|$ is the diameter of π, then $|\boldsymbol{x} - \boldsymbol{x}^*| \to 0$ for all $\boldsymbol{x} \in \pi$ and the continuity of F implies that $F(\boldsymbol{x}) \to F(\boldsymbol{x}^*)$. For example, the well-known mean-stress theorem, following from (4) and (6), yields

$$MV(\boldsymbol{P}) = \int_{\partial\pi} \boldsymbol{p} \otimes \boldsymbol{x}\, da, \tag{54}$$

which vanishes in the free-swelling problem. Accordingly, $\boldsymbol{P}(\boldsymbol{x}) \to \boldsymbol{0}$ as $d(\pi) \to 0$, implying that the pointwise values of the stress field may be brought arbitrarily close to zero by making the diameter of π correspondingly small against any available length scale. In the same way, by integrating $\mu\boldsymbol{m}\cdot\boldsymbol{n}(=0)$ over $\partial\pi$, we find, using the divergence theorem and the diffusive balance, that

$$MV(\mu\dot{c}) = MV(\boldsymbol{m}\cdot D\mu) \le 0. \tag{55}$$

Therefore, if $d(\pi) \to 0$ it follows that

$$\mu\dot{c} \le 0, \tag{56}$$

pointwise in π.

In the present setting, the condition on stress reduces to $\boldsymbol{\sigma} = \boldsymbol{0}$; i.e.,

$$\Big(\frac{\partial\psi}{\partial i_1} + i_1\frac{\partial\psi}{\partial i_2}\Big)\boldsymbol{I} - \frac{\partial\psi}{\partial i_2}\boldsymbol{U} = q\boldsymbol{U}^*. \tag{57}$$

The trace yields

$$qi_2 = 3\frac{\partial\psi}{\partial i_1} + 2i_1\frac{\partial\psi}{\partial i_2}, \tag{58}$$

which furnishes q in terms of i_1, i_2 and c. If this state is at least asymptotically stable then it minimizes the energy; i.e.,

$$MV[\Psi(\overline{\boldsymbol{F}}, \bar{c}) - \Psi(\boldsymbol{F}, c)] \ge 0, \tag{59}$$

where $\bar{\boldsymbol{F}}(\boldsymbol{x})$ and $\bar{c}(\boldsymbol{x})$ are any deformation and concentration fields consistent with the balance laws and boundary conditions. Accordingly, if $d(\pi) \to 0$ then

$$\Psi(\bar{\boldsymbol{F}}, \bar{c}) \geq \Psi(\boldsymbol{F}, c), \tag{60}$$

again pointwise.

Two definitions of the free-swelling problem are evident:

(a) In the first, which is tacitly adopted in [Pence and Tsai 2005; Chester and Anand 2011], the flux $\boldsymbol{m} \cdot \boldsymbol{n}$ is unrestricted on $\partial\pi$; this requires μ to vanish there, and the local inequality (56) is satisfied if $d(\pi) \to 0$, with $\dot{c}$ unrestricted. This in turn requires that $\mu = 0$ pointwise, ensuring that the condition on the boundary is satisfied. Equation (34) gives $q = -\partial\psi/\partial c$, while (58) reduces to

$$3\frac{\partial\varphi}{\partial i_1} + 2\frac{\partial\varphi}{\partial i_2} + i_2\frac{\partial\varphi}{\partial i_3} = 0, \tag{61}$$

where we have used the connection (36) pertaining to the equivalent compressible material, in which the stress is given by (33) with ψ replaced by φ and c by $i_3 - 1$. This is effectively the free-swelling condition given in Equation (7) of [Deng and Pence 2010], yielding an equation for c. We note that our polyconvexity criteria $(43)_2$ then require that $\partial\varphi/\partial i_3 \leq 0$, which is compatible with $(43)_1$.

(b) In the second definition of free swelling, proposed here, the diffusive flux $\boldsymbol{m} \cdot \boldsymbol{n}$ vanishes pointwise on $\partial\pi$ and the argument leading to (55), with $\mu\boldsymbol{m}$ replaced by $\boldsymbol{m}$, yields

$$\frac{d}{dt}MV(c) = 0, \quad \text{so that} \quad MV(c) = c^*, \tag{62}$$

a constant. Accordingly, if $d(\pi) \to 0$ then

$$c = c^*, \tag{63}$$

pointwise. This implies $\dot{c} = 0$, ensuring that (56) is satisfied. Inequality (60) reduces to

$$\Psi_{(c^*)}(\bar{\boldsymbol{F}}) \geq \Psi_{(c^*)}(\boldsymbol{F}); \quad J_{\bar{F}} = J_F = 1 + c^*, \tag{64}$$

where $\Psi_{(c)}$ is defined by (47).

In the literature on isotropic elastomers [Treloar 1975; Pence and Tsai 2005; Hong et al. 2008; Duda et al. 2010; Chester and Anand 2011] we find the universal assumption that the free-swelling deformation is a pure equi-triaxial stretch; i.e., that

$$\boldsymbol{F} = \lambda\boldsymbol{I} \tag{65}$$

for some $\lambda > 0$, yielding $\boldsymbol{U} = \boldsymbol{F}$. Then, Equation (58) reduces to

$$\frac{\partial\psi}{\partial i_1} + 2\lambda\frac{\partial\psi}{\partial i_2} = \lambda^2 q, \tag{66}$$

in which the derivatives are evaluated at

$$i_1 = 3\lambda \quad \text{and} \quad i_2 = 3\lambda^2. \tag{67}$$

The swelling constraint yields

$$\lambda^3 = 1 + c^* \tag{68}$$

and (66) and (34) yield unique values of q and μ. We have thus established the existence of a state that satisfies the local free-swelling problem in which concentration is assigned.

If the solution described by (65)–(68) is to be stable, it must satisfy (64). To investigate this we follow [Ogden 1984, p. 110] and decompose the deformation $\bar{\boldsymbol{F}}$ of the comparison state in the form

$$\bar{\boldsymbol{F}} = \bar{\boldsymbol{R}}\bar{\boldsymbol{U}}, \tag{69}$$

in which $\bar{\boldsymbol{R}} \in \mathrm{Orth}^+$ and

$$\begin{aligned}\bar{\boldsymbol{U}} &= \sum \bar{\lambda}_i \boldsymbol{u}_i \otimes \boldsymbol{u}_i \qquad (70)\\ &= (\bar{\lambda}\boldsymbol{I})(s\boldsymbol{u}_1 \otimes \boldsymbol{u}_1 + s^{-1}\boldsymbol{u}_2 \otimes \boldsymbol{u}_2 + \boldsymbol{u}_3 \otimes \boldsymbol{u}_3)[t^{-1/2}(\boldsymbol{u}_1 \otimes \boldsymbol{u}_1 + \boldsymbol{u}_2 \otimes \boldsymbol{u}_2) + t\boldsymbol{u}_3 \otimes \boldsymbol{u}_3],\end{aligned}$$

where $\bar{\lambda}_i (> 0)$ are the principal stretches, $\{\boldsymbol{u}_i\}$ are the orthonormal principal axes of $\bar{\boldsymbol{U}}$ and the factors correspond to a pure equi-triaxial stretch of amount $\bar{\lambda}(> 0)$, a pure shear of amount $s(> 0)$ and an isochoric uniaxial extension of amount $t(> 0)$ with accompanying lateral contraction. These are coaxial and so may be composed in any order. The principal stretches are

$$\bar{\lambda}_1 = \bar{\lambda} s t^{-1/2}, \quad \bar{\lambda}_2 = \bar{\lambda} s^{-1} t^{-1/2}, \quad \bar{\lambda}_3 = \bar{\lambda} t, \tag{71}$$

which may be inverted to yield

$$\bar{\lambda} = (\bar{\lambda}_1 \bar{\lambda}_2 \bar{\lambda}_3)^{1/3}, \quad t = \bar{\lambda}_3^{2/3}(\bar{\lambda}_1 \bar{\lambda}_2)^{-1/3}, \quad s = (\bar{\lambda}_1/\bar{\lambda}_2)^{1/2}. \tag{72}$$

Accordingly, (71) affords a general representation of any state of stretch and the decomposition (70) entails no loss of generality [Ogden 1984]. The corresponding invariants are

$$\bar{\imath}_1 = \bar{\lambda}[(s + s^{-1})t^{-1/2} + t], \quad \bar{\imath}_2 = \bar{\lambda}^2[(s + s^{-1})t^{1/2} + t^{-1}], \quad \bar{\imath}_3 = \bar{\lambda}^3. \tag{73}$$

It is instructive to insert these expressions into the polyconvexity criteria. Because c is fixed at the value c^* in inequality (64), the relevant polyconvexity condition is (49), in which $\bar{\lambda} = \lambda$. Accordingly, (52) yields

$$\Psi_{(c^*)}(\bar{\boldsymbol{F}}) - \Psi_{(c^*)}(\boldsymbol{F}) \geq \lambda f(s,t)\frac{\partial \psi}{\partial i_1} + \lambda^2 f(s,u)\frac{\partial \psi}{\partial i_2}, \tag{74}$$

where $u = t^{-1}$, and, for x and y positive,

$$f(x, y) = (x + x^{-1})y^{-1/2} + y - 3. \tag{75}$$

The derivatives in (74) are evaluated at the invariants given in (67), with (68).

The function f is stationary at $(x, y) = (1, 1)$. Its Hessian matrix there is easily shown to be positive definite, so that f has a local minimum, equal to zero, at $(1, 1)$. At this point both the pure shear and uniaxial extension of (70) reduce to the identity, and $\overline{\boldsymbol{U}} = \boldsymbol{U}$ $(= \lambda \boldsymbol{I})$. It then follows from (74) that $\Psi_{(c^*)}$ has a local minimum at the equi-triaxial stretch defined by (67) and (68). In fact the minimum is global, as $f \geq 0$ for all values of its arguments. This claim is easily proved by observing that it is equivalent to the inequality $x + x^{-1} \geq (3 - y)y^{1/2}$, the truth of which follows from the fact that the left-hand side has a strict global minimum, equal to 2, at $x = 1$; whereas the right-hand side has a strict global maximum, also equal to 2, at $y = 1$. Thus, $f(1, 1) = 0$ and $f(x, y) > 0$ for all $(x, y) \neq (1, 1)$ with x and y positive. In particular, inequality (74) and the strict versions of (53) then imply that (64) holds in the strict sense for all $\overline{\boldsymbol{U}} \neq \boldsymbol{U}$ such that $J_{\overline{F}} = J_F$.

We have shown that the polyconvexity criteria (51) imply that a state defined by (57) and (65) furnishes an optimal solution to the free swelling problem associated with a given concentration. This result is perhaps surprising. For, although polyconvexity implies quasiconvexity, the latter is not known to imply (64). In fact the optimality of the solution (65) follows from the weaker restriction of rank-one convexity. This is demonstrated in the Appendix.

Our result justifies the local decomposition [Chester and Anand 2011]

$$\boldsymbol{F} = \boldsymbol{H}\boldsymbol{G}, \tag{76}$$

in which

$$\boldsymbol{G} = (1 + c)^{1/3}\boldsymbol{I} \quad \text{with } c \geq 0, \tag{77}$$

is the free-swelling deformation associated with c, yielding a swollen stress-free local configuration κ_c followed by an elastic deformation $\boldsymbol{H}$ that gives rise to stress. Indeed, if such $\boldsymbol{G}$ were not energetically optimal, then theories based on (76) might well yield predictions that are sub-optimal and perhaps even unstable. We observe that (76) and (77) are consistent with the swelling constraint (16) if and only if

$$J_H = 1. \tag{78}$$

Using the foregoing procedure, the free-swelling state can be achieved at all points of the gel provided that κ is divided into an arbitrarily large number of subvolumes, each of which satisfies the free-swelling problem. In general, the local configurations resulting from this operation cannot be made congruent in three-dimensional space in the absence of strain; that is, they do not necessarily fit

together to form a connected whole in Euclidean space. Instead, the union of such states is to be regarded as a smooth three-dimensional manifold, whose tangent space at a given value of c is identified with κ_c. The fact that this manifold is generally non-Euclidean implies that $\boldsymbol{G}$ is not the gradient of any position field. Accordingly, it does not satisfy the compatibility condition which follows from the existence of such a field. Of course, $\boldsymbol{F}$, being the gradient of the map $\boldsymbol{\chi}$, is necessarily compatible. The incompatibility of $\boldsymbol{G}$ then implies that $\boldsymbol{H}$ is likewise incompatible. That such a formalism is necessary follows from the fact that $\boldsymbol{G}$ is compatible if and only if the field $c(\boldsymbol{x}, t)$ is uniform. In this case the equilibrium equation, diffusive balance and boundary/initial conditions yield an overdetermined problem for the fields $\boldsymbol{\chi}(\boldsymbol{x}, t)$ and $q(\boldsymbol{x}, t)$, having no solution except in special circumstances.

2.5. ***Using κ_c as reference.*** The decomposition (76) suggests the use of κ_c as reference when formulating constitutive equations [Chester and Anand 2011]. For example, given a deformation field $\boldsymbol{\chi}(\boldsymbol{x}, t)$ the free-energy function per unit volume of κ_c is $W(\boldsymbol{F}\boldsymbol{K}, c) = J_K \Psi(\boldsymbol{F}, c)$, where $\boldsymbol{K} = \boldsymbol{G}^{-1}$. Accordingly,

$$\Psi(\boldsymbol{F}, c) = (1+c) W(\boldsymbol{H}, c), \quad \text{where} \quad \boldsymbol{H} = (1+c)^{-1/3} \boldsymbol{F}. \tag{79}$$

This decomposition shows that the present model is subsumed, at fixed c, under Noll's theory of materially uniform bodies [Noll 1967], which has dramatically advanced the development of theories of plasticity and continuously distributed defects. Here, because the local change of reference from κ to κ_c entails a pure dilation (cf. (77)), it does not affect the symmetry group of the gel. This is a simple consequence of Noll's rule [Truesdell 1977] connecting the symmetry groups associated with local references. Accordingly, the material is isotropic with respect to both κ and κ_c. This fact, the constraint (16) and the invariance of W under superposed rotations — a property inherited from Ψ — imply that the strain energy W depends on $\boldsymbol{H}$ via the list $\{h_1, h_2\}$, where

$$h_1 = \operatorname{tr} \boldsymbol{U}_H \quad \text{and} \quad h_2 = \operatorname{tr} \boldsymbol{U}_H^*. \tag{80}$$

Here $\boldsymbol{U}_H$ is the symmetric, positive right-stretch tensor in the polar factorization of $\boldsymbol{H}$. From $(79)_2$, the rotation in this factorization is simply $\boldsymbol{R}$, the rotation associated with $\boldsymbol{F}$. It follows that $\boldsymbol{U}_H = (1+c)^{-1/3} \boldsymbol{U}$; therefore,

$$h_1 = (1+c)^{-1/3} i_1 \quad \text{and} \quad h_2 = (1+c)^{-2/3} i_2. \tag{81}$$

These in turn yield $W(\boldsymbol{H}, c) = w(h_1, h_2, c)$ for some function w, which may be used to write (79) in the form

$$\psi(i_1, i_2, c) = (1+c) w(h_1, h_2, c). \tag{82}$$

Using this with (81), it is straightforward to show that the polyconvexity criteria (51) are satisfied *if and only if*

$$\begin{aligned} &w \text{ is a convex function of } h_1 \text{ and } h_2 \text{ jointly, and} \\ &w \text{ is a nondecreasing function of } h_1 \text{ and } h_2, \end{aligned} \tag{83}$$

and this in turn implies that the function $W_{(c)}(\boldsymbol{H}) = W(\boldsymbol{H}, c)$, associated with fixed c, is polyconvex [Steigmann 2003]. Thus, there is a function $G_{(c)}(\boldsymbol{H}, \boldsymbol{H}^*) = W_{(c)}(\boldsymbol{H})$ such that

$$\begin{aligned} &W_{(c)}(\overline{\boldsymbol{H}}) - W_{(c)}(\boldsymbol{H}) \\ &\qquad \geq G_{(c)\boldsymbol{H}}(\boldsymbol{H}, \boldsymbol{H}^*) \cdot (\overline{\boldsymbol{H}} - \boldsymbol{H}) + G_{(c)\boldsymbol{H}^*}(\boldsymbol{H}, \boldsymbol{H}^*) \cdot (\overline{\boldsymbol{H}}^* - \boldsymbol{H}^*). \end{aligned} \tag{84}$$

In the present circumstances we take $G_{(c)}(\boldsymbol{H}, \boldsymbol{H}^*) = w(h_1, h_2, c)$ (cf. (50)), yielding $G_{(c)\boldsymbol{H}} = (\partial w/\partial h_1)\boldsymbol{R}$ and $G_{(c)\boldsymbol{H}^*} = (\partial w/\partial h_2)\boldsymbol{R}$. Indeed, using (81), (82) and (41) it is straightforward to show that (84) is *equivalent* to inequality (49). This furnishes an explicit example of the general fact that polyconvexity is preserved under *any* fixed change of reference [Neff 2003, Lemma 6.5, p. 260]. Using the results of the previous subsection, we conclude that any *distortion*; i.e., any $\boldsymbol{U}_H \neq \boldsymbol{I}$ with $J_H = 1$, entails an energetic cost if w satisfies (83).

2.6. ***Constitutive specification of the diffusive flux.*** A well-known representation theorem for isotropic functions [Noll 1970] furnishes the canonical form of the mobility tensor for materials exhibiting holohedral (as distinct from hemihedral) isotropy. Here, we combine this theorem with the Cayley–Hamilton formula to conclude that

$$\boldsymbol{M} = \alpha_0 \boldsymbol{I} + \alpha_1 \boldsymbol{U} + \alpha_2 \boldsymbol{U}^2, \tag{85}$$

where $\alpha_{0,1,2}$ are functions of $i_{1,2}$ and c, arranged to ensure that the mobility is negative definite. The swelling constraint implies that i_3 is redundant. This representation furnishes the referential diffusive flux $\boldsymbol{m}$ defined by (22).

In [Chester and Anand 2011] the diffusive flux is expressed in terms of variables pertaining to the local stress-free swollen configuration κ_c. The operative diffusive flux is the push-forward

$$\boldsymbol{m}_c = J_G^{-1} \boldsymbol{G}\boldsymbol{m} \tag{86}$$

of the referential flux $\boldsymbol{m}$. Using (22) and (77), this is easily shown to yield

$$\boldsymbol{m}_c = (\gamma_0 \boldsymbol{I} + \gamma_1 \boldsymbol{U}_H + \gamma_2 \boldsymbol{U}_H^2) \boldsymbol{H}^t (\operatorname{grad} \mu), \tag{87}$$

where grad is the spatial gradient based on position $\boldsymbol{y}$ and $\gamma_{0,1,2}$ are functions of the invariants $h_{1,2}$ and c. This is an isotropic function jointly of $\boldsymbol{U}_H$ and the push-forward to κ_c of the referential chemical-potential gradient, namely, $\boldsymbol{H}^t(\operatorname{grad} \mu) = \boldsymbol{G}^{-t} D\mu$.

Chester and Anand [2011] state that isotropy of the constitutive equation for a *single* diffusive flux vector should be imposed relative to *both* κ and κ_c. This restriction implies that the mobility tensor is purely spherical; i.e., that it is proportional to the identity, as they point out. However, we have shown that isotropy is preserved under the transformation $\kappa \to \kappa_c$, provided that the appropriate flux vector is used. Thus we conclude that the imposition in [Chester and Anand 2011] of the dual requirement on the constitutive equation for a single diffusive flux vector is not appropriate. Instead, we observe that the preservation of isotropy is due to the fact that the transformation $\kappa \to \kappa_c$ is a pure dilation. According to Noll's theorem [Truesdell 1977], these do not alter the symmetry group. When using this theorem it is essential to account for the induced change in the referential variables (cf. (86)).

3. Two-dimensional theory

The two-dimensional version of the foregoing theory, applicable to plane-strain deformations, is substantially simpler. Here we use the same notation as before with the stipulation that all tensors are regarded as linear maps from E^2 to itself. Thus the model discussed is inherently two dimensional. We present a synopsis of the main results in this case, emphasizing those features that differ from the corresponding three-dimensional theory.

The arguments leading to (19) and (32) remain unaltered, and thus yield

$$\boldsymbol{P} = \boldsymbol{R}\boldsymbol{\sigma}, \quad \text{with } \boldsymbol{\sigma} = \frac{\partial \varphi}{\partial i}\boldsymbol{I} - q\boldsymbol{U}^*, \tag{88}$$

where

$$i = \operatorname{tr} \boldsymbol{U} \tag{89}$$

is one of the two independent isotropic invariants. The other, $\det \boldsymbol{U}$, is fixed by the swelling constraint, which carries over without modification, except of course that all reference to *volume* is replaced here by *area*. The relation (34) connecting the chemical potential, free energy and Lagrange multiplier is also unaltered.

In two dimensions we have [Steigmann 2002]

$$i\boldsymbol{R} = \boldsymbol{F} + \boldsymbol{F}^*, \tag{90}$$

and therefore

$$\boldsymbol{P} = i^{-1}\Big(\frac{\partial \psi}{\partial i}\Big)(\boldsymbol{F} + \boldsymbol{F}^*) - q\boldsymbol{F}^*, \tag{91}$$

which is required to satisfy $\operatorname{div} \boldsymbol{P} = \boldsymbol{0}$ in κ together with any traction data on $\partial\kappa$, where div is the *two-dimensional* divergence on κ.

The diffusive balance (2) takes the form

$$\dot{c} + \operatorname{div}[\boldsymbol{M}(\boldsymbol{U}, c)\nabla\mu] = 0, \tag{92}$$

in which ∇ is the two-dimensional gradient and, by the two-dimensional representation theorem for isotropic functions,

$$\boldsymbol{M}(\boldsymbol{U}, c) = \beta_0(i, c)\boldsymbol{I} + \beta_1(i, c)\boldsymbol{U}, \tag{93}$$

in which $\boldsymbol{I}$ is the identity for 2-space and the functions $\beta_{0,1}$ are restricted by the requirement that $\boldsymbol{M}$ be negative definite. To obtain $\boldsymbol{U}$ from the deformation gradient we use the two-dimensional Cayley–Hamilton formula

$$\boldsymbol{U} = i^{-1}[(1 + c)\boldsymbol{I} + \boldsymbol{C}], \quad \text{where } \boldsymbol{C} = \boldsymbol{F}^t\boldsymbol{F}, \tag{94}$$

in which the swelling constraint has been imposed.

The operative polyconvexity condition, replacing (51), is

$$\psi(i, c) \text{ is a convex, nondecreasing function of } i. \tag{95}$$

This is necessary and sufficient for $\Psi_{(c)}(\boldsymbol{F}) = \psi(i(\boldsymbol{F}), c)$ to be polyconvex [Steigmann 2003].

The free-swelling problem, in which c is assigned, is again solved by deformations of the form $\boldsymbol{F} = \lambda\boldsymbol{I}$, but now with

$$\lambda^2 = 1 + c. \tag{96}$$

Equation (66) is replaced by

$$q = \lambda^{-1}\frac{\partial\psi}{\partial i}, \tag{97}$$

in which the derivative is evaluated at $i = 2\lambda$, and (34) then furnishes a unique value of μ.

Polyconvexity again ensures the optimality of this solution to the free-swelling problem. To see this we use (95) to obtain

$$\Psi_{(c)}(\bar{\boldsymbol{F}}) - \Psi_{(c)}(\boldsymbol{F}) \geq (\bar{\imath} - i)\frac{\partial\psi}{\partial i}, \tag{98}$$

in which the derivative is evaluated at $i = 2\lambda$. The gradient $\bar{\boldsymbol{F}}$ may, without loss of generality, be decomposed in the form (69), where

$$\bar{\boldsymbol{U}} = (\bar{\lambda}\boldsymbol{I})(s\boldsymbol{u}_1 \otimes \boldsymbol{u}_1 + s^{-1}\boldsymbol{u}_2 \otimes \boldsymbol{u}_2) \tag{99}$$

is the composition of an areal dilation of amount $\bar{\lambda}$ and a pure shear of amount $s(> 0)$. The associated principal stretches are $\bar{\lambda}_1 = \bar{\lambda}s$ and $\bar{\lambda}_2 = \bar{\lambda}s^{-1}$, yielding

$$\bar{\lambda} = \bar{\lambda}_1\bar{\lambda}_2 \quad \text{and} \quad s = (\bar{\lambda}_1/\bar{\lambda}_2)^{1/2}.$$

Here we impose $\bar{\lambda} = \lambda$ in accordance with the swelling constraint, obtaining $\bar{\imath} = \lambda(s + s^{-1})$. Inequality (98) reduces to

$$\Psi_{(c)}(\bar{\boldsymbol{F}}) - \Psi_{(c)}(\boldsymbol{F}) \geq \lambda(s + s^{-1} - 2)\frac{\partial \psi}{\partial i}, \tag{100}$$

which is non-negative by virtue of (95) and the fact that $s + s^{-1}$ has an isolated minimum, equal to 2, at $s = 1$, corresponding to $\bar{\boldsymbol{U}} = \boldsymbol{U}$.

This result justifies the decomposition (cf. (76))

$$\boldsymbol{F} = \boldsymbol{H}\boldsymbol{G}, \tag{101}$$

in which

$$\boldsymbol{G} = (1 + c)^{1/2}\boldsymbol{I} \quad \text{with } c \geq 0, \tag{102}$$

and

$$J_H = 1, \tag{103}$$

by virtue of the swelling constraint.

The free energy, per unit reference area, is then expressible in the form

$$\psi(i, c) = (1 + c)w(h, c), \quad \text{where } h = i/\sqrt{1 + c}, \tag{104}$$

is the trace of $\boldsymbol{U}_H$, the stretch factor in the polar decomposition of $\boldsymbol{H}$, and w is the free energy per unit area of the swollen elastomer. It is easy to show that the polyconvexity criterion (95) holds if and only if the same restriction applies to w; that is, $w(h, c)$ is a convex, nondecreasing function of h.

Appendix

In [Steigmann 2003] the conditions (43) are shown to be sufficient for polyconvexity. The proof of necessity is given here.

The necessity of (44) follows immediately from (39), (40) and (41) by selecting a deformation $\bar{\boldsymbol{F}}$ for which $\boldsymbol{R}(\bar{\boldsymbol{F}}) = \boldsymbol{R}(\boldsymbol{F})$, and hence $\boldsymbol{R}(\bar{\boldsymbol{F}}^*) = \boldsymbol{R}(\boldsymbol{F}^*)$; these yield $\boldsymbol{R} \cdot (\bar{\boldsymbol{F}} - \boldsymbol{F}) = \bar{\imath}_1 - i_1$ and $\boldsymbol{R} \cdot (\bar{\boldsymbol{F}}^* - \boldsymbol{F}^*) = \bar{\imath}_2 - i_2$, respectively, thereby reducing (39) to (44).

To demonstrate the necessity of (45) we use the fact that polyconvexity implies rank-one convexity [Ball 1977], which is equivalent to the inequality

$$\Psi'(\bar{\boldsymbol{F}}) - \Psi'(\boldsymbol{F}) \geq \boldsymbol{P}(\boldsymbol{F}) \cdot (\bar{\boldsymbol{F}} - \boldsymbol{F}), \quad \text{with } \bar{\boldsymbol{F}} - \boldsymbol{F} = \boldsymbol{a} \otimes \boldsymbol{b}, \tag{A1}$$

for arbitrary $\boldsymbol{a}$ and $\boldsymbol{b}$. For isotropic materials, (32), (33), (34) and (36) may be used to express this in the form

$$\Psi'(\bar{\boldsymbol{F}}) - \Psi'(\boldsymbol{F}) \geq \Big[\Big(\frac{\partial \varphi}{\partial i_1} + i_1\frac{\partial \varphi}{\partial i_2}\Big)\boldsymbol{R} - \frac{\partial \varphi}{\partial i_2}\boldsymbol{F} + \frac{\partial \varphi}{\partial i_3}\boldsymbol{F}^*\Big] \cdot \boldsymbol{a} \otimes \boldsymbol{b}. \tag{A2}$$

Consider $\boldsymbol{a}$ and $\boldsymbol{b}$ such that $\boldsymbol{F}^{-1}\boldsymbol{a}\cdot\boldsymbol{b}=0$. For this choice (A2) reduces to

$$\Psi'(\bar{\boldsymbol{F}})-\Psi'(\boldsymbol{F})\geq\Big(\frac{\partial\varphi}{\partial i_1}+i_1\frac{\partial\varphi}{\partial i_2}\Big)\boldsymbol{a}\cdot\boldsymbol{R}\boldsymbol{b}-\frac{\partial\varphi}{\partial i_2}\boldsymbol{a}\cdot\boldsymbol{F}\boldsymbol{b},\quad\text{with } J_{\bar{F}}=J_F. \tag{A3}$$

If $\boldsymbol{b}$ is an eigenvector of $\boldsymbol{U}$ with eigenvalue λ, then we have the further reduction

$$\Psi'(\bar{\boldsymbol{F}})-\Psi'(\boldsymbol{F})\geq\Big[\frac{\partial\varphi}{\partial i_1}+(i_1-\lambda)\frac{\partial\varphi}{\partial i_2}\Big]\boldsymbol{a}\cdot\boldsymbol{R}\boldsymbol{b}. \tag{A4}$$

Because $\boldsymbol{U}$ and $\boldsymbol{U}^*$ have the same eigenvectors, the restrictions on $\boldsymbol{a}$ and $\boldsymbol{b}$ imply that $\boldsymbol{a}\cdot\boldsymbol{R}\boldsymbol{b}=0$, yielding

$$\Psi'(\bar{\boldsymbol{F}})-\Psi'(\boldsymbol{F})\geq 0,\quad\text{with } J_{\bar{F}}=J_F. \tag{A5}$$

To interpret this result in the present context, we use $\bar{\boldsymbol{F}}-\boldsymbol{F}=\boldsymbol{a}\otimes\boldsymbol{b}$, with $\boldsymbol{F}^{-1}\boldsymbol{a}\cdot\boldsymbol{b}=0$, in the convexity conditions (46), obtaining $\bar{\boldsymbol{F}}^*-\boldsymbol{F}^*=-\boldsymbol{F}^*\boldsymbol{b}\otimes\boldsymbol{F}^{-1}\boldsymbol{a}$ and

$$i_1(\bar{\boldsymbol{F}})-i_1(\boldsymbol{F})\geq\boldsymbol{a}\cdot\boldsymbol{R}(\boldsymbol{F})\boldsymbol{b}\quad\text{and}\quad i_2(\bar{\boldsymbol{F}}^*)-i_2(\boldsymbol{F}^*)\geq-\boldsymbol{U}^*\boldsymbol{b}\cdot\boldsymbol{F}^{-1}\boldsymbol{a}, \tag{A6}$$

where the rotation invariance of the inner product has been used in the second inequality. The right-hand sides of these inequalities vanish for the choice of $\boldsymbol{b}$ leading to (A5). Thus, (A5) is equivalent to the restriction

$$\varphi(\bar{\imath}_1,\bar{\imath}_2,i_3)\geq\varphi(i_1,i_2,i_3)\quad\text{for all } \bar{\imath}_1\geq i_1 \text{ and } \bar{\imath}_2\geq i_2. \tag{A7}$$

We observe that the restrictions $\bar{\imath}_1\geq i_1$ and $\bar{\imath}_2\geq i_2$ are satisfied in the formulas (73) pertaining to the free-swelling problem. Consequently the optimality of the solution (65) to that problem is a consequence of rank-one convexity, which of course is weaker than polyconvexity.

In a general deformation, the stretch invariants satisfy the restrictions $i_1^2\geq 3i_2$ and $i_2^2\geq 3i_1i_3$ [Podio-Guidugli and Vergara-Caffarelli 1991], with equality if and only if $\boldsymbol{U}$ is purely spherical. These impose the limits

$$\sqrt{3i_2}\leq i_1\leq\frac{i_2^2}{3i_3}\quad\text{and}\quad\sqrt{3i_1i_3}\leq i_2\leq\frac{i_1^2}{3} \tag{A8}$$

on the invariants; the first applying when i_2 and i_3 are specified and the second when i_1 and i_3 are specified. Fixing $\bar{\imath}_2=i_2$ and choosing $\bar{\imath}_1>i_1$ in the first interval, we write (A7) in the form

$$\epsilon\Big(\frac{\partial\varphi}{\partial i_1}+\epsilon^{-1}o(\epsilon)\Big)\geq 0,\quad\text{with } \epsilon=\bar{\imath}_1-i_1, \tag{A9}$$

which yields $(45)_1$ on dividing by $\epsilon(>0)$ and passing to the limit. In the same way (A7) is seen to imply $(45)_2$. Thus the necessity of (45) for polyconvexity has been

demonstrated. A discussion of necessary and sufficient conditions for polyconvexity in isotropic elasticity in terms of principal stretches is given in [Mielke 2005]. The idea for the present proof of necessity may be found in [Steigmann and Pipkin 1988], where it is applied to a special class of materials.

We note that if the free-energy function is strictly polyconvex, then the strict inequalities (44) and (45) follow. Moreover, if inequality (44) is strict and *at least one* of inequalities (45) is strict, then the free-energy function is *strictly* polyconvex. These statements are valid despite the fact that inequalities (46) are nonstrict [Steigmann 2003]. Finally, all of the foregoing remarks remain valid in the presence of a constraint on J_F, exemplified here by the swelling constraint.

References

[Adkins 1964] J. E. Adkins, "Non-linear diffusion, III: Diffusion through isotropic highly elastic solids", *Philos. Trans. Roy. Soc. London Ser. A* **256**:1071 (1964), 301–316.

[Baek and Srinivasa 2004] S. Baek and A. R. Srinivasa, "Diffusion of a fluid through an elastic solid undergoing large deformations", *Int. J. Non-Linear Mech.* **39**:2 (2004), 201–218.

[Ball 1977] J. M. Ball, "Convexity conditions and existence theorems in nonlinear elasticity", *Arch. Rational Mech. Anal.* **63**:4 (1977), 337–403.

[Carroll 1988] M. M. Carroll, "Finite strain solutions in compressible isotropic elasticity", *J. Elasticity* **20**:1 (1988), 65–92.

[Chester and Anand 2011] S. A. Chester and L. Anand, "A thermo-mechanically coupled theory for fluid permeation in elastomeric materials: application to thermally responsive gels", *J. Mech. Phys. Solids* **59**:10 (2011), 1978–2006.

[Deng and Pence 2010] H. Deng and T. J. Pence, "Shear induced loss of saturation and in a fluid infused swollen hyperelastic cylinder", *Int J. Engng. Sci.* **48**:6 (2010), 624–646.

[Duda et al. 2010] F. P. Duda, A. C. Souza, and E. Fried, "A theory for species migration in a finitely strained solid with application to polymer network swelling", *J. Mech. Phys. Solids* **58**:4 (2010), 515–529.

[Duda et al. 2011] F. P. Duda, A. C. Souza, and E. Fried, "Solvent uptake and cavitation", *J. Mech. Phys. Solids* **59**:11 (2011), 2341–2354.

[Gurtin 2000] M. E. Gurtin, *Configurational forces as basic concepts of continuum physics*, Applied Mathematical Sciences **137**, Springer, New York, 2000.

[Gurtin et al. 2010] M. E. Gurtin, E. Fried, and L. Anand, *The mechanics and thermodynamics of continua*, Cambridge University Press, 2010.

[Hong et al. 2008] W. Hong, X. Zhao, J. Zhou, and Z. Suo, "A theory of coupled diffusion and large deformation in polymeric gels", *J. Mech. Phys. Solids* **56**:5 (2008), 1779–1793.

[Liu 2002] I.-S. Liu, *Continuum mechanics*, Springer, Berlin, 2002.

[Mielke 2005] A. Mielke, "Necessary and sufficient conditions for polyconvexity of isotropic functions", *J. Convex Anal.* **12**:2 (2005), 291–314.

[Neff 2003] P. Neff, "Some results concerning the mathematical treatment of finite multiplicative elasto-plasticity", pp. 251–274 in *Deformation and failure in metallic and granular structures*,

edited by K. Hutter and H. Baaser, Lecture Notes in Applied and Computational Mech. **10**, Springer, Berlin, 2003.

[Noll 1967] W. Noll, "Materially uniform simple bodies with inhomogeneities", *Arch. Rational Mech. Anal.* **27**:1 (1967), 1–32.

[Noll 1970] W. Noll, "Representations of certain isotropic tensor functions", *Arch. Math.* (*Basel*) **21**:1 (1970), 87–90.

[Ogden 1984] R. W. Ogden, *Non-linear elastic deformations*, Ellis Horwood, Chichester, 1984. Reprinted Dover, New York, 1997.

[Pence and Tsai 2005] T. J. Pence and H. Tsai, "On the cavitation of a swollen compressible sphere in finite elasticity", *Int. J. Non-Linear Mech.* **40**:2-3 (2005), 307–321.

[Pence and Tsai 2006] T. J. Pence and H. Tsai, "Swelling-induced cavitation of elastic spheres", *Math. Mech. Solids* **11**:5 (2006), 527–551.

[Podio-Guidugli 1991] P. Podio-Guidugli, "Polyconvex energies and symmetry requirements", *J. Elasticity* **26**:3 (1991), 223–237.

[Podio-Guidugli and Vergara-Caffarelli 1991] P. Podio-Guidugli and G. Vergara-Caffarelli, "Extreme elastic deformations", *Arch. Rational Mech. Anal.* **115**:4 (1991), 311–328.

[Prasad and Rajagopal 2006] S. C. Prasad and K. R. Rajagopal, "On the diffusion of fluids through solids undergoing large deformations", *Math. Mech. Solids* **11**:3 (2006), 291–305.

[Rajagopal 2003] K. R. Rajagopal, "Diffusion through polymeric solids undergoing large deformations", *Mat. Sci. Technol.* **19**:9 (2003), 1175–1180.

[Rajagopal and Srinivasa 2012] K. R. Rajagopal and A. Srinivasa, "Restrictions placed on constitutive relations by angular momentum balance and Galilean invariance", *Z. Angew. Math. Phys.* (2012).

[Rivlin 1977] R. S. Rivlin, "Some research directions in finite elasticity theory", *Rheol. Acta* **16**:2 (1977), 101–112. Reprinted as pp. 418–429 of his *Collected papers*, vol. 1, edited by G. I. Barenblatt and D. D. Joseph, Springer, Berlin, 1997.

[Shi et al. 1981] J. J. Shi, K. R. Rajagopal, and A. S. Wineman, "Application of the theory of interacting continua to the diffusion of a fluid through a nonlinear elastic medium", *Int. J. Engng. Sci.* **19**:6 (1981), 871–889.

[Steigmann 2002] D. J. Steigmann, "Invariants of the stretch tensors and their application to finite elasticity theory", *Math. Mech. Solids* **7**:4 (2002), 393–404.

[Steigmann 2003] D. J. Steigmann, "On isotropic, frame-invariant, polyconvex strain-energy functions", *Quart. J. Mech. Appl. Math.* **56**:4 (2003), 483–491.

[Steigmann and Pipkin 1988] D. J. Steigmann and A. C. Pipkin, "Stability of harmonic materials in plane strain", *Quart. Appl. Math.* **46**:3 (1988), 559–568.

[Treloar 1975] L. R. G. Treloar, *The physics of rubber elasticity*, 3rd ed., Oxford University Press, 1975.

[Truesdell 1962] C. Truesdell, "Mechanical basis of diffusion", *J. Chem. Phys.* **37**:10 (1962), 2336–2344.

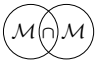
[Truesdell 1977] C. A. Truesdell, *A first course in rational continuum mechanics, 1: General concepts*, Academic Press, New York, 1977.

[Weitsman 1987] Y. Weitsman, "Stress assisted diffusion in elastic and viscoelastic materials", *J. Mech. Phys. Solids* **35**:1 (1987), 73–94.

Received 8 May 2012. Revised 25 Jul 2012. Accepted 26 Sep 2012.

GARY J. TEMPLET: gjtempl@berkeley.edu
Department of Mechanical Engineering, University of California, Berkeley, Berkeley, CA 94720, United States

DAVID J. STEIGMANN: steigman@me.berkeley.edu
Department of Mechanical Engineering, University of California, Berkeley, Berkeley, CA 94720, United States



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. 1 no. 1 2013

Dislocations, imperfect interfaces and interface cracks in anisotropic elasticity for quasicrystals 1
Xu Wang and Peter Schiavone

Localization of point vortices under curvature perturbations 19
Roberto Garra

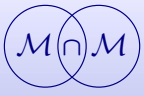
Contraction of the proximal map and generalized convexity of the Moreau–Yosida regularization in the 2-Wasserstein metric 33
Eric A. Carlen and Katy Craig

Ptolemy's longitudes and Eratosthenes' measurement of the earth's circumference 67
Lucio Russo

TV-min and greedy pursuit for constrained joint sparsity and application to inverse scattering 81
Albert Fannjiang

On the theory of diffusion and swelling in finitely deforming elastomers 105
Gary J. Templet and David J. Steigmann

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(2013)1:1;1-I