Springer INdAM Series 4

Franco Brezzi
Piero Colli Franzone
Ugo Gianazza
Gianni Gilardi *Editors*

Analysis and Numerics of Partial Differential Equations

Springer

Springer INdAM Series

Volume 4

For further volumes:
www.springer.com/series/10283

Franco Brezzi • Piero Colli Franzone •
Ugo Gianazza • Gianni Gilardi
Editors

Analysis and Numerics of Partial Differential Equations

Springer

Editors
Franco Brezzi
Istituto di Matematica Applicata e Tecnologie Informatiche (IMATI)
CNR Pavia
Pavia, Italy

Ugo Gianazza
Dipartimento di Matematica "F. Casorati"
Università degli Studi di Pavia
Pavia, Italy

Piero Colli Franzone
Dipartimento di Matematica "F. Casorati"
Università degli Studi di Pavia
Pavia, Italy

Gianni Gilardi
Dipartimento di Matematica "F. Casorati"
Università degli Studi di Pavia
Pavia, Italy

ISSN 2281-518X
ISSN 2281-5198 (electronic)
Springer INdAM Series
ISBN 978-88-470-2591-2
ISBN 978-88-470-2592-9 (eBook)
DOI 10.1007/978-88-470-2592-9
Springer Milan Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012951305

In memory of Enrico Magenes

Enrico Magenes (courtesy of the family)

Preface

On November 2nd 2010 Enrico Magenes passed away.

One year later, some of his friends, collaborators and former students organized a three-day conference, in order to celebrate his memory and give a first assessment of his deep influence on contemporary Mathematics. All the speakers were experts in the analysis and numerics of partial differential equations, who had directly interacted with Magenes, during his long career.

The present volume is a direct offshoot of that meeting, and it collects the main contributions offered in that occasion, properly revised and expanded. It consists of two parts: the first one gives a wide historical perspective of Magenes' work; the second one contains original research or survey papers, and shows how ideas, methods, and techniques introduced by Magenes and his collaborators still have an impact on the current research in Mathematics. As agreed between Springer and UMI (Unione Matematica Italiana), some of the papers appearing in the second part will be fully published on Bollettino UMI as well.

Although it is still too early to fully appreciate Magenes' legacy, nonetheless the volume is a first attempt to present a comprehensive survey of his activity in Mathematics. At the same time, from Magenes' peculiar point of view, it is a broad perspective of the research in partial differential equations and their applications developed in Italy in the period 1950–2000.

The editors are grateful to Francesca Bonadei of Springer Italy for the unique opportunity offered with the publication of this volume, and for her constant support during its preparation.

Pavia, Italy

Franco Brezzi
Piero Colli Franzone
Ugo Gianazza
Gianni Gilardi

Contents

List of Contributors

Luigi Ambrosio Scuola Normale Superiore, Pisa, Italy

Douglas N. Arnold School of Mathematics, University of Minnesota, Minneapolis, MN, USA

Claudio Baiocchi Dipartimento di Matematica "F. Casorati", Università di Pavia, Pavia, Italy; Dipartimento di Matematica "G. Castelnuovo", Università di Roma "La Sapienza", Rome, Italy

Katayun Barmak Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA

Andrea Bonito Department of Mathematics, Texas A&M University, College Station, TX, USA

Franco Brezzi Istituto di Matematica Applicata e Tecnologie Informatiche "E. Magenes", Consiglio Nazionale delle Ricerche, Pavia, Italy

Luis Caffarelli Department of Mathematics, University of Texas at Austin, Austin, TX, USA

Claudio Canuto Dipartimento di Scienze Matematiche, Politecnico di Torino, Turin, Italy

J. Manuel Cascón Departamento de Economía e Historia Económica, Universidad de Salamanca, Salamanca, Spain

Piero Colli Franzone Dipartimento di Matematica "F. Casorati", Università di Pavia, Pavia, Italy

Eva Eggeling Visual Computing, Fraunhofer Austria Research GmbH, Graz, Austria

Maria Emelianenko Department of Mathematical Sciences, George Mason University, Fairfax, VA, USA

Yekaterina Epshteyn Department of Mathematics, The University of Utah, Salt Lake City, UT, USA

Mario Ferrari Dipartimento di Matematica "F. Casorati", Università di Pavia, Pavia, Italy

Giuseppe Geymonat LMS, Laboratoire de Mécanique des Solides UMR 7649, École Polytechnique, Palaiseau Cedex, France

Ugo Gianazza Dipartimento di Matematica "F. Casorati", Università di Pavia, Pavia, Italy

Nicola Gigli Université de Nice, Mathématiques, Nice, France

David Kinderlehrer Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

Toni Lassila MATHICSE-CMCS, Modelling and Scientific Computing, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Yvon Maday Laboratoire Jacques-Louis Lions, UMR 7598, UPMC Univ Paris 06, Paris, France; Division of Applied Mathematics, Institut Universitaire de France and Brown University, Providence, RI, USA

Andrea Manzoni SISSA Mathlab, SISSA—International School for Advanced Studies, Trieste, Italy

Pedro Morin Instituto de Matemática Aplicada del Litoral (IMAL), Güemes 3450 and Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santa Fe, Argentina

Umberto Mosco Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA, USA

Olga Mula Laboratoire Jacques-Louis Lions, UMR 7598, UPMC Univ Paris 06, Paris, France; CEA Saclay—DEN/DANS/DM2S/SERMA/LLPR, Gif-Sur-Yvette Cedex, France

Ricardo H. Nochetto Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA

Alfio Quarteroni MATHICSE-CMCS, Modelling and Scientific Computing, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; MOX, Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy

Gianluigi Rozza MATHICSE-CMCS, Modelling and Scientific Computing, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; SISSA Mathlab, SISSA—International School for Advanced Studies, Trieste, Italy

Giuseppe Savaré Dipartimento di Matematica "F. Casorati", Università di Pavia, Pavia, Italy

Richard Sharp Microsoft Corporation, Redmond, WA, USA

Shlomo Ta'asan Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

Enrico Valdinoci Dipartimento di Matematica "F. Enriques", Università di Milano, Milan, Italy; Istituto di Matematica Applicata e Tecnologie Informatiche "E. Magenes", Consiglio Nazionale delle Ricerche, Pavia, Italy

Marco Verani MOX-Dipartimento di Matematica, Politecnico di Milano, Milan, Italy

Claudio Verdi Dipartimento di Matematica "F. Enriques", Università di Milano, Milan, Italy

Augusto Visintin Dipartimento di Matematica, Università di Trento, Trento, Italy

Part I
A Historical Perspective

Personal Memories

Franco Brezzi

Abstract Few personal memories about Enrico Magenes as scientific mentor, chairman of the Institute of Numerical Analysis, President of the Italian Mathematical Union.

Enrico Magenes taught me A LOT, both as a mathematician and as a man.

This is not obviously the place to tell what I learned in Mathematics. It is a fact, however, that I tried to learn, at least in part, many of his human gifts: his commitment at work, his sincerity, his love for the talent, his warm-heartedness, a lot of things. Maybe the most important thing that I tried to learn from him, was "not to hate." As a matter of fact, I saw him a lot of times losing patience, but I never had the impression he hated anybody. There were things he disapproved of. There were persons he did not think much of. However, he never hated anybody, not even those, who had put him in a concentration camp.

For me, this was particularly important, and I did my best, in order to learn it.

Under a more general point of view, for us, his students, he was an example and a stimulus. In my opinion, his best scientific talent was one of those gifts, which are not frequently extolled. However, still in my opinion, it is perhaps the most important talent in a mentor: the ability to recognize the important problems. He could sense the scientific directions, that would generate a lot of important developments, and those that would extinguish after a couple of papers. It is in this way, with extreme far-sightedness, that he could open new ways, combining mathematical rigor with interest for applications, and starting a lot of fruitful collaborations with engineers, biologists and physicians. These are things, that at the time were almost revolutionary.

As a matter of fact, he never forced anybody to work on a specific topic. He just limited himself to suggesting the problems he considered important enough to be solved. All of us were obviously EXTREMELY careful in following his advices.

F. Brezzi (✉)
Istituto di Matematica Applicata e Tecnologie Informatiche "E. Magenes", Consiglio Nazionale delle Ricerche, via Ferrata 1, 27100, Pavia, Italy
e-mail: brezzi@imati.cnr.it

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*,
Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_1,

I was his (unworthy) successor in two positions, which I consider important: the chairmanship of the Institute of Numerical Analysis of the Italian Research Council (CNR) (later on merged in the current Institute of Applied Mathematics and Information Technologies, still of the CNR) which I took in 1992, immediately after he finished his term, and the Presidency of the Italian Mathematical Union, which I took thirty years later (he finished his term in 1976, I started mine in 2006).

In the first instance, notwithstanding my commitment, the comparison was merciless. I try and console myself, telling me that nobody could possibly rival his enthusiasm, his rigor, his vision about Mathematics, and his humanity. Moreover, I tell myself that few would have been able, so to speak, to lose by a narrow margin. Anyway, the comparison was really hard, and I was lucky that he was there, always at disposal for suggestions, advices, warnings. As a matter of fact, also through the activities of the Institute (which originally was called Laboratory), Magenes could give birth to a large part of the Italian Applied Mathematics (in particular, Numerical Analysis), placing it among the highest ranking international positions: just preserving this ranking was very difficult.

The Italian Mathematical Union (UMI) owes him a lot, not only for the work done as a member of the Scientific Committee (from 1967 to 1979) and then as President in difficult and stormy times (from 1973 to 1976). For UMI Magenes was always a stimulus, a source of ideas and initiatives, and at the same time a balanced and wise presence, witness of a clear and solid vision, of what was important, and what not. Even when he did not hold any institutional office, at national level, all the same he was a reference point, somebody who could encourage or warn, a sort of Guardian Angel.

With him a large piece of history passes away, an important piece, not only for Mathematics. Good Bye Enrico!

Some Aspects of the Research of Enrico Magenes in Partial Differential Equations

Giuseppe Geymonat

Abstract The author traces the initial stage of Enrico Magenes's research, with a particular emphasis on his work in Partial Differential Equations. The very fruitful collaborations with G. Stampacchia and J.-L. Lions are clearly presented.

1 The Beginnings in Modena

The first researches of Enrico Magenes in Partial Differential Equations date to 1952, [14, 15] (and in the same year he became professor of Mathematical Analysis at the University of Modena). Their argument is the application to the heat equation of a method that in the Italian School is called "Picone's Method". The basic idea of the method is to transform the boundary value problem into a system of integral equations of Fischer-Riesz type. This idea was introduced by Picone around 1935 and then deeply applied by Amerio, Fichera, and many others to elliptic equations. For simplicity, we present the method in the simplest case of a non-homogeneous Dirichlet problem in a smooth, bounded domain $\Omega \subset \mathbb{R}^N$:

$$A(u) = f \quad \text{in } \Omega, \qquad \gamma_0 u = g \quad \text{on } \Gamma := \partial\Omega \tag{1}$$

where $A(u)$ is a second order linear elliptic operator with smooth coefficients and $\gamma_0 u$ denotes the trace of u on Γ; for simplicity one can also assume that uniqueness holds true for this problem. Let $A^\star$ be the formal adjoint of A and let $\frac{\partial}{\partial \nu}$ denote the so-called co-normal derivative (when $A = \triangle$, then $A^\star = \triangle$ and $\nu = \mathbf{n}$, the outgoing normal to Γ). The Green formula states

$$\int_\Omega A(u) w dx - \int_\Omega u A^\star(w) dx = \int_\Gamma \left(\frac{\partial u}{\partial \nu} w - u \frac{\partial w}{\partial \nu} \right) d\Gamma. \tag{2}$$

Hence, if one knows a sequence w_n of smooth enough functions, such that $A^\star(w_n)$ and $\gamma_0(w_n)$ both converge, then, thanks to (2), the determination of the vector

G. Geymonat (✉)
LMS, Laboratoire de Mécanique des Solides UMR 7649, École Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France
e-mail: giuseppe.geymonat@lms.polytechnique.fr

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*, Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_2,

$(u, \frac{\partial u}{\partial \nu})$ with u the solution of (1) is reduced to the solution of a system of linear equations of Fischer-Riesz type.

The difficulty was naturally to find such a sequence, to prove its completeness (in a suitable functional space) and hence the space where the corresponding system is solvable and so the problem (1).

Following Amerio [1], let the coefficients of A be smoothly extended to a domain $\widehat{\Omega} \supset \Omega$ and for every fixed $R \in \widehat{\Omega}$ let $F(P, R)$, as a function of P, be the fundamental solution of $A^\star(w) = 0$ (moreover, such a fundamental function can be chosen so that, as a function of R, it also satisfies $A(u) = 0$). Then from (1) and (2) it follows that for every $Q \in \widehat{\Omega} \setminus \Omega$

$$0 = \int_\Gamma \left(u(x) \frac{\partial F(x, Q)}{\partial \nu} - \frac{\partial u(x)}{\partial \nu} F(x, Q) \right) d\Gamma - \int_\Omega f(x) F(x, Q) dx. \tag{3}$$

This equation gives a necessary compatibility condition between $\gamma_0 u$ and $\frac{\partial u(x)}{\partial \nu}$. Moreover, if φ_n is a sequence of "good" functions defined in $\overline{\widehat{\Omega} \setminus \Omega}$, then one can take $w_n(P) = \int_{\overline{\widehat{\Omega} \setminus \Omega}} \varphi_n(x) F(P, x) dx$. Two problems remain:

(i) the choice of the sequence φ_n, in order that the procedure can be applied;
(ii) the determination of the good classes of data f, g, Ω, and solutions u to which the procedure can be applied.

In this context it is also useful to recall that for every $P \in \Omega$ it holds

$$\frac{2\pi^{N/2}}{\Gamma(N/2)} u(P) = \int_\Gamma \left(u(x) \frac{\partial F(x, P)}{\partial \nu} - \frac{\partial u(x)}{\partial \nu} F(x, P) \right) d\Gamma - \int_\Omega f(x) F(x, P) dx. \tag{4}$$

In order to study the previous problems, one has to study the fine properties of the simple and double layer potentials, appearing in (3) and (4). See for instance Fichera's paper [3], where many properties are studied, and in particular results of completeness are proved. (The modern potential theory studies the fine properties of the representation (4) for general Lipschitz domains and in a L_p framework.)

Magenes applied the method to the heat operator $E(u) = \triangle u - \frac{\partial u}{\partial t}$ in $\Omega \times (0, T)$, whose formal adjoint is $E^\star(u) = \triangle u + \frac{\partial u}{\partial t}$; the Green formula (2) becomes

$$\begin{aligned} &\int_0^T \int_\Omega E(u) w dx dt - \int_0^T \int_\Omega u E^\star(w) dx dt \\ &\quad = \int_0^T \int_\Gamma \left(\frac{\partial u}{\partial \mathbf{n}} w - u \frac{\partial w}{\partial \mathbf{n}} \right) d\Gamma dt \\ &\qquad + \int_\Omega u(T) w(T) dx - \int_\Omega u(0) w(0) dx. \end{aligned} \tag{5}$$

Following the approach of Amerio and Fichera, Magenes used the fundamental solution of the heat equation, defined by $F(x, t; x', t') = \frac{1}{t'-t} \exp(-\frac{\|x'-x\|}{4(t'-t)})$ for $t' > t$ and $F(x, t; x', t') = 0$ for $t' \leq t$. He also defined a class of solutions of the heat

equation $E(u) = f$, assuming the boundary value in a suitable way and represented by potentials of simple and double layer.

These researches were followed [17] by the study of the so-called mixed problem, where the boundary is splitted in two parts: in the first one the boundary condition is of Dirichlet type, and in the other one the datum is the co-normal derivative. This problem was of particular difficulty for the presence of discontinuity in the data, even in the stationary case: see [16] (where the results are stated for $N = 2$, although they are valid for arbitrary N) and has stimulated many researches using potential theory not only of Magenes (see e.g. [18, 19]) but also of Fichera, Miranda, Stampacchia, ...

2 The Years in Genoa with G. Stampacchia

From the historical point of view, these researches show the change of perspective that occurred in Italy at that time in the study of these problems with the use of

- some first type of trace theorems (e.g. inspired by the results of Cimmino [2]);
- the introduction of the concept of weak solution;
- the use of general theorems of functional analysis (see e.g. [4]).

Under this point of view, the following summary of a conference of Magenes gives a typical account (see [20]). *Breve esposizione e raffronto dei più recenti sviluppi della teoria dei problemi al contorno misti per le equazioni alle derivate parziali lineari ellittiche del secondo ordine, soprattutto dal punto di vista di impostazioni "generalizzate" degli stessi* (*A short presentation and comparison of the most recent developments in the theory of mixed boundary value problems for second order elliptic linear partial differential equations, mainly from the point of view of "generalized" approaches to them*).

At the end of 1955 Magenes left the University of Modena for the University of Genoa, where he had G. Stampacchia as colleague. Stampacchia was a very good friend of Magenes from their years as students at Scuola Normale, since both where antifascist. Moreover, Magenes and Stampacchia were well aware of the fundamental change induced by the distribution theory and the Sobolev spaces in the calculus of variations and in the study of partial differential equations, particularly in the study of boundary value problems for elliptic equations (see for instance the bibliography of [20]).

They studied the works of L. Schwartz and its school, and specially the results on the mixed problem in the Hadamard sense. At the first *Réunion des mathématiciens d'expression latine*, in September 1957, Magenes and Stampacchia met J.-L. Lions. It was the beginning of a friendship, that would never stop. During the Spring 1958, J.-L. Lions gave at Genoa a series of talks on the mixed problems [5, 6], and in June 1958 Magenes and Stampacchia completed a long paper [22], that would have a fundamental influence on the Italian researches on elliptic partial differential equations. Indeed, that paper gives a general presentation of the results obtained up to that moment in France, United States, Sweden, Soviet Union by N. Aronszajn,

F.E. Browder, G. Fichera, K.O. Friedrichs, L. Gårding, O. Ladyzenskaja, J.-L. Lions, S.G. Mikhlin, C.B. Morrey Jr., L. Nirenberg, M.I. Visik, ...

It is worth giving the titles of the four chapters: *I General notions, II Boundary value problems for linear elliptic equations, methods with finite "Dirichlet integral," III Problems of regularization, IV Other approaches to boundary value problems*. Then, in the following few years Magenes tried to increase the audience of this methodology in the Italian mathematical community, giving lectures in various universities (see e.g. [21]), for instance organizing with Stampacchia a CIME course on distribution theory in 1961, ...

At the end of 1959 Magenes left Genoa and went to the University of Pavia. During the year 1959, the collaboration with J.-L. Lions became more active and they started a long series of joint works [7–10], whose results were summarized and fully developed in a series of books [11–13] translated in Russian, English and Chinese.

3 The Collaboration with J.-L. Lions in the Study of Boundary Value Problems

Following an idea of J. Hadamard, Courant and Hilbert (*Methods of Mathematical Physics*, volume II, Interscience, 1962, p. 227) state that *a mathematical problem, which must correspond to a physical reality, should satisfy the following basic requirements*:

1. *The solution must exist.*
2. *The solution should be uniquely determined.*
3. *The solution should depend continuously on the data* (*requirement of stability*).

In order to satisfy these requirements, one has to identify the functional spaces where the problems are well-posed. The distribution theory and the Sobolev spaces give a natural framework and the instruments to study partial differential equations. The results collected in the first three chapters of [22] allow to prove that the elliptic boundary value problems *with homogeneous boundary data* are well posed in Sobolev spaces $W^{m,2}(\Omega)$ *with m big enough*. For non-homogeneous boundary data, the situation was more difficult, since at first it was necessary to give a *good* definition of the trace $\gamma_0 u$ of an element $u \in W^{m,p}(\Omega)$ on $\Gamma := \partial\Omega$. The good definitions and the corresponding characterizations were given (under various conditions on $p \geq 1$, on $m > 1 - \frac{1}{p}$ and on the regularity of the domain Ω, i.e. of its boundary Γ) by E. Gagliardo, J.-L. Lions, and P.I. Lizorkin, G. Prodi, ... In particular it was proved that *the trace operator cannot be continuously defined on* $L^2(\Omega)$.

However, for many problems coming from the applications (e.g. mechanics, engineering, ...) the natural setting is in Sobolev spaces of low order and sometimes of negative order. Therefore, it is necessary to define a *weak or generalized* solution of a non-homogeneous boundary value problem and hence, to give a *good definition of trace in a weak sense*.

Inspired by the theory of distributions, Lions and Magenes [7–10] tackled the problem by duality. More precisely let us consider the map $u \to \mathcal{A}u := \{Au, B\gamma_0 u\}$, where A is a linear elliptic operator with smooth coefficients defined in a domain $\Omega \subset \mathbf{R}^{\mathbf{N}}$ with smooth boundary Γ, and $B\gamma_0$ is a linear differential operator with smooth coefficients, defined on Γ, and compatible with A in a suitable sense.

Such a general framework is a "natural" extension of the Dirichlet problem (1). Thanks to known regularity results (described for instance in Chap. III of [22]), the map $\mathcal{A} : E(\Omega) \to F(\Omega) \times G(\Gamma)$ is an isomorphism (for simplicity, and in general a finite index operator) between the Sobolev spaces $E(\Omega)$ and $F(\Omega) \times G(\Gamma)$, where these spaces are of big enough positive order.

In the case of (1), one can take for instance $E(\Omega) = H^{m+2}(\Omega)(= W^{m+2,2}(\Omega))$ with $m \geq 0$, and then $F(\Omega) = H^m(\Omega)$ and $G(\Gamma) = H^{m+3/2}(\Gamma)$. By restriction to the case of homogeneous boundary data and to the space $F_0(\Omega)$ (closure of $\mathcal{D}(\Omega)$ into $F(\Omega)$), it is possible to define the isomorphism $\mathcal{A}_\sharp : X(\Omega) \to F_0(\Omega)$, where $X(\Omega)$ is a subspace of $E(\Omega)$.

By transposition, for every linear and continuous form $L(v)$ on $X(\Omega)$, there exists $u \in (F_0(\Omega))'$ such that

$$\left\langle u, \mathcal{A}_\sharp(v) \right\rangle = L(v) \quad \text{for all } v \in X(\Omega). \tag{6}$$

Let us point out that $(F_0(\Omega))'$ is a Sobolev space of *negative* order (in the case of (1) $F_0(\Omega) = H_0^m(\Omega)$ and $(F_0(\Omega))' = H^{-m}(\Omega)$). In order to get the wanted result, Lions and Magenes chose $L = L_1 + L_2$ in such a way that L_1 gives rise to the equation $A^\star u = f$, where $A^\star$ is the linear elliptic operator formally adjoint to A, and L_2 corresponds to the non-homogeneous boundary conditions $B^\star u = g$ in the most natural way.

Perhaps the most interesting contribution of Lions and Magenes was the optimal choice of L_2. It was obtained thanks to a clever use of the Green formula, that allows to naturally define the traces of every element $u \in (F_0(\Omega))'$, such that $A^\star u$ belongs to a suitable distribution space on Ω.

For instance, in the case of (1) with $A = A^\star = \triangle$, $B\gamma_0 = \frac{\partial}{\partial \mathbf{n}} := \gamma_1$ and $m = 0$, one can define the trace $\gamma_0 u \in H^{-1/2}(\Gamma)$ for every $u \in L^2(\Omega)$, such that $A^\star u = \triangle u \in L^2(\Omega)$. The main steps of the proof are the following:

1. One proves the density of $\mathcal{D}(\overline{\Omega})$ into the space $Y(\Omega) := \{u \in L^2(\Omega); \triangle u \in L^2(\Omega)\}$, equipped with the natural graph norm.
2. Let us define $X(\Omega) = \{v \in H^2(\Omega); \gamma_0 u = 0\}$ and let us remark that the map $v \longrightarrow \gamma_1 v$ is a linear and continuous map of $X(\Omega)$ onto $H^{1/2}(\Gamma)$, whose kernel is $H_0^2(\Omega)$.
3. For every $(u, \phi) \in Y(\Omega) \times H^{1/2}(\Gamma)$, one defines the bilinear and bi-continuous map $L_2(u, \phi)$ with

$$L_2(u, \phi) = \int_\Omega u \triangle v_\phi dx - \int_\Omega \triangle u v_\phi dx,$$

where $v_\phi \in X(\Omega)$ is such that $\gamma_1 v_\phi = \phi$ (it is easy to verify that indeed $L_2(u, \phi) = 0$ when $v_\phi \in H_0^2(\Omega)$ and hence, $L_2(u, \phi)$ does not depend on the particular choice of v_ϕ).

4. One can do the identification
$$L_2(u,\phi) = \langle Tu, \phi \rangle,$$
where $\langle \bullet, \bullet \rangle$ denotes the duality pairing between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$, and $u \longrightarrow Tu$ is linear and continuous from $Y(\Omega)$ to $H^{-1/2}(\Gamma)$.
5. When $u \in \mathcal{D}(\overline{\Omega})$, then the Green formula (2) implies
$$L_2(u,\phi) = \int_\Gamma u\phi d\Gamma$$
and hence, the map T can be identified with the trace map.

The books [11] and [12] present the general theory, not only for elliptic operators, but also for linear evolution equations of parabolic type, both in distributions spaces and also [13] in ultra-distributions of Gevrey classes.

References

1. Amerio, L.: Sul calcolo delle soluzioni dei problemi al contorno per le equazioni lineari del secondo ordine di tipo ellittico. Am. J. Math. **69**, 447–489 (1947)
2. Cimmino, G.: Sulle equazioni lineari alle derivate parziali di tipo ellittico. Rend. Semin. Mat. Fis. Milano **23**, 1–23 (1952)
3. Fichera, G.: Teoremi di completezza sulla frontiera di un dominio per taluni sistemi di funzioni. Ann. Mat. Pura Appl. (4) **27**, 1–28 (1948)
4. Fichera, G.: Methods of functional linear analysis in mathematical physics: "a priori" estimates for the solutions of boundary value problems. In: Proceedings of the International Congress of Mathematicians, vol. III, Amsterdam, 1954, pp. 216–228. North-Holland, Amsterdam (1956)
5. Lions, J.-L.: Problemi misti nel senso di Hadamard classici e generalizzati. Rend. Semin. Mat. Fis. Milano **28**, 149–188 (1958)
6. Lions, J.-L.: Problemi misti nel senso di Hadamard classici e generalizzati. Rend. Semin. Mat. Fis. Milano **29**, 235–239 (1959)
7. Lions, J.-L., Magenes, E.: Problemi ai limiti non omogenei. I. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (3) **14**, 269–308 (1960)
8. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes. II. Ann. Inst. Fourier (Grenoble) **11**, 137–178 (1961)
9. Lions, J.-L., Magenes, E.: Problemi ai limiti non omogenei. III. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (3) **15**, 41–103 (1961)
10. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes. VII. Ann. Mat. Pura Appl. (4) **63**, 201–224 (1963)
11. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes et applications, vol. 1. Dunod, Paris (1968)
12. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes et applications, vol. 2. Dunod, Paris (1968)
13. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes et applications, vol. 3. Dunod, Paris (1970)
14. Magenes, E.: Sull'equazione del calore: teoremi di unicità e teoremi di completezza connessi col metodo di integrazione di M. Picone, Nota I. Rend. Semin. Mat. Univ. Padova **21**, 99–123 (1952)
15. Magenes, E.: Sull'equazione del calore: teoremi di unicità e teoremi di completezza connessi col metodo di integrazione di M. Picone, Nota II. Rend. Semin. Mat. Univ. Padova **21**, 136–170 (1952)

16. Magenes, E.: Sui problemi al contorno misti per le equazioni lineari del secondo ordine di tipo ellittico. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (3) **8**, 93–120 (1954)
17. Magenes, E.: Problemi al contorno misti per l'equazione del calore. Rend. Semin. Mat. Univ. Padova **24**, 1–28 (1955)
18. Magenes, E.: Problema generalizzato di Dirichlet e teoria del potenziale. Rend. Semin. Mat. Univ. Padova **24**, 220–229 (1955)
19. Magenes, E.: Sulla teoria del potenziale. Rend. Semin. Mat. Univ. Padova **24**, 510–522 (1955)
20. Magenes, E.: Recenti sviluppi nella teoria dei problemi misti per le equazioni lineari ellittiche. Rend. Semin. Mat. Fis. Milano **27**, 75–95 (1957)
21. Magenes, E.: Sui problemi al contorno per i sistemi di equazioni differenziali lineari ellittici di ordine qualunque. Univ. Politec. Torino. Rend. Semin. Mat. **17**, 25–45 (1957/1958)
22. Magenes, E., Stampacchia, G.: I problemi al contorno per le equazioni differenziali di tipo ellittico. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (3) **12**, 247–358 (1958)

Enrico Magenes and the Dam Problem

Claudio Baiocchi

Abstract In a very vivid way, the author describes how Magenes was introduced to the Dam Problem, and how the group around Magenes, made up by pure analysts, numerical analysts, mathematical physicists, worked hard to provide a solution that could be satisfactory both from a theoretical and practical point of view.

Enrico Magenes, Director of the Institute of Numerical Analysis of the Italian Research Council (IAN), which he had founded, was always looking out for new problems, that might inspire the Institute research.

In 1970 the School of Engineering of the University of Pavia hired Ugo Maione, an engineer, as professor of Hydraulics. Maione had a very good reputation, and Magenes immediately seized the opportunity: he contacted the new professor, and suggested a meeting, in order to examine the possibility of starting a collaboration between the two Institutes.

Maione gladly accepted, and at the first meeting, that took place in Magenes's office, amongst others were present also Valeriano Comincioli, as numerical analyst, and Luciano Guerri, as mathematical physicist; as for me, I was supposed to be the pure mathematician: indeed, one of Magenes's strongest conviction was always the idea, that within the Institute research activities, the numerical treatment of every single problem should come together and rely upon a deep theoretical analysis.

Maione commenced saying that he had several irons in the fire, but at the moment, the topic he was most interested in was the numerical treatment of a Hydraulics problem, that can be summarized in this way: a pile of dirt separates two water basins at different heights; due to the gravity force, water flows through the pile and, by inserting a proper concrete bulkhead (omitted in Fig. 1), one wants to reduce the dam flow.

C. Baiocchi (✉)
Dipartimento di Matematica "F. Casorati", Università di Pavia, via Ferrata 1, 27100 Pavia, Italy
e-mail: bici.nando@gmail.com

C. Baiocchi
Dipartimento di Matematica "G. Castelnuovo", Università di Roma "La Sapienza", Piazzale Aldo Moro 5, 00185 Rome, Italy

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*,
Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_3,

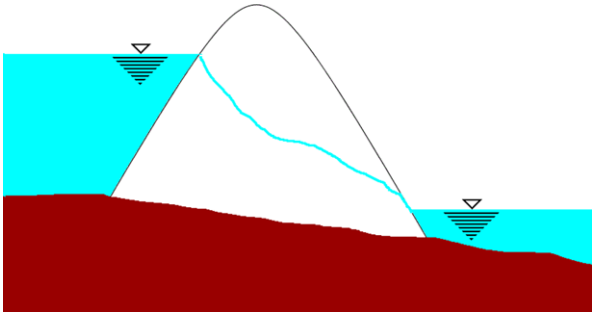


Fig. 1 A dam separates two water basins at different heights

The real problem, which had to do with some dams in Sudan, if I remember correctly, was obviously much more complicated, but Maione was willing to postpone the main difficulties for the moment: for example, assuming the dirt to be homogeneous, the dam vertical section to be constant and geometrically simple, that capillarity and viscosity are negligible, that after a short transient state, the evolution of the phenomenon reaches a steady state, and therefore the *Darcy Law* can be applied ...

Thus, we got to a two-dimensional problem, whose mathematical statement reads in this way (see Fig. 2): the two basins and the rectangular dam, which separates them, lie on a horizontal waterproof basis (the x axis); the curve φ (the so-called *free surface*) bounds from above the wet section Ω of the dam; in such a set, one looks for a function (the so-called *piezometric height*), which must satisfy both the equations in Fig. 2 and further qualitative properties (for example $u > 0$ in Ω); as a matter of fact, from a purely applicative point of view, it suffices to estimate the function φ, as one can easily deduce all the other relevant quantities (*pressure, flow, etc.*) from it.

The only "complication" Maione was not willing to get rid of, was the presence of the waterproof concrete bulkhead: indeed, he claimed that without it, the problem was largely studied (mainly in the Russian literature) by means of conformal transforms, and there was no need to reinvent the wheel.

Somehow we succeeded in convincing him, that starting with the problem without the bulkhead was worth the effort: the numerical results obtained by means of conformal transforms could be a test for a numerical treatment of a different kind, which we could then hopefully adapt to the problem with the bulkhead ...

As a matter of fact, still during the first meeting, a possible numerical approach was sketched by Comincioli and Guerri: on a domain bounded by a "first approximation" $y = \varphi_o(x)$ of the unknown boundary, one solves (obviously numerically) the problem just with the Neumann condition on the portion φ_o of such a boundary; the other condition, namely $u(x, \varphi_o(x)) = \varphi_o(x)$, in general will not be satisfied, but one can define the quantity $\varphi_1(x) = u(x, \varphi_o(x))$ as a new approximation of the unknown function; one then iterates, and hopefully the numerical procedure converges. Such a convergence should actually suggest the possibility that a fixed point theorem provides both an existence result and a numerical justification of the numerical treatment.

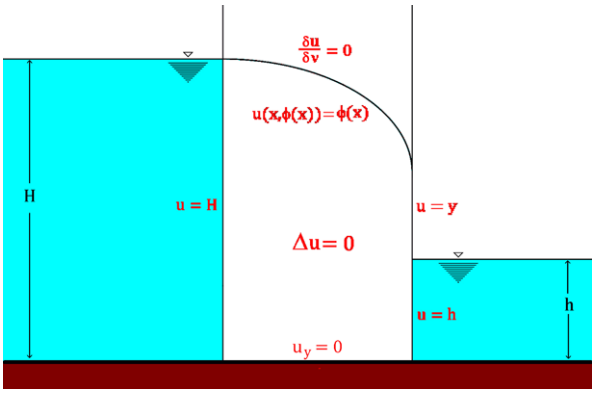


Fig. 2 The two-dimensional, simplified dam problem

Few days later, while I was struggling with the papers that made use of the conformal transform, suddenly I remembered an episode that had taken place a couple of years before: in 1968 Magenes had "invited" me (so to say), one of his young students, to go to Rome with him, in order to take part to a Symposium organized by the Institute of Higher Mathematics.

It is still very fresh in my memory the lecture hall that hosted the symposium: it was filled up with a number of sacred cows (not only Italian) of Mathematics; amongst the names that I can remember, let me mention Felix Browder, Ennio De Giorgi, Gaetano Fichera, Jacques Louis Lions, Enrico Magenes, Carlo Miranda, Louis Nirenberg, Laurent Schwartz, Guido Stampacchia, ...

In particular, a communication presented by Carlo Miranda was about a problem he had been suggested by a colleague of Hydraulics, Russo Spena, and he had dealt with some time before in collaboration with Renato Caccioppoli.

More or less, Miranda started in this way:

> *In a domain whose boundary is given by four rectifiable arcs, we look for a harmonic function, which is constant on two noncontiguous arcs; on the remaining two, the normal derivative vanishes, and finally, on one of the two, a Dirichlet-type condition is given ...*

At that point, from the first row, at the same time two voices called loud, and stopped the presentation: Fichera turned towards his student De Vito, and asked him to take detailed notes; Magenes cried out: "You cannot do that, you have already imposed a different condition on that curve." Miranda replied: "If you let me finish, you will understand," and explained that the said curve was not given, but it was one of the unknowns of the problem ...

When this episode came back to my mind, I rushed to the library and I found Miranda's work printed in the proceedings of the symposium. At a first reading, I realized it was exactly our problem, although without the bulkhead.

In Miranda's approach, the existence result relied on a fixed-point-type theorem, exactly as we hoped we could do with the numerical treatment we had devised in our first meeting; in the meantime, under the guidance of Comincioli and Guerri, this problem had been assigned to a very smart student, Giampiero Volpi, as his thesis dissertation. Let me say here that the student was indeed so smart, that later on IBM stole him from us and hired him ...

Magenes was very happy he had a theoretical basis, on which the numerical treatment could rest, and Maione was immediately summoned, in order to give him the good news, but ... disaster: in Miranda's approach, there was no room for the suspended source! In other words, because of unavoidable physical reasons, the unknown curve must end at a point that lies strictly above the lower basin, whereas in Miranda's work the curve ended exactly at the height of the lower basin.

We were in a deadlock: Magenes was keen to agree with Miranda, not only for the deep esteem he had for him, but also because it was a redrafting of an idea originally due to Caccioppoli; on the other hand, Maione invited us to visit his Hydraulics laboratory; there, we saw an analogical device, where the Hydraulics physical phenomenon was simulated by means of a capillarity phenomenon (Hele-Shaw effect), and the suspended source was clearly visible!

A couple of months later, an idea, that turned out to be very fruitful, allowed me to give a weak formulation to the problem, and independently of the existence or not of the suspended source, everything was recast as a variational inequality; in turn, this provided an existence and uniqueness result for the solution. Then, relying on ideas that were already present in the Russian literature, and by means of a particular technique that Miranda himself had used, I succeeded in proving that the weak solution was indeed a strong solution, and that the suspended source had to exist ...

We were all excited, as the tool represented by variational inequalities lent itself to a numerical approach of the problem, which was at the same time extremely simple under the point of view of programming (an aspect I will talk about in a moment), and offered the possibility to provide a priori estimates for the error; Magenes considered this particular possibility an essential condition in every single numerical approach. However, we were left with a sort of "diplomatic" problem: how could we tell Miranda that his paper contained a mistake? And where exactly was this mistake?

I have to confess that even nowadays I cannot answer the second question, as some points of Miranda's work were somehow confusing; as for the first question, it was Magenes who, with special care, told Miranda what was going on. Miranda's reaction was extremely reasonable: probably Caccioppoli had kept those ideas in the drawer, as some of the details still needed to be polished; Miranda's decision to publish these results, should not damage his Master's reputation! On the other hand, I had used some ideas from that paper, and I wanted to cite it. Promising himself to get back to the problem later on, Miranda begged me to limit myself to an unspecific citation, of the kind "for a similar problem, see ...," and I gladly accepted this suggested solution.

From Maione's point of view, although still not completely satisfactory, as the bulkhead was not there, the result was very much appreciated: even if he did not have the necessary expertise to appreciate the correctness of the approximation, he quickly realized that there was no more any need to discretize the differential equation on a sequence of domains, in each of whom the shape of the domain depended on the solution of the previous problem: the solution on the whole rectangle of a single variational inequality provided all the looked-for quantities, free-surface curve included!

Moreover, the comparison with the discretization on variable domains was extremely favorable under every point of view; besides largely improving the memory management and the processing time, the new approach cut the programming complexity down to the bone: even I could write down such a program!

We were all euphoric for this achievement, and the argument looked very promising; it was therefore time to devote ourselves to the study of more sophisticated problems (primarily the bulkhead). Magenes and Maione quickly agreed on a series of joint seminars: we would explain Sobolev spaces to engineers, and engineers would present us the main points one needs to know in the study of the motion of fluids. Unfortunately, in a short time, the number of engineers quickly shrank just to Maione; on the contrary, probably because Magenes had been extremely clear in inviting us to take part to the seminars, the mathematicians' group remained compact. Frankly speaking, I cannot say how many of the notions Maione explained us really entered the participants' heads (in my case I have some vague remembrances about Bernoulli's law and the water hammer effect), but from the point of view of scientific productivity, the results were noteworthy, and achieved in a very short time: indeed, we had to scoop the competition, because Magenes, as he used to do, had set up a real advertising campaign (organization of a CIME Summer course, a set of seminars at SISSA, talks at meetings and conferences ...), so that in a short time this research topic was brought to the attention of the Italian mathematical community; this was particularly easy, because variational inequalities and their applications were at that time a peak topic.

I cannot mention here all the names of the persons that gave a contribution to this enterprise, and I strongly apologize for this with all the colleagues; I will limit myself to say that the "Pavia School" quickly succeeded in getting rid of many of the initial "simplifications", providing in this way a solution to almost all the problems, we had originally set apart: not only the insertion of a bulkhead, but also the presence of capillarity, the lack of homogeneity of the building material, the evolution phenomenon ...

Amongst the problem we solved, one, which was particularly valued by Maione, was the extraction of fresh water from a stratum close to the coast (see Fig. 3): the salty water seeps below the fresh water stratum (which is lighter); however, the difference in density of the two fluids is so small, that a small perturbation runs the risk to mix them together, making the precious good unusable.

It is interesting to remark that the context where these results can be applied is much larger: when I presented this problem in an American university, I heard the audience noisily grumbling, till one of the presents raised her hand and asked me: "Where is oil?"

Amongst all the problems Maione has suggested us, only the case of a more complicated geometry took more time to be solved: variational inequalities did not suffice here, and we had to resort to "quasi-variational inequalities" a tool that the French School headed by Jacques Louis Lions was introducing and developing right at that time.

I am going to conclude, mentioning two episodes, that I still remember with great pleasure.

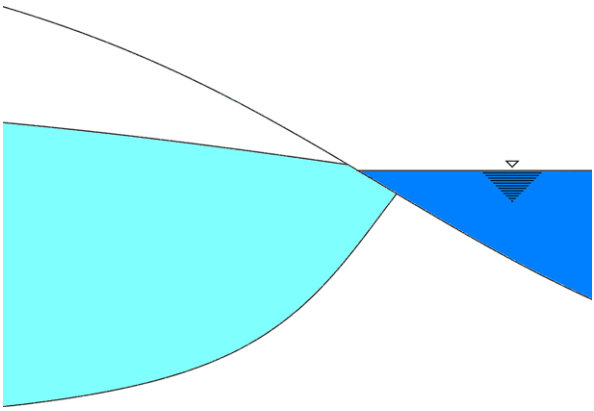


Fig. 3 A fresh water stratum close to the coast

At a Hydraulics meeting, Maione gave a talk, and presented our results: in order to give a more thorough, mathematical description, he registered me too. Nobody knew me there, and therefore, during Maione's seminar, mixed among the participants, I could hear some nipping comments about the importance of mathematical theorems, without which water could not flow ... However, when Maione presented all the details of the numerical simulations and the comparison with different methods, sarcasm disappeared, and also my seminar was listened to, with great attention.

The second episode took place during a mathematical meeting, where I got the following comment by Ennio De Giorgi:

> *It is really true that Mathematics is changing; once upon a time 90 % of the speakers started drawing a potato-like object, and continued saying "Let us consider a domain D ..." Nowadays almost everybody starts, talking about a given domain D, but then she draws half a potato!*

Inverse Problems in Electrocardiology

Piero Colli Franzone

Abstract The author outlines from an historical point of view the mathematical problems, which were the main topics of his scientific collaboration with *Enrico Magenes*. A full account of the joint work on mathematical models for electrocardiology is given. The paper ends commenting few original letters Magenes wrote to the author: it gives a clear description of the work method, and also of the "gentle" pressure to which Magenes's collaborators were exposed.

1 Introduction

In this work, I will outline from an historical point of view the mathematical problems which were the main topics of my scientific collaboration with *Enrico Magenes*. I must first recall and emphasize *Enrico Magenes'* crucial role as promoter and organizer of research activities in the fields that today are known as Scientific Computing, Applied Mathematics, and in particular of Biomedical Applications. He carried out this role in particular during his tenure as Director of the *Laboratorio di Analisi Numerica* (**LAN**, Laboratory of Numerical Analysis), from 1970 to 1993.

Peculiar qualities of *Enrico Magenes*, shared with *Jacques-Louis Lions*, were the scientific open-mindedness and the curiosity toward interesting mathematical models formulated as systems of partial differential equations having a strong impact in significant applications. After having productively collaborated on several important research projects, both scientists reached beyond their own fields of expertise, with the objective to set up and develop the numerical analysis of partial differential equations (PDEs) in a modern functional framework.

Starting from 1970 as the Director of **LAN**, *Magenes* encouraged the initiation of research activities on:

P. Colli Franzone (✉)
Dipartimento di Matematica "F. Casorati", Università di Pavia, via Ferrata 1, 27100 Pavia, Italy
e-mail: colli@imati.cnr.it

P. Colli Franzone
e-mail: piero.collifranzone@unipv.it

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*,
Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_4,

- the analysis of Finite Element **FE** method for the approximation of second order elliptic boundary value problems and of control problems of distributed systems;
- the implementation of **FE** methods and the planning of numerical simulations on bi-dimensional test problems (with simple structured meshes on rectangular domains), in order to evaluate the optimality of the theoretical estimates related to the FE order of convergence.

2 Cardiac Body Surface Maps and Inverse Potential Problems

During the mid-seventies, the Italian Research National Council (**CNR**) launched a set of feasibility studies focused on thematic research projects between C.N.R. Laboratories, Universities and various Industries.

In 1976, the first so-called *Progetti Finalizzati del CNR* (i.e. *Goal-Oriented Projects*) were started and *Magenes* was appointed member of the Scientific Council of the *Progetto Finalizzato Tecnologie Biomediche 1976–1980*, coordinated by Prof. *Luigi Donato*, Director of the CNR *Institute of Clinical Physiology in Pisa*.

Magenes encouraged *Luciano Guerri, Carla Viganotti* and myself to elaborate a project concerning the relationship between electrocardiograms and the excitation process of the heart. As a first step, we considered the so-called **Inverse Problem of Electrocardiography**, consisting in computing the electric potential map on a surface near and surrounding the heart volume from the electric potential measured in various points distributed on the body surface.

The main goal of the project submitted by the **LAN** research team, was to develop numerical algorithms for solving the inverse problem and subsequently to apply the inverse procedure to data measured in experiments on animals. The data were collected by Prof. *Bruno Taccardi* in experiments on isolated dog hearts using an advanced electrical and digital equipment conceived and designed by Prof. *Emilio Gatti* of the *Politecnico di Milano* and assembled at the Centro *C.I.S.E.*. At that time, the results of two experiments on animals were available, performed on two dogs called *Tristano* and *Isotta*.

The research proposal was accepted and included in the subproject *Bio-Immagini 2* concerning the analysis of *Cardiac Body Surface Maps*. The achievement of the **LAN** project goals entailed reaching new methodological and computational goals:

- the development of finite element numerical codes for solving second order elliptic problem on three-dimensional domains;
- the numerical solution of *ill-posed* minimum problems related to quadratic functional subjected to distributed constraints;
- the analysis and application of numerical FE codes to synthetic data associated to test problems, in order to estimate the accuracy of the results and subsequently the application to large experimental data sets.

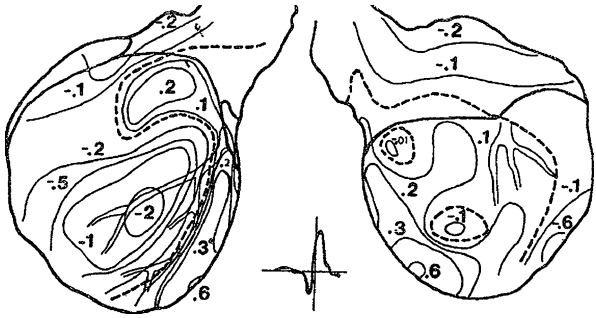


Fig. 1 Epicardial electric equipotential lines in millivolts, drawn manually on a projected anterior and posterior heart surface in dog experiments

Facing these demanding and challenging tasks, at that time I expressed my perplexity to *Magenes* regarding the team effective capability to reach the project goals but, as people acquainted with *Magenes* would have known, his expected answer was: "*Senta, Colli, non faccia la solita mammoletta* (i.e. *Listen, Colli, do not behave as the usual shrinking violet*)."

The data handling of the measured potential body and heart surface maps required, for its analysis and interpretation, the drawing of equipotential lines. At that time, these level lines were drawn manually on the printed potential values in millivolt; an example is displayed in Fig. 1. Subsequently, the **LAN** laboratory acquired a Tektronix 4010 graphical equipment for automatically tracing the equipotential lines on the projected body and heart surfaces, see Fig. 2.

At this point, I would like to insert a letter, which Prof. *Bruno Taccardi* sent me few years ago, containing several memories about *Enrico Magenes*:

"Dear Piero, it is my pleasure to acknowledge the important role played by Professor Magenes *and his school, in the 70's and subsequent years, in promoting advances in mathematics and computer science in the area of normal and pathological cardiac electrophysiology. Our group of electrophysiologists, first at the Free University of Brussels (1951–1959), then at the Simes Institute in Milano and at the Universities of Pavia and Parma* (Taccardi, De Ambroggi, Macchi, Musso et al.) *had observed that the classical 12-lead electrocardiogram, universally used in clinical cardiology, contained only a fraction of the electrical information that is present on the surface of the human body. Meticulous investigations in experimental animals had shown that it was possible to obtain maps that depict the distribution of the electric potential generated by the heart in extracardiac volume conductors (Brussels, 1951, 1958) and on the entire body surface (1962, 1963, Circulation Research). These maps revealed heart abnormalities even in patients with normal 12-lead electrocardiograms. Unfortunately, hundreds of measurements were necessary to obtain the maps, involving a month of manual work for every patient. The measurements were then automated, thanks to the electronic instruments developed at the CISE Institute, under the direction of Prof.* Emilio Gatti. *Furthermore, to convert the signals collected from 240 body surface sites into images readable by a clinician, we needed advanced computing and graphic methods that were not available in any electrophysiological labora-*

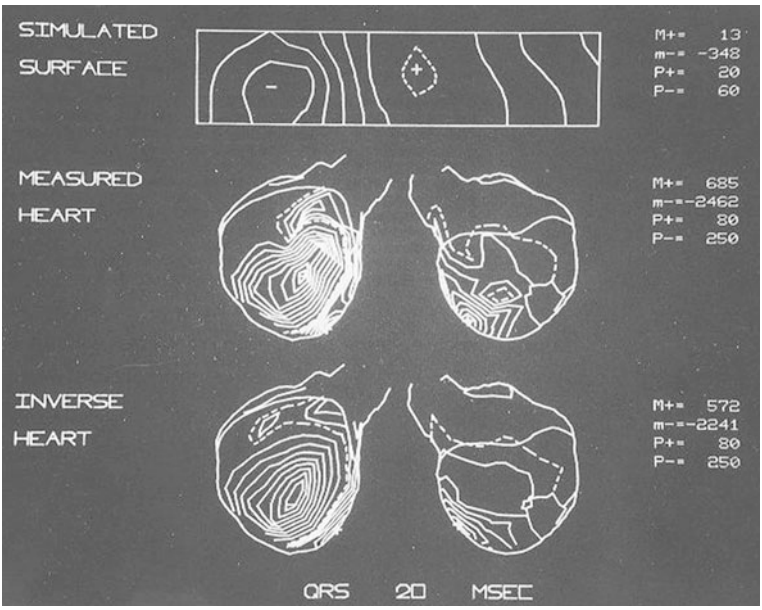


Fig. 2 Epicardial electric equipotential lines in millivolts automatically drawn on a projected anterior and posterior heart surface in a dog experiments using a Tektronix 4010 graphical equipment

tory. Through the intervention of Prof. Casella *(University of Pavia) we were introduced to Prof. Magenes, who gave a decisive impulse to our research. Professor* Magenes *understood that it was important to move from the body-surface information to the knowledge of the intracardiac electrical events that trigger the mechanical heart beat. The theoretical feasibility of this transition, predicted by Prof.* Pilkington *(Duke University, 1968), see [12], was confirmed by Prof. Magenes's group* (Colli Franzone et al.) *who published the first cardiac maps obtained with the "inverse" procedure. The cooperation between our group of physiologists and the bio-mathematicians in Pavia continued for about 40 years, and resulted in a number of publications. Prof.* Magenes's group *is still active and productive in the field."*

Let me now state the **Forward or Direct Problem of Electrocardiography**: given the potential at the surface Γ_h bounding the heart and close to it, compute the potential map on the body surface Γ_t.

For simplicity, we assume that the body volume Ω outside the heart H is an isotropic and homogeneous conductor, see Fig. 3. Since the current sources lie only in the heart muscle and the body surface Γ_t in contact with air is insulated, then, at any time instant t of the heart beat, the potential field $U(\mathbf{x}, t)$, under the quasi-static

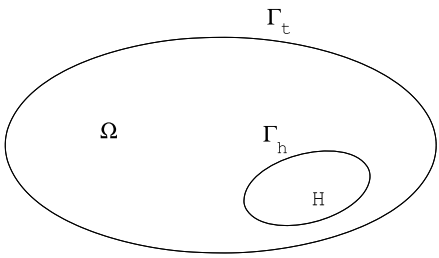


Fig. 3 Sketch of the geometrical domains

assumption, satisfies the following elliptic problem with mixed boundary conditions:

$$\begin{cases} -\Delta U(\mathbf{x},t) = 0 & \text{in } \Omega, \\ U(\mathbf{x},t) = v(\mathbf{x},t) & \text{on } \Gamma_h, \\ \mathbf{n}^T \nabla U(\mathbf{x},t) = 0 & \text{on } \Gamma_t. \end{cases} \quad \Rightarrow \quad z = U_{|\Gamma_t}$$

For any time instant t and $v = U(\mathbf{x},t)$ on Γ_h, let z be the trace on the thorax surface Γ_t of the solution of the previous mixed problem.

We then have a **Transfer Operator** relating the heart to the body surface potentials:

$$\mathscr{T} : v = U_{|\Gamma_h} \to z = U_{|\Gamma_t}.$$

The **Inverse Potential Problem of Electrocardiography** consists in finding the potential distribution on a surface near and surrounding the heart surface, corresponding to a given body surface potential map z on $\Sigma \subset \Gamma_t$, where Σ denotes the part of the torso surface where the potential $U(\mathbf{x},t)$ is measured. This inverse problem consists in solving the following Cauchy problem for the elliptic operator:

$$\begin{cases} -\Delta U(\mathbf{x},t) = 0, & \text{in } \Omega, \\ U(\mathbf{x},t) = z(\mathbf{x},t) & \text{on } \Sigma, \\ \mathbf{n}^T \nabla U(\mathbf{x},t) = 0 & \text{on } \Gamma_t. \end{cases} \quad \Rightarrow \quad v = U_{|\Gamma_h}$$

For t fixed, we have a corresponding potential $v = U(\mathbf{x},t)$ on a surface Γ_h encircling the heart and very close to it, which I call *epicardial* potential. Since Γ_h is close to the heart surface, the potential $v = U_{|\Gamma_h}$ for a sequence of time instants, yields more detailed information on the bioelectric activity of the heart than $z = U_{|\Gamma_t}$. Unfortunately, the problem is known to be strongly ill-conditioned, i.e. for compatible data the inverse operator of $\mathscr{T}$ is unbounded in the usual Sobolev spaces.

The inverse problem can be solved by means of suitable regularization techniques, but a first step of any stabilization technique consists in building an approximation of the transfer operator $\mathscr{T}$. Two approaches are available:

- one based on the **FE** method for solving the 3D direct problem;
- another based on **Boundary Element** methods applied to a surface integral representation of the direct problem, usually adopted by the biomedical engineering community, see e.g. [12].

For a given time instant t, the vector $\mathbf{z}(t) = \mathbf{U}_T(t)$ represents the potential measured on a set of M points lying on the thorax surface and $\mathbf{v}(t) = \mathbf{U}_H(t)$ represents the corresponding potential on N points located on the heart surface Γ_h.

The *transfer matrix* $\mathbf{T} : \mathbf{U}_H(t) \to \mathbf{U}_T(t)$ approximating the operator $\mathscr{T}$ can be built either:

- by solving the following **N** three-dimensional elliptic mixed problems assigning *elementary potentials* $v_i(\mathbf{x})$ on the cardiac surface $U_{\Gamma_h}(\mathbf{x}, t) = \sum_{i=1}^{N} U^i_{\Gamma_h}(t) v_i(x)$:

$$-\Delta T_i(\mathbf{x}) = 0, \quad T_i = v_i \quad \text{on } \Gamma_h, \quad \frac{\partial T_i}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma_t, \quad \mathbf{T} = \{T_i(x_j)\},$$

or

- by solving a surface potential formulation of the forward problem consisting of a system of two surface integral equations on the heart and body surfaces. For instance, using the double Green formula, the following integral representation of the solution of the Direct problem holds:

$$U(\mathbf{x}, t; \psi, \phi) = -\int_{\Gamma_h} v(\xi, t) \frac{\partial s(\mathbf{x}, \xi)}{\partial \mathbf{n}_\xi} d\sigma_\xi + \int_{\Gamma_h} \Psi(\xi, t) s(\mathbf{x}, \xi) d\sigma_\xi - \int_{\Gamma_t} \phi(\xi, t) \frac{\partial s(\mathbf{x}, \xi)}{\partial \mathbf{n}_\xi} d\sigma_\xi,$$

with $v := U_{|\Gamma_h}$, $\Phi := U_{|\Gamma_t}$, $\psi := \frac{\partial U}{\partial \mathbf{n}}{}_{|\Gamma_h}$ and $s(\mathbf{x}, \xi) = \frac{1}{\|\mathbf{x} - \xi\|}$. Applying the limit formulae for the traces on Γ_h and Γ_t, we obtain the following surface integral system:

$$\begin{aligned}
&\int_{\Gamma_H} \Psi(\mathbf{y}, t) s(\mathbf{x}, \xi) d\sigma_\xi - \int_{\Gamma_T} \phi(\xi, t) \frac{\partial s(\mathbf{x}, \xi)}{\partial \mathbf{n}_\xi} d\sigma_\xi \\
&\quad = \frac{v(\mathbf{x}, t)}{2} + \int_{\Gamma_H} v(\xi, t) \frac{\partial s(\mathbf{x}, \xi)}{\partial \mathbf{n}_y} d\sigma_y, \quad \text{in } H^{1/2}(\Gamma_h), \\
&\frac{\phi(\mathbf{x}, t)}{2} + \int_{\Gamma_T} \phi(\xi, t) \frac{\partial s(\mathbf{x}, \xi)}{\partial \mathbf{n}_\xi} d\sigma_\xi - \int_{\Gamma_H} \Psi(\xi, t) s(\mathbf{x}, \xi) d\sigma_\xi \\
&\quad = -\int_{\Gamma_H} v(\xi, t) \frac{\partial s(\mathbf{x}, \xi)}{\partial \mathbf{n}_\xi} d\sigma_\xi, \quad \text{in } H^{1/2}(\Gamma_t).
\end{aligned}$$

In the years 1976–1978, *Nedelec* and collaborators, see e.g. [13], developed *variational formulations* of exterior and interior harmonic problems with Dirichlet or Neumann boundary conditions, in two- or three-dimensional domains based on integral representations of the solution, different from the *classical* simple or double layer potentials. Following this approach, *Magenes* and I considered various variational formulations of the direct problem with their boundary element approximations and the results were reported in [1]. For instance for the previous integral system in the unknowns ($\phi = \frac{\partial U}{\partial \mathbf{n}}{}_{|\Gamma_h}$, $\Psi = U_{|\Gamma_t}$) in $H^{-1/2}(\Gamma_h) \times H^{1/2}(\Gamma_t)$, a-priori error estimates for boundary element approximations were established.

The results of these integral approaches were collected in [1], in the thesis of *Stefania Tentoni* [16], and in a work presented at the Italy-France-Russia workshop in Moscow, published in Cyrillic (see [7]). The **transfer matrix T**, approximating the transfer operator $\mathscr{T}$, was derived by block solving the linear system associated with the discretization of the surface integral system.

Both approaches **3D-FE** and **2D-BE** yielded transfer matrices exhibiting a ratio between the greatest and smallest non zero singular values of the order of 10^7, indicating a strong ill-conditioning of the inverse problem.

A new **Goal-oriented Project** of the C.N.R., called *Tecnologie Biomediche e Sanitarie*, i.e. *Biomedical and Health Technologies*, lasted from 1982 until 1987, still coordinated by Prof. *Luigi Donato*, Director of the CNR *Institute of Clinical Physiology* in Pisa. Once more, *Enrico Magenes* was appointed member of the Scientific Board of this project. A new research unit of the *Istituto di Analisi Numerica* (**IAN**, Institute of Numerical Analysis), the new name for the **LAN** laboratory, composed by *Luciano Guerri, Stefania Tentoni, Carla Viganotti and myself*, submitted a project included in the subproject on *Elettromappe Cardiache*, i.e. *Cardiac Body Surface Maps*. During this project, the inverse problem in term of epicardial potential was carried out by applying the inverse procedure to new experimental data on animals collected by Prof. **Bruno Taccardi** at the University of Parma, see Fig. 4 and subsequently also to human body surface maps related both to normal subjects and to subjects with the Wolff-Parkinson-White syndrome, see Figs. 5–6. These figures display examples of the epicardial potential distributions inversely computed from body surface maps. Equipotential lines are trace using a Tektronix 4010 graphical computer; later on, the **IAN** acquired the new 3D graphical workstation Tektronix 4330 released in 1988, see Fig. 7.

3 Cardiac Excitation Sources Models and Inverse Excitation Wavefront Problems

Another research topic in my collaborations with *Magenes*, was related to the *macroscopic representation of the cardiac electric sources* during the excitation phase of the ventricular myocardium. During this phase, a moving layer with thickness less than 1 mm, sweeps all the ventricular wall. Disregarding the layer thickness, we have an excitation wavefront surface separating the resting tissue from the activated one. The *classical* model of the excitation sources is the *uniform dipole layer source* distributed on the excitation wavefront. In 1976, experimental measurements [6] disproved the *classical* model, showing the important role played by anisotropic cardiac fibers of the ventricular wall in determining the features of the potential patterns. Thus, we began to develop models of the anisotropic cardiac sources. In [2], predictions of the *oblique dipole layer* model were compared with experimental data and a first mathematical investigation of this anisotropic source model was performed in [3].

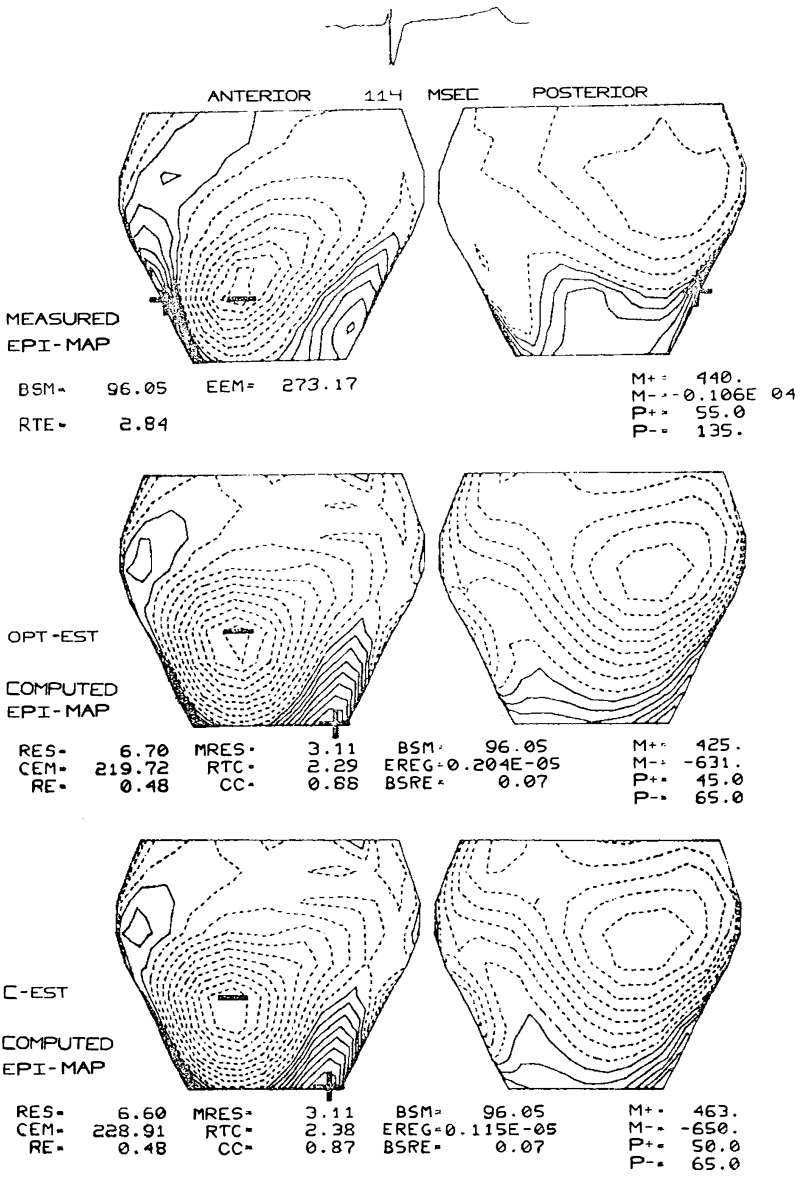


Fig. 4 Inverse problem: comparison between measured and inversely computed *epicardial* potential maps related to a time instant of the QRS complex

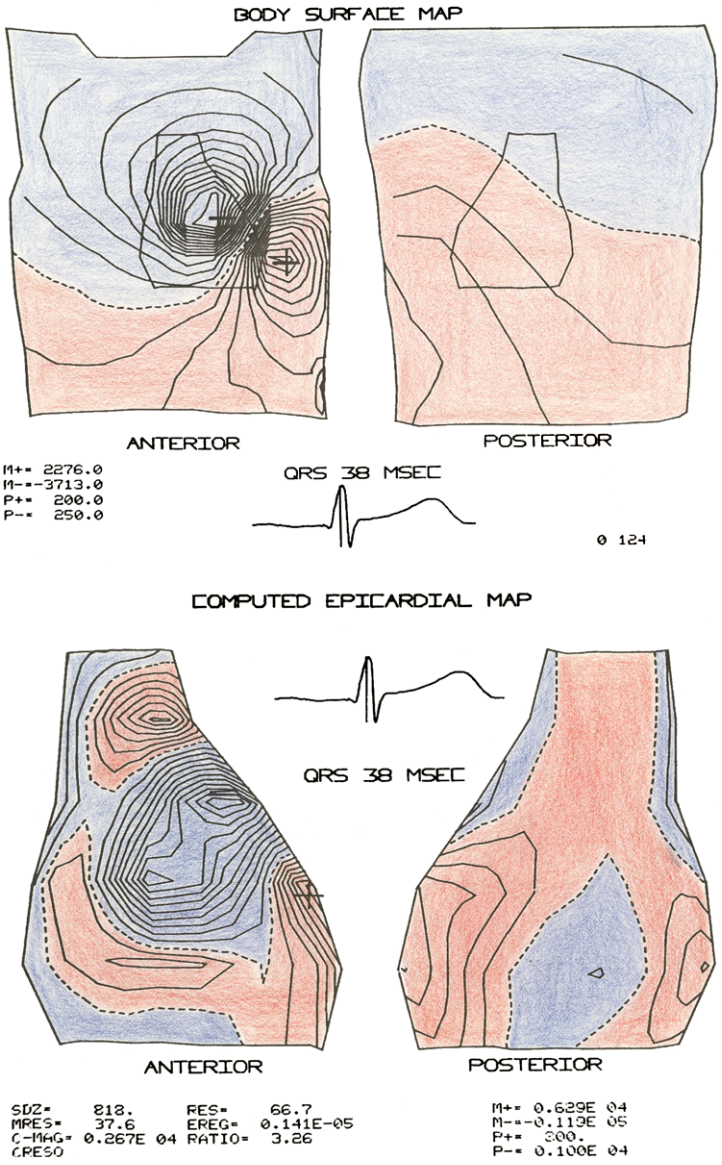


Fig. 5 Body and *epicardial* heart surface mesh. Measured body surface map and the inversely computed *epicardial* potential map for a time instant in the QRS complex of a normal human subject

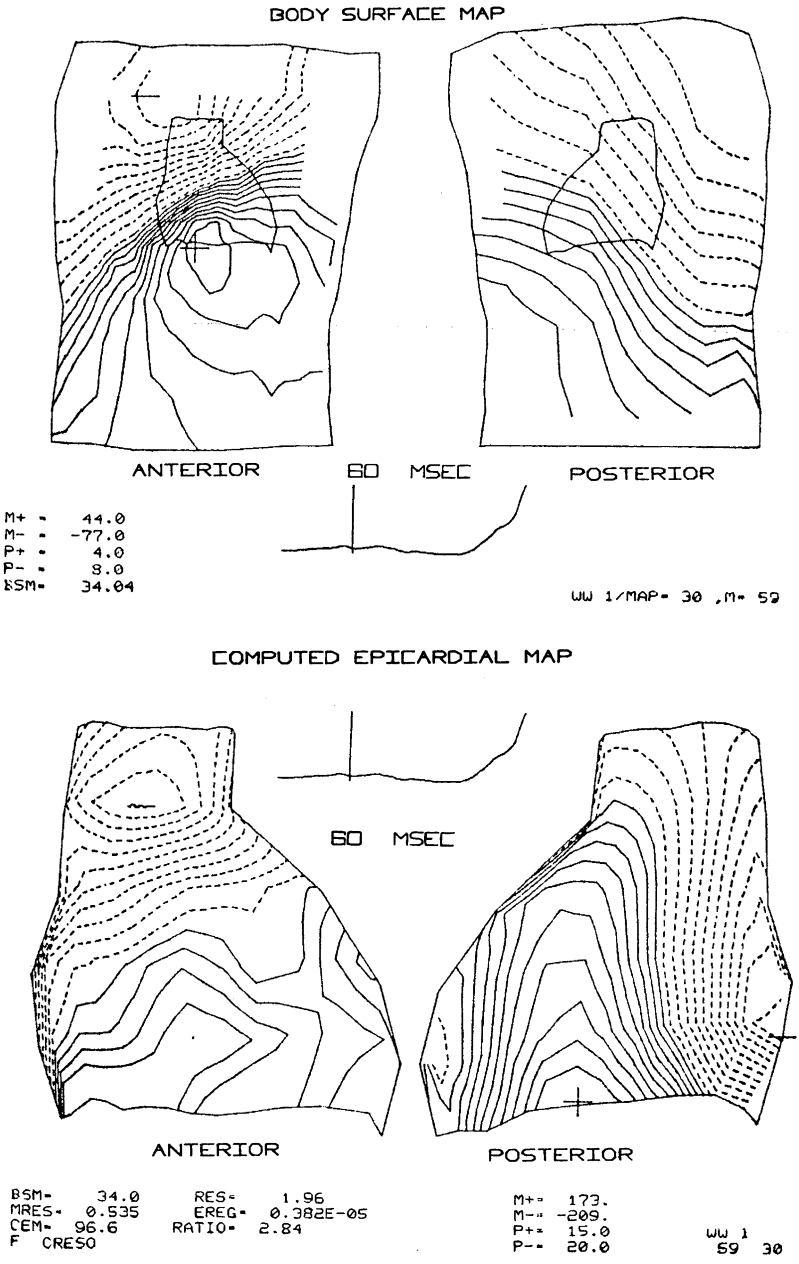


Fig. 6 Body and *epicardial* heart surface mesh. Measured body surface map and the inversely computed *epicardial* potential map for a time instant in the QRS complex of a human subject with WPW syndrome

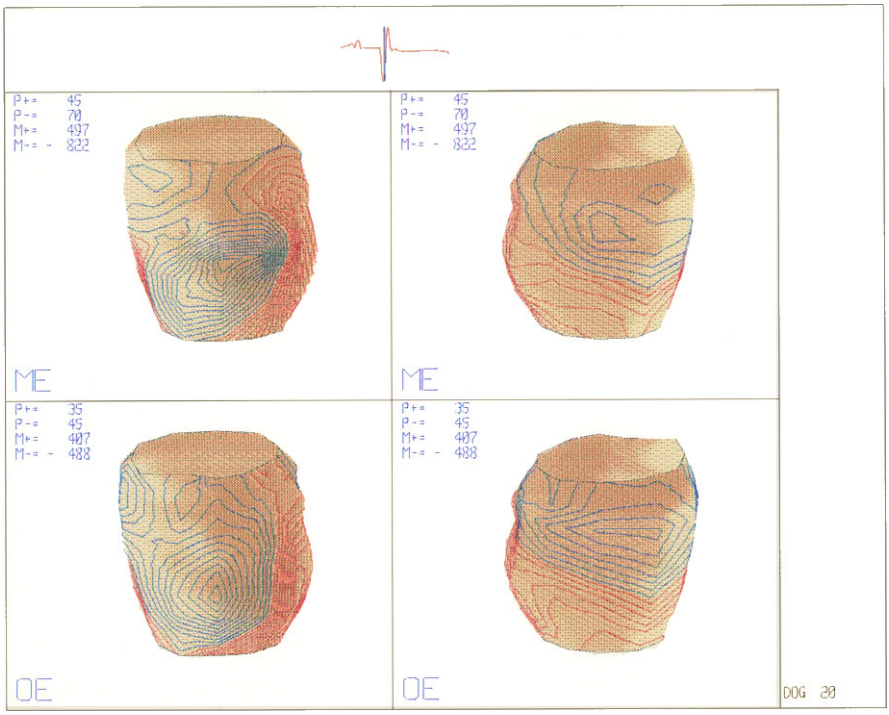


Fig. 7 Inverse problem: comparison between measured and inversely computed *epicardial* potential maps displayed on a Tektronix 4330 3D graphical terminal

Given the excitation wavefront S at the time instant t, assumed embedded in an infinite homogeneous Ohmic conductor medium with conductivity σ, and given the intracellular tensor $M_i(\mathbf{x})$ related to the fiber structure of the cardiac tissue, the cardiac electric field at point $\mathbf{x}$ can be derived from the following integral representation:

$$U(\mathbf{x},t) = \frac{v_J}{4\pi\sigma}\int_S \mathbf{n}_\xi^T M_i(\xi)\nabla_\xi s(\mathbf{x},\xi)d\xi, \quad \mathbf{x}\in R^3 - S, \quad \text{with } s(\mathbf{x},\xi) = \frac{1}{\|\mathbf{x}-\xi\|},$$

where $\mathbf{n}$ denotes the unit vector normal to S pointing toward the resting tissue, v_J is the jump of the transmembrane potential from the excited tissue to the resting one, and M_i is the intracellular conductivity tensor.

The presence of the conductivity tensor M_i takes into account the anisotropic structure of the ventricular tissue due to the fiber structure. Assuming axisymmetric anisotropy, the intracellular tensor is given by $M_i(\mathbf{x}) = \sigma_t^i \mathbf{I} + (\sigma_l^i - \sigma_t^i)\mathbf{a}_l(\mathbf{x})\mathbf{a}_l(\mathbf{x})^T$, where the unit vector $\mathbf{a}_l$ is parallel to the local fiber direction. This yields an *anisotropic cardiac sources model* composed by oblique dipoles, i.e. directed as $M_i\mathbf{n}$ and distributed on the excitation wavefront S.

Let Ω be the body volume with boundary $\Gamma = \partial\Omega$, which is insulated, since it is in contact with the air. Assuming for simplicity that the body conductor is isotropic and homogeneous, the potential field generated by the previous *oblique dipole layer sources* is the solution of the following boundary value problem:

$$\begin{gathered}\Delta U = 0 \quad \text{in } \Omega - S, \qquad \frac{\partial U}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma, \\ [[U]]_S = \alpha, \qquad \left[\left[\frac{\partial U}{\partial \mathbf{n}}\right]\right]_S = \operatorname{div}^S \beta, \\ \text{with } \alpha = \mathbf{n}^T M_i \mathbf{n} \text{ and } \beta = M_i \mathbf{n} - \alpha \mathbf{n}.\end{gathered} \tag{1}$$

In collaboration with *Magenes*, we established in [4] the following first result for the well-posedness of the direct problem (1) in terms of the excitation cardiac sources:

For any open and regular surface S, there exits a unique solution $U(\mathbf{x}, t)$ of (1) *in* $\mathbf{H}^{\frac{1}{2}-\varepsilon}$, $\forall \varepsilon > 0$, *up to an additive constant; moreover, any bounded solution admits a surface integral representation.*

The uniqueness of the inverse problem in term of sources was also investigated in [4] and under some technical geometrical assumptions the following result was obtained:

Given the intracellular tensor M_i related to the fiber structure and a body surface potential map z on $\Sigma \subset \Gamma_t$ generated by the excitation wavefront S, S is the unique surface able to elicit the body surface potential map z.

During last year Spring, I decided to repaint some rooms in my house, and I substituted a bookcase, thus transferring a lot of paper sheets in some boxes. At the end of July 2011, I examined the paper sheets before finally discarding them, and I rediscovered three letters sent to me by *Magenes* in the summer of 1982. During the 80s, I used to spend my holidays at the seashore in Liguria, in the last weeks of July. Indeed, the dates of two letters are July 29 and 30, and the third is August 14. All the letters are related to the initial phase of the work on the *macroscopic models of the cardiac electric sources*.

I begin from the last letter, written in the days before the mid-August holiday, in which the direct and inverse problem in term of the cardiac sources are discussed:

- *Cari* Colli e Guerri, *Pavia 14-8-82*
  spero siate tornati riposati e "pimpanti" dalle vacanze.
  Io ho preso un raffreddore molto forte appena arrivato e per una settimana sono rimasto istupidito e non ho certo riposato. Comunque in attesa di giorni migliori, nei pochi momenti di lucidità ho ripensato ai nostri problemi, arrivando a queste conclusioni per ora. (solite notazioni)

Mi sembra utile introdurre una definizione dicendo che dato Ω (torace), una terna $(H, M, \mathscr{F})$ è ammissibile se (mi riferisco al modello semplice):
(**a**) *il problema differenziale (vedi (1)) ammette una ed una sola soluzione in un opportuno senso*,
(**b**) *il problema inverso ammette unicità, cioè se u_1 e u_2 sono le soluzioni di (a) relative ad S_1 e $S_2 \in \mathscr{F}$ allora*:

$$u_1(\mathbf{x}) = u_2(\mathbf{x}), \quad \mathbf{x} \in \Gamma \quad \rightarrow \quad S_1 = S_2.$$

Per quanto riguarda (**a**) *non abbiamo più problemi nel senso che* (**a**) *vale in ipotesi molto larghe su $(H, M, \mathscr{F})$.*
Per quanto riguarda (**b**) *sono arrivato a concludere così un Teorema possibile e ormai quasi dimostrato.*
Teorema: *$(H, M, \mathscr{F})$ è ammissibile se ... 7 dense pages follow.*

- *Dear* Colli and Guerri, *Pavia 14-8-82*
I hope you came back from the holidays rested and jaunty.
I caught a strong cold upon my arrival and for a week I remained dazed and certainly I did not rest. However, waiting for better days, in the few lucid moments, I have looked back to our problems, reaching for the moment the following conclusions (in the usual notations).
I think it would be useful to introduce the following definition, stating that given Ω (the torax), a triple $(H, M, \mathscr{F})$ is admissible if (I refer here to the simple model):
(**a**) *the differential problem (see (1)) admits one and only one solution in a proper sense*,
(**b**) *the inverse problem admits uniqueness, i.e. if u_1 and u_2 are solutions of* (**a**) *associated with the surfaces S_1 and $S_2 \in \mathscr{F}$ then*:

$$U_1(\mathbf{x}) = U_2(\mathbf{x}), \quad \mathbf{x} \in \Gamma \quad \rightarrow \quad S_1 = S_2.$$

Concerning (**a**), *we no longer have problems, meaning that* (**a**) *holds under weak hypotheses on $(H, M, \mathscr{F})$.*
Concerning (**b**), *I have reached a formulation of the following possible Theorem, by now almost proved.*
Theorem: *$(H, M, \mathscr{F})$ is an admissible triple if ... 7 dense pages follow.*

The first letter begins with

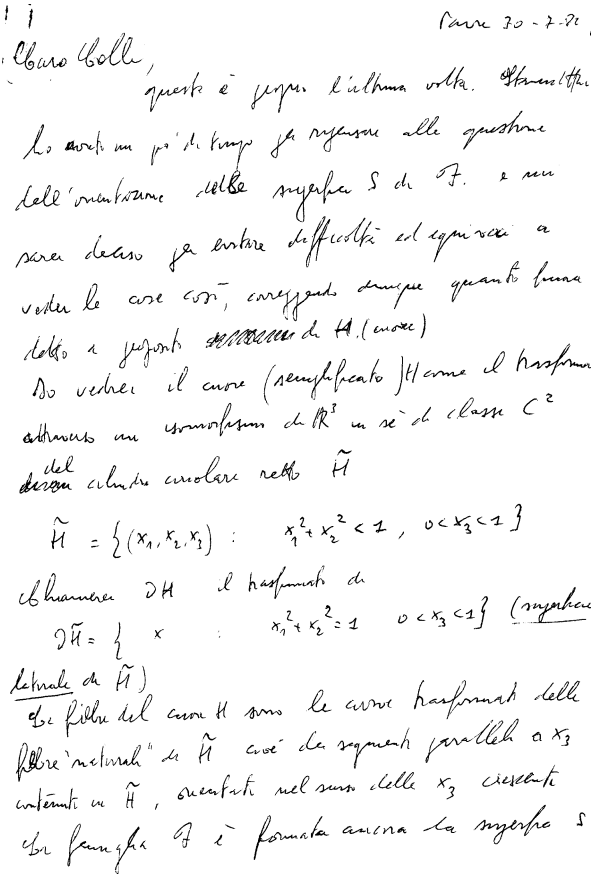
- *Caro Colli*, *Pavia 29-7-82*
prima di partire per la montagna ti invio quel poco che ho potuto fare, per lo più con il **Guerri** *nel giorno in cui è passato da Pavia tra (beato lui che può!) il mare e la montagna.*
Anzitutto con **Guerri** *abbiamo chiarito le ipotesi su* **M** *e sulla famiglia di superficie S ammissibili nel caso del modello semplificato ...*

A translation attempt is:

- *Dear Colli,* *Pavia 29-7-82*
  before leaving for the mountains, I send you the little material that I have produced, mostly with Guerri during the day when he stopped by in Pavia, while traveling from the seaside to the mountains (lucky him, as he can do that!).
  First of all, with **Guerri** *we clarified the hypotheses on* **M** *and on the family of admissible surfaces S for the simplified model . . .*

and the following letter:

- *Caro Colli,* *Pavia 30-7-82*
  questa è proprio l'ultima volta. Stamattina ho avuto un pò di tempo per ripensare alla questione dell'orientazione delle superficie S di $\mathcal{F}$ e mi sarei deciso per evitare difficoltà ed equivoci a vedere le cose così, correggendo quanto finora detto a proposito di H (cuore).
  Io vedrei il cuore (semplificato) H come il trasformato attraverso un isomorfismo . . . di un cilindro circolare retto . . .



Pavia 30-7-82

Caro Colli,

questa è proprio l'ultima volta. Stamattina ho avuto un pò di tempo per ripensare alla questione dell'orientazione della superficie S di $\mathcal{F}$ e mi sarei deciso per evitare difficoltà ed equivoci a vedere le cose così, correggendo dunque quanto finora detto a proposito di H (cuore)

Io vedrei il cuore (semplificato) H come il trasformato attraverso un isomorfismo di $\mathbb{R}^3$ in sè di classe C^2 del cilindro circolare retto $\tilde{H}$

$$\tilde{H} = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 < 1, \ 0 < x_3 < 1\}$$

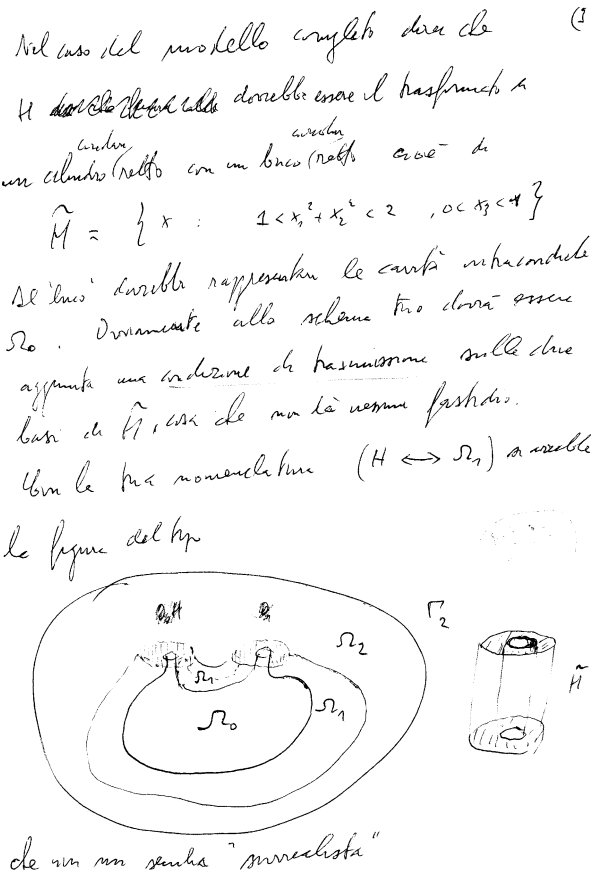
chiamerei ∂H il trasformato di

$$\partial\tilde{H} = \{x : x_1^2 + x_2^2 = 1 \ \ 0 < x_3 < 1\}$$ (superficie laterale di $\tilde{H}$)

Le fibre del cuore H sono le curve trasformate delle fibre "naturali" di $\tilde{H}$ cioè da segmenti paralleli a x_3 contenuti in $\tilde{H}$, orientate nel senso delle x_3 crescenti

La famiglia $\mathcal{F}$ è formata ancora da superfici S

Then Magenes goes on writing:
Nel caso del modello completo direi che H dovrebbe essere il trasformato di un cilindro circolare retto con un buco circolare retto ...
il buco dovrebbe rappresentare le cavità intracardiache Ω_o. Ovviamente allo schema tuo dovrà essere aggiunta una condizione di trasmissione sulle due basi di $\bar{H}$, cosa che non dà fastidio.
Con la tua nomenclatura ($H \leftrightarrow \Omega_1$) si avrebbe una figura del tipo:
che non mi sembra "surrealista".



A translation attempt is:

- *Dear Colli,* *Pavia 30-7-82*
this is really the last time. This morning I had some time to think again about to the issue concerning the orientation of the surface S of the family $\mathcal{F}$ and, in

order to avoid difficulties and misunderstandings, I decided to view things in the following way, correcting what we have introduced so far about H (heart).
I would suggest to see the heart H (simplified) as the image through an isomorphism ... of a right circular cylinder ...
Then *Magenes* goes on writing:
For the full model, I think that H should be a right circular cylinder with a right circular hole ...
the hole could represent the intracardiac cavities Ω_o. Obviously, a transmission condition on the two basis of $\bar{H}$ should be added to your scheme, which does not create any trouble.
With your notation ($H\ leftrightarrow\ \Omega_1$), we would have a picture of the following type:
which does not seem "*surrealist*" *to me.*

The solution of the inverse problem was in part based on the study of *regular oblique derivative elliptic problems*. This research topic had been tackled by *Magenes* in the Fifties when he was a professor at the University of Modena and subsequently at the University of Genoa in the papers [8–10].

The uniqueness results for the inverse problem, developed in the first paper [3] on the *Oblique double layer Sources*, were indeed based on uniqueness results concerning *regular oblique derivative elliptic problems*.

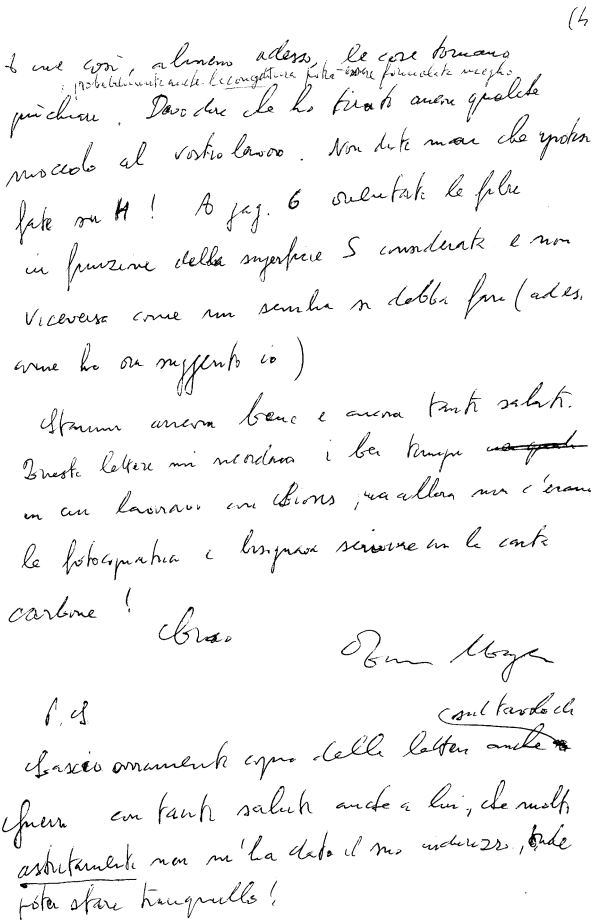
Moreover, a deeper analysis of the uniqueness of the inverse problem in two dimensions was developed by *Magenes* in [11].

A final excerpt from the end of the letter of August 14:

- *... A me così, almeno adesso le cose tornano più chiare e probabilmente anche la congettura potrà essere formulata meglio. Devo dire che ho tirato ancora qualche moccolo al vostro lavoro. Non dite mai che ipotesi fate su H! A pag. 6 orientate le fibre in funzione della superficie considerata e non viceversa come mi sembra si debba fare (ad es. come ora ho suggerito io).*
Stammi ancora bene e ancora tanti saluti.
Queste lettere mi ricordano i bei tempi in cui lavoravo con **Lions**, *ma allora non c'erano le fotocopiatrici e bisognava scrivere con la carta carbone!*

Ciao **Enrico Magenes**

P.S. *Lascio ovviamente copia della lettera anche sul tavolo di* **Guerri**, *con tanti saluti anche a lui, che molto astutamente non mi ha dato il suo indirizzo, onde poter stare tranquillo!*



A translation attempt is:

- *. . . From my point of view, now things are clearer and probably also the conjecture could be better formulated. I must say that I still swore your work. You do not state clearly the hypothesis on H! On page 6, you assign the fiber orientation as a function of the chosen surface and not vice-versa, as I think it should be (for example as I have suggested).*
  Take care and again best regards.
  These letters remind me of the good old times, when I worked with **Lions**, *but back then we did not have xerox machines and we had to write duplicates with carbon copy!*

  ciao *Enrico Magenes*

  P.S. *I leave a copy of the letter also on Guerri's desk, with many regards also for him (who, very cleverly, did not leave me his address so as not to be disturbed!).*

Dear Enrico, thanks for all.

References

1. Colli Franzone, P., Magenes, E.: On the inverse potential problem of electrocardiology. Calcolo **XVI**(IV), 459–538 (1979)
2. Colli Franzone, P., Guerri, L., Viganotti, C., Macchi, E., Baruffi, S., Spaggiari, S., Taccardi, B.: Potential fields generated by oblique dipole layers modeling excitation wavefronts in the anisotropic myocardium. Comparison with potential fields elicited by paced dog heart in a volume conductor. Circ. Res. **51**, 330–346 (1982)
3. Colli Franzone, P., Guerri, L., Viganotti, C.: Oblique dipole layer potentials applied to electrocardiology. J. Math. Biol. **17**, 93–124 (1983)
4. Colli Franzone, P., Guerri, L., Magenes, E.: Oblique dipole layer potential for the direct and inverse problems of electrocardiology. Math. Biosci. **68**, 23–55 (1984)
5. Colli Franzone, P., Guerri, L., Viganotti, C., Taccardi, B.: Finite element approximation of regularized solution of the inverse potential problem of electrocardiography and applications to experimental data. Calcolo **XXII**(I), 91–186 (1985)
6. Corbin, L.V. II, Scher, A.M.: The canine heart as an electrocardiographic generator: dependence on cardiac cell orientation. Circ. Res. **41**, 58–67 (1977)
7. Guerri, L., Magenes, E.: On the inverse problem of electrocardiology. In: Alekseev, A.S. (ed.) Current Problems of Numerical and Applied Mathematics. Collection of Articles (Aktual'nye Problemy Vychislitel'noj i Prikladnoj Matematiki. Sbornik Statej), pp. 59–72. Izdatel'stvo "Nauka" Sibirskoe Otdelenie, Novosibirsk (1983) (in Russian). [B] Akademiya Nauk SSSR, Sibirskoe Otdelenie. Vychislitel'nyj Tsentr., 208 p. R. 2.00
8. Magenes, E.: Sui problemi di derivata obliqua regolare per le equazioni lineari del secondo ordine di tipo ellittico. Ann. Mat. Pura Appl. (4) **XL**, 143–160 (1955)
9. Magenes, E.: Su alcune recenti impostazioni dei problemi al contorno, in particolare misti, per le equazioni lineari ellittiche del secondo ordine. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (3), **X**(I–II), 75–84 (1956)
10. Magenes, E.: Il problema della derivata obliqua regolare per le equazioni lineari ellittico-paraboliche del secondo ordine in m variabili. Rend. Mat. **16**(3–4), 363–414 (1957)
11. Magenes, E.: Su un problema inverso di teoria dei potenziali logaritmici. Calcolo **XXII**(I), 31–46 (1985)
12. Martin, R.O., Pilkington, T.C.: Statistically constrained inverse electrocardiology. IEEE Trans. Biomed. Eng. **22**(6), 487–492 (1975)
13. Nedelec, C.: Curved finite element methods for the solution of singular integral equations on surfaces in R^3. Comput. Methods Appl. Mech. Eng. **8**, 61–80 (1976)
14. Rudy, Y., Oster, H.S.: The Electrocardiographic Inverse Problem. CRC Press, Boca Raton (1993)
15. Taccardi, B., Macchi, E., Lux, R.L., Ershler, P.R., Spaggiari, S., Baruffi, S., Vyhmeister, Y.: Effect of myocardial fiber direction on epicardial potentials. Circulation **90**, 3076–3090 (1994)
16. Tentoni, S.: Analisi numerica di problemi ai limiti ellittici, connessi con l'Elettrocardiologia, mediante potenziali di superficie. Master Thesis, Pavia (1980)

Stefan Problems and Numerical Analysis

Claudio Verdi

Abstract We outline the main contributions of Prof. Enrico Magenes to the analysis and numerical approximation of mathematical models of phase transition processes. Starting from the 1980's, a semigroup approach to Stefan problems, optimal rates of convergence for the nonlinear Chernoff formula, regularity properties of solutions, theoretical and numerical aspects of Stefan models in a concentrated capacity, were investigated by Enrico Magenes. His expertise was fundamental for developing numerical analysis of evolutionary free boundary problems and applications in a modern framework.

1 Stefan Problems and Semigroups: Analysis and Numerical Approximation

Phase transitions occur in many relevant processes in natural sciences and industrial applications. The basic Stefan model represents phase transitions in a rather simplified way by coupling heat diffusion and exchange of latent heat between phases. It has been extensively studied in the last 60 years: the existing literature includes tenths of thousand of papers and a number of meetings has been devoted to this model and its extensions. Following the *International Seminar on Free Boundary Problems* held in Pavia in September–October 1979 [33], a regular series of *International Symposium on Free Boundary Problems: Theory and Applications* took place in: Montecatini, 1981 [23]; Maubuisson, 1983 [5]; Irsee, 1987 [27]; Montreal, 1990 [11]; Toledo, 1993 [17]; Zakopane, 1995 [58]; Hiraklion, 1997 [1]; Chiba, 1999 [31]; Trento, 2002 [13]; Coimbra, 2005 [25]; etc. A series of conferences focusing on numerical methods and applications started with the *Workshop on Generalized Stefan Problems: Analysis and Numerical Methods* held in Pavia in 1995 and took place in: Freiburg, 1995; Lamoura, 1996; Faro, 1996; Berlin, 1996; Ittingen, 1997; Madeira, 1998; Braga, 1998; Hiraklion, 1999; etc.

C. Verdi (✉)
Dipartimento di Matematica "F. Enriques", Università di Milano, via Saldini 50, 20133 Milan, Italy
e-mail: claudio.verdi@unimi.it

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*, Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_5,

Existence, uniqueness, and regularity properties of the solution of Stefan problems were obtained by L. Rubinstein [79], S. Kamin [30], O. Oleinik [75], A. Friedman [26], D. Kinderlehrer and L. Niremberg [32], L.A. Caffarelli [8] and L.C. Evans [10], E. Di Benedetto [16], A.M. Meirmanov [54], A. Damlamian [15], A. Visintin [89], M. Niezgodka and I. Pawlow [57], A. Fasano and M. Primicerio [20–22], and many others (we refer to the monographs [55, 80, 91–93]; a huge bibliography can be found in [83]). Numerical analysis and applications were developed by G. Meyer [56], J. Nitsche [59], J.F. Ciavaldini [12], J.W. Jerome and M.E. Rose [29], C.M. Elliott [19], R.H. Nochetto [63] among others (see the survey [86]).

The interest of Enrico Magenes for the Stefan model and its numerical approximation received an impulse during the *International Seminar on Free Boundary Problems* held in Pavia in 1979 and yielded a series of seminars that he delivered at the *V Seminario di Analisi Funzionale e Applicazioni* held in Catania on September 17–24, 1981 (these lecture notes were published in [34], an interesting overview of the state of the art for the multidimensional two-phase Stefan problem). Enrico Magenes worked on this subject for twenty years, with many original contributions but, more relevantly, by stimulating his students and collaborators with continuous discussions and suggestions.

The enthalpy formulation, a fixed domain formulation were the interface or free boundary can be recovered *a posteriori* as level-set of temperature variable, reads in weak or variational form as follows:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta\beta(u) = 0 & \text{in } \Omega \times (0,T), \\ \beta(u) = 0 & \text{on } \partial\Omega \times (0,T), \\ u(\cdot,0) = u_0(\cdot) & \text{in } \Omega, \end{cases} \tag{1}$$

where $\Omega \subset \mathbf{R}^d$ and $\beta(s) = (s-1)^+ - s^-$ is the constitutive relation between enthalpy u and temperature $\theta = \beta(u)$. Existence and uniqueness of the solution can be proved in suitable functional spaces. The interest of Magenes was focused on the possibility to formulate the Stefan problem (1) as an m-accretive semigroup of contraction in $L^1(\Omega)$ in the sense of M.G. Crandall and T. Ligget [14], and Ph. Bénilan [3], which allows to numerically approximate (1), either by backward finite differences (here τ denotes the time-step)

$$U^n - U^{n-1} - \tau\Delta\beta\big(U^n\big) = 0, \tag{2}$$

or by the nonlinear Chernoff formula, as observed by A.E. Berger, H. Brézis, and J.C.W. Rogers [4]

$$\begin{cases} V^n - \tau\Delta V^n = \beta\big(U^{n-1}\big), \\ U^n = U^{n-1} - \beta\big(U^{n-1}\big) + V^n. \end{cases} \tag{3}$$

In many significant applications, the boundary condition $\beta(u) = 0$ on $\partial\Omega \times (0,T)$ could be replaced by nonlinear flux conditions (e.g. Stefan-Boltzmann law) [88]

$$\frac{\partial\beta(u)}{\partial\nu} + g\big(\beta(u)\big) = 0 \quad \text{on } \partial\Omega \times (0,T).$$

In [48], Magenes *et al.* extended the semigroup approach to the Stefan problem with nonlinear flux conditions, by proving that the operator $A : w \to -\Delta\beta(w)$ with domain $D(A) = \{w \in L^1(\Omega) : \beta(w) \in L^1(\Omega), \Delta\beta(w) \in L^1(\Omega), \frac{\partial\beta(w)}{\partial\nu} + g(\beta(w)) = 0 \text{ on } \partial\Omega\}$ is m-accretive in $L^1(\Omega)$, whence the solution exists and is unique in a suitable weak sense. This results was a theoretical step for justifying the convergence properties of the numerical algorithms studied in [49, 84].

The backward Euler method (2) requires the solution of a nonlinear elliptic PDE at each time-step. Combined with a finite element method for spatial approximation and numerical quadrature, it leads to an effective numerical scheme. Stability and *a priori* error estimates under minimal regularity properties on data have been proved in [65, 85] (see [67, 68] for *a posteriori* error estimates and an adaptive implementation). On the other hand, the nonlinear Chernoff formula (3) requires the solution of a linear elliptic PDE at each time step followed by an algebraic correction to recover discrete enthalpy. It turns out that Chernoff is a stable linearization procedure in the spirit of the Laplace-modified forward Galerkin method for non-degenerate parabolic problems introduced by J. Douglas and T. Dupont [18]. Despite convergence was guaranteed by the theory of nonlinear contraction semigroups in Banach space [6], error estimates remained open until the paper by Magenes *et al.* [50]. The key argument is a combination of the following three features:

- the use of a variational technique first applied by R.H. Nochetto [60, 61];
- the possibility of dealing with minimal regularity properties $u_0 \in L^2(\Omega)$ as shown in [65];
- the relationship between the nonlinear Chernoff formula and the discrete-time scheme studied in [87] for the approximation of Stefan problems with phase relaxation introduced by A. Visintin [90].

By denoting $\theta = \beta(u)$ the temperature and $\chi = u - \theta$ the phase variable, the PDE in (1) reads

$$\frac{\partial(\theta + \chi)}{\partial t} - \Delta\theta = 0, \quad \chi \in H(\theta), \tag{4}$$

where H stands for the Heaviside graph. Being $\varepsilon > 0$ a time-relaxation parameter, the constitutive relation in (4) can be approximated with the phase relaxation equation introduced by A. Visintin [90]

$$\varepsilon \frac{\partial \chi}{\partial t} + H^{-1}(\chi) \ni \theta. \tag{5}$$

After coupling this equation (5) with the PDE in (4), and discretizing in time [87] we get the following algorithm

$$\begin{cases} \left(\Theta^n - \Theta^{n-1}\right) + \left(X^n - X^{n-1}\right) - \tau\,\Delta\Theta^n = 0, \\ \dfrac{\varepsilon}{\tau}\left(X^n - X^{n-1}\right) + H^{-1}\left(X^n\right) \ni \Theta^{n-1}, \end{cases} \tag{6}$$

with stability constraint $\tau \le \varepsilon$. Now it is not difficult to see that this scheme (6) reduces to (3) by choosing $\varepsilon = \tau$ and setting $U^n = \Theta^n + X^n$. With the tools above, Magenes *et al.* [50] completely answered the question of how accurate the nonlinear

Chernoff formula is both for degenerate and non-degenerate parabolic problems. Let $E_\tau = \|\theta - \Theta^n\|_{L^2(\Omega\times(0,T))}$ be the error for temperature in energy norm, then the following *optimal a priori* error estimates were proved for Stefan problems: $E_\tau = O(\tau^{1/4})$ if $u_0 \in L^2(\Omega)$ or $E_\tau = O(\tau^{1/2})$ if $u_0 \in D(A) \cap L^\infty(\Omega)$.

Combined with a finite element method for spatial approximation, the Chernoff formula (3) leads to a very efficient numerical algorithm [64, 66] (see [69, 70] for *a posteriori* error estimates and an adaptive implementation). See also [28] for a refinement of the stabilization parameter in the Chernoff formula.

In a series of papers [35–38] Magenes addressed various extensions of the previous results for the nonlinear Chernoff formula to more general operators and problems, in particular to evolutionary equations on the boundary of a domain. In [51] Magenes *et al.* proved new regularity results in Nikolskiĭ spaces for the multidimensional two-phase Stefan problem with general source terms and, as a consequence, error estimates for enthalpy in energy spaces for the implicit Euler algorithm.

2 Stefan Problems in a Concentrated Capacity

The Stefan problems in a *concentrated capacity* [24] arise in heat diffusion phenomena involving phase changes in two adjoining bodies Ω and Γ, when assuming that the thermal conductivity along the direction normal to the boundary of Ω is much greater than in the others, whence Γ can be considered as the boundary of Ω. The mathematical model describing phase change process in both bodies reads [42]:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta_g \beta(u) = \dfrac{\partial \beta(u)}{\partial \nu} & \text{on } \partial\Omega \times (0,T), \\ u(\cdot,0) = u_0(\cdot) & \text{on } \partial\Omega, \\ \dfrac{\partial v}{\partial t} - \Delta \gamma(v) = 0 & \text{in } \Omega \times (0,T), \\ v(\cdot,0) = v_0(\cdot) & \text{in } \Omega, \\ \gamma(v) = \beta(u) & \text{on } \partial\Omega \times (0,T), \end{cases} \tag{7}$$

where Δ_g is the Laplace-Beltrami operator on $\partial\Omega$ with respect to the Riemannian structure g inferred by the tangential conductivity, β and γ are the constitutive relations between enthalpies u and v and temperature $\theta = \beta(u) = \gamma(v)$. Existence and uniqueness of the solution of (7) was proved in suitable functional spaces [42]. In a series of papers [39–41, 43–46] Magenes addressed the theory of heat conduction with phase change in a concentrated capacity for various operators; see also [82]. The relevance of these models in a number of physical applications motivates their numerical analysis, which was developed in [47]. The implicit Euler scheme reads:

$$\begin{cases} \beta\big(U^n\big) = \gamma\big(V^n\big) \quad \text{on } \partial\Omega, \\ \displaystyle\int_{\partial\Omega} \big(U^n - U^{n-1}\big)\varphi + \tau\big(d\beta\big(U^n\big), d\varphi\big)_g \\ \qquad + \displaystyle\int_\Omega \big(V^n - V^{n-1}\big)\eta + \tau \int_\Omega \nabla\gamma\big(V^n\big)\cdot\nabla\varphi = 0 \quad \forall\varphi. \end{cases} \tag{8}$$

The linear scheme based on the nonlinear Chernoff formula reads:

$$\begin{cases} \Xi^n = \Theta^n \quad \text{on } \partial\Omega, \\ \displaystyle\int_{\partial\Omega} \Xi^n \varphi + \tau \big(d\,\Xi^n, d\varphi\big)_g + \int_\Omega \Theta^n \eta + \tau \int_\Omega \nabla\Theta^n \cdot \nabla\varphi \\ \qquad = \displaystyle\int_{\partial\Omega} \beta\big(U^{n-1}\big)\varphi + \int_\Omega \gamma\big(V^{n-1}\big)\varphi \quad \forall\varphi, \\ U^n = U^{n-1} - \beta\big(U^{n-1}\big) + \Xi^n \quad \text{on } \partial\Omega, \\ V^n = V^{n-1} - \gamma\big(V^{n-1}\big) + \Theta^n \quad \text{in } \Omega. \end{cases} \tag{9}$$

Both algorithms (8) and (9) are well posed. The latter is linear in the unknowns Ξ^n and Θ^n, the nonlinearity reducing to pointwise corrections for U^n and V^n, whence it is expected to be more efficient than (8) from a numerical viewpoint. Stability and error estimates in the natural energy spaces were proved for both schemes in [47].

3 Approximation of Interfaces, Adaptivity and Applications

Optimality of error estimates for the proposed algorithms was one of the main aspects attracting Magenes' interest; in this direction let me mention [74], where R.H. Nochetto *et al.* proved *optimal a posteriori* error estimates for variable time-step discretizations of nonlinear evolution equations; see also [81].

Thanks to stimulating discussions with E. Magenes, C. Baiocchi, F. Brezzi, and L.A. Caffarelli, [2, 7, 9] convergence and accuracy estimates for the approximation of the free boundary of parabolic phase-change problems under suitable condition of non-degeneracy at the interface were proved [62, 78].

The intense interest of Enrico Magenes for the numerical approximation of phase-change models and their applications involved the Istituto di Analisi Numerica of the CNR in Pavia and his collaborators in various national and international projects focused on phase transition problems; I will mention, e.g., the European projects "Phase Transition Problems" (1986–1988), "Mathematical Treatment of Free Boundary Problems" (1993–1996), "Phase Transition and Surface Tension" (1995–1997), "Viscosity Solutions and their Applications" (1998–2000). Within the framework of the national projects of the CNR "Software: Ricerche di Base e Applicazioni; Software Matematico" (1986–1987) and "Sistemi Informatici e Calcolo Parallelo: Calcolo Scientifico per Grandi Sistemi" (1988–1993), a computational code for solving general parabolic free boundary problems with the finite elements algorithms studied above was implemented [77]; some interesting collaborations took place with Himont (Ferrara), a leading factory in polymer production [52], and Istituto Ortopedico Rizzoli (Bologna) [53].

Numerical approximation of geometric motion of interfaces, a first step for the approximation of more complex phase transitions problems including surface tension effects, was addressed by R.H. Nochetto *et al.* [71–73, 76], also thanks to helpful discussions and suggestions of E. De Giorgi, L.A. Caffarelli, A. Visintin, and Magenes himself.

References

1. Athanasopoulos, I., Makrakis, G., Rodrigues, J.-F. (eds.): Free Boundary Problems: Theory and Applications. Pitman Res. Notes Math., vol. 409. Chapman & Hall, Boca Raton (1999)
2. Baiocchi, C.: Estimations d'erreur dans L^∞ pour les inéquations à obstacle. In: Galligani, I., Magenes, E. (eds.) Mathematical Aspects of Finite Element Methods. Lect. Notes Math., vol. 606, pp. 27–34. Springer, Berlin (1977)
3. Bénilan, Ph.: Equation d'évolution dans un espace de Banach quelconque et applications. Pubbl. Math. Orsay **25** (1972)
4. Berger, A.E., Brézis, H., Rogers, J.C.W.: A numerical method for solving the problem $u_t - \Delta f(u) = 0$. RAIRO Modél. Math. Anal. Numér. **13**, 297–312 (1979)
5. Bossavit, A., Damlamian, A., Frémond, M. (eds.): Free Boundary Problems: Applications and Theory, III–IV. Res. Notes Math., vol. 120–121. Pitman, Boston (1985)
6. Brézis, H., Pazy, A.: Convergence and approximation of semigroups of nonlinear operators in Banach spaces. J. Funct. Anal. **9**, 63–74 (1972)
7. Brezzi, F., Caffarelli, L.A.: Convergence of the discrete free boundaries for finite element approximations. RAIRO Modél. Math. Anal. Numér. **4**, 385–395 (1983)
8. Caffarelli, L.A.: The regularity of free boundaries in higher dimensions. Acta Math. **139**, 155–184 (1977)
9. Caffarelli, L.A.: A remark on the Hausdorff measure of a free boundary and the convergence of the coincidence sets. Boll. Unione Mat. Ital., C Anal. Funz. Appl. **18**, 109–113 (1981)
10. Caffarelli, L.A., Evans, L.C.: Continuity of the temperature in the two-phase Stefan problem. Arch. Ration. Mech. Anal. **81**, 199–220 (1983)
11. Chadam, J.M., Rasmussen, H. (eds.): Emerging Applications in Free Boundary Problems, Free Boundary Problems Involving Solids, Free Boundary Problems in Fluid Flow with Applications. Pitman Res. Notes Math., vol. 280–282. Longman, Harlow (1993)
12. Ciavaldini, J.F.: Analyse numérique d'un problème de Stefan à deux phases par une méthode d'éléments finis. SIAM J. Numer. Anal. **12**, 464–487 (1975)
13. Colli, P., Verdi, C., Visintin, A. (eds.): Free Boundary Problems: Theory and Applications. Int. Series Numer. Math., vol. 147. Birkhäuser, Basel (2004)
14. Crandall, M.G., Ligget, T.: Generation of semigroups of nonlinear transformations on general Banach spaces. Am. J. Math. **93**, 265–298 (1971)
15. Damlamian, A.: Some results in the multiphase Stefan problem. Commun. Partial Differ. Equ. **2**, 1017–1044 (1977)
16. Di Benedetto, E.: Regularity properties of the solution of a n-dimensional two-phase Stefan problem. Boll. Unione Mat. Ital. Suppl. **1**, 129–152 (1980)
17. Diaz, J.I., Herrero, M.A., Linan, A., Vasquez, J.L. (eds.): Free Boundary Problems: Theory and Applications. Pitman Res. Notes Math., vol. 323. Longman, Harlow (1995)
18. Douglas, J., Dupont, T.: Alternating-direction Galerkin methods on rectangles. In: Hubbard, B. (ed.) Numerical Solutions of Partial Differential Equations, II, pp. 133–214. Academic Press, New York (1971)
19. Elliott, C.M.: Error analysis of the enthalpy method for the Stefan problem. IMA J. Numer. Anal. **7**, 61–71 (1987)
20. Fasano, A., Primicerio, M.: General free boundary problems for the heat equation, I. J. Math. Anal. Appl. **57**, 694–723 (1977)
21. Fasano, A., Primicerio, M.: General free boundary problems for the heat equation, II. J. Math. Anal. Appl. **58**, 202–231 (1977)
22. Fasano, A., Primicerio, M.: General free boundary problems for the heat equation, III. J. Math. Anal. Appl. **59**, 1–14 (1977).
23. Fasano, A., Primicerio, M. (eds.): Free Boundary Problems: Theory and Applications, I–II. Res. Notes Math., vol. 78–79. Pitman, London (1983)
24. Fasano, A., Primicerio, M., Rubinstein, L.: A model problem for heat conduction with a free boundary in a concentrated capacity. J. Inst. Math. Appl. **26**, 327–347 (1980)

25. Figueiredo, I.N., Rodrigues, J.F., Santos, L. (eds.): Free Boundary Problems: Theory and Applications. Int. Series Numer. Math., vol. 154. Birkhäuser, Basel (2007)
26. Friedman, A.: The Stefan problem in several space variables. Trans. Am. Math. Soc. **133**, 51–87 (1968)
27. Hoffmann, K.-H., Sprekels, J. (eds.): Free Boundary Problems: Theory and Applications, I–II. Pitman Res. Notes Math., vol. 185–186. Longman, Harlow (1990)
28. Jäger, W., Kacur, J.: Solution of porous medium type systems by linear approximation schemes. Numer. Math. **60**, 407–427 (1991)
29. Jerome, J.W., Rose, M.E.: Error estimates for the multidimensional two-phase Stefan problem. Math. Comput. **39**, 377–414 (1982)
30. Kamenomostskaya, S.L.: On Stefan Problem. Nauch. Dokl. Vyss. Shkoly **1**, 60–62 (1958)
31. Kenmochi, N. (ed.): Free Boundary Problems: Theory and Applications, I–II. Gakuto Int. Ser. Math. Sci. Appl., vol. 14. Gakkötosho, Tokyo (2000)
32. Kinderlehrer, D., Nirenberg, L.: Regularity in free boundary problems. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (4) **4**, 373–391 (1977)
33. Magenes, E. (ed.): Free Boundary Problems. Istituto Nazionale di Alta Matematica, Roma (1980)
34. Magenes, E.: Problemi di Stefan bifase in più variabili spaziali. Matematiche **XXXVI**, 65–108 (1981)
35. Magenes, E.: Remarques sur l'approximation des problèmes paraboliques non linéaires. In: Analyse Mathématique et Applications. Contributions en l'honneur de Jacques-Louis Lions, pp. 297–318. Gauthier-Villars, Paris (1988)
36. Magenes, E.: A time-discretization scheme approximating the non-linear evolution equation $u_t + ABu = 0$. In: Colombini, F., et al. (eds.) Partial Differential Equations and the Calculus of Variations. Essays in Honor of Ennio de Giorgi, II, pp. 743–765. Birkhäuser, Boston (1989)
37. Magenes, E.: Numerical approximation of non-linear evolution problems. In: Dautray, R. (ed.) Frontiers in Pure and Applied Mathemathics. A Collection of Papers Dedicated to Jacques-Louis Lions on the Occasion of His Sixties Birthday, pp. 193–207. North-Holland, Amsterdam (1991)
38. Magenes, E.: On the approximation of some non-linear evolution equations. Ric. Mat. **40**(Suppl.), 215–240 (1991)
39. Magenes, E.: On a Stefan problem in a concentrated capacity. In: Ambrosetti, A., Marino, A. (eds.) Nonlinear Analysis: A Tribute in Honour of Giovanni Prodi, pp. 217–229. Scuola Normale Superiore, Pisa (1991)
40. Magenes, E.: Some new results on a Stefan problem in a concentrated capacity. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (9) **3**, 23–34 (1992)
41. Magenes, E.: On a Stefan problem on the boundary of a domain. In: Miranda, M. (ed.) Partial Differential Equations and Relates Subjects. Pitman Res. Notes Math., vol. 269, pp. 209–226. Longman, Harlow (1992)
42. Magenes, E.: The Stefan problem in a concentrated capacity. In: Ricci, P.E. (ed.) Problemi attuali dell'analisi e della fisica matematica: Atti del Simposio internazionale dedicato a Gaetano Fichera nel suo 70° compleanno, Roma, pp. 156–182 (1993)
43. Magenes, E.: Regularity and approximation properties for the solution of a Stefan problem in a concentrated capacity. In: Chicco, M., et al. (eds.) Variational Methods, Nonlinear Analysis and Differential Equations: Proceedings of the International Workshop on the Occasion of the 75th Birthday of J.P. Cecconi, pp. 88–106. E.C.I.G., Genova (1994)
44. Magenes, E.: Stefan problems in a concentrated capacity. In: Alekseev, A.S., Bakhvalov, N.S. (eds.) Advanced Mathematics, Computations and Applications: Proceedings of the International Conference AMCA-95, pp. 82–90. NCC Publisher, Novosibirsk (1995)
45. Magenes, E.: On a Stefan problem in a concentrated capacity. In: Marcellini, P., et al. (eds.) Partial Differential Equations and Applications: Collected Papers in Honor of Carlo Pucci, pp. 237–253. Dekker, New York (1996)
46. Magenes, E.: Stefan problems in a concentrated capacity. Boll. Unione Mat. Ital. Suppl. **8**, 71–81 (1998)

47. Magenes, E., Verdi, C.: Time discretization schemes for the Stefan problem in a concentrated capacity. Meccanica **28**, 121–128 (1993)
48. Magenes, E., Verdi, C., Visintin, A.: Semigroup approach to the Stefan problem with nonlinear flux. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (8) **75**, 24–33 (1983)
49. Magenes, E., Verdi, C.: The semigroup approach to the two-phase Stefan problem with nonlinear flux conditions. In: Bossavit, A., et al. (eds.) Free Boundary Problems: Applications and Theory, III. Res. Notes Math., vol. 120, pp. 28–39. Pitman, Boston (1985)
50. Magenes, E., Nochetto, R.H., Verdi, C.: Energy error estimates for a linear scheme to approximate nonlinear parabolic problems. RAIRO Modél. Math. Anal. Numér. **21**, 655–678 (1987)
51. Magenes, E., Verdi, C., Visintin, A.: Theoretical and numerical results on the two-phase Stefan problem. SIAM J. Numer. Anal. **26**, 1425–1438 (1989)
52. Mazzullo, S., Paolini, M., Verdi, C.: Polymer crystallization and processing: free boundary problems and their numerical approximation. Math. Eng. Ind. **2**, 219–232 (1989)
53. Mazzullo, S., Paolini, M., Verdi, C.: Numerical simulation of thermal bone necrosis during cementation of femoral prostheses. J. Math. Biol. **29**, 475–494 (1991)
54. Meirmanov, A.M.: On the classical solution of the multidimensional Stefan problem for quasilinear parabolic equations. Math. USSR Sb. **40**, 157–178 (1981)
55. Meirmanov, A.M.: The Stefan Problem. de Gruyter, Berlin (1992)
56. Meyer, G.H.: Multidimensional Stefan problem. SIAM J. Numer. Anal. **10**, 512–538 (1973)
57. Niezgodka, M., Pawlow, I.: A generalized Stefan problem in several space variables. Appl. Math. Optim. **9**, 193–224 (1983)
58. Niezgodka, M., Strzelecki, P. (eds.): Free Boundary Problems: Theory and Applications. Pitman Res. Notes Math., vol. 363. Longman, Harlow (1996)
59. Nitsche, J.A.: A finite element method for parabolic free boundary problems. In: Magenes, E. (ed.) Free Boundary Problems, I, pp. 277–318. Istituto Nazionale di Alta Matematica, Roma (1980)
60. Nochetto, R.H.: Error estimates for two-phase Stefan problems in several space variables, I: linear boundary conditions. Calcolo **22**, 457–499 (1985)
61. Nochetto, R.H.: Error estimates for two-phase Stefan problems in several space variables, II: nonlinear flux conditions. Calcolo **22**, 501–534 (1985)
62. Nochetto, R.H.: A note on the approximation of free boundaries by finite element methods. RAIRO Modél. Math. Anal. Numér. **20**, 355–368 (1986)
63. Nochetto, R.H.: Error estimates for multidimensional parabolic problems. Jpn. J. Ind. Appl. Math. **4**, 111–138 (1987)
64. Nochetto, R.H., Verdi, C.: The combined use of a nonlinear Chernoff formula with a regularization procedure for two-phase Stefan problems. Numer. Funct. Anal. Optim. **9**, 1177–1192 (1987–1988)
65. Nochetto, R.H., Verdi, C.: Approximation of degenerate parabolic problems using numerical integration. SIAM J. Numer. Anal. **25**, 784–814 (1988)
66. Nochetto, R.H., Verdi, C.: An efficient linear scheme to approximate parabolic free boundary problems: error estimates and implementation. Math. Comput. **51**, 27–53 (1988)
67. Nochetto, R.H., Paolini, M., Verdi, C.: An adaptive finite element method for two-phase Stefan problems in two space dimensions. Part I: stability and error estimates. Supplement. Math. Comp. **57**, 73–108 (1991), S1–S11
68. Nochetto, R.H., Paolini, M., Verdi, C.: An adaptive finite element method for two-phase Stefan problems in two space dimensions. Part II: implementation and numerical experiments. SIAM J. Sci. Stat. Comput. **12**, 1207–1244 (1991)
69. Nochetto, R.H., Paolini, M., Verdi, C.: A fully discrete adaptive nonlinear Chernoff formula. SIAM J. Numer. Anal. **30**, 991–1014 (1993)
70. Nochetto, R.H., Paolini, M., Verdi, C.: Linearization and adaptivity for FBPs. In: Baiocchi, C., Lions, J.L. (eds.) Boundary Value Problems for Partial Differential Equations and Applications: Dedicated to E. Magenes. Res. Notes Appl. Math., vol. 29, pp. 443–448. Masson, Paris (1993)

71. Nochetto, R.H., Paolini, M., Verdi, C.: Geometric motion of interfaces. In: Baiocchi, C., Lions, J.L. (eds.) Boundary Value Problems for Partial Differential Equations and Applications: Dedicated to E. Magenes. Res. Notes Appl. Math., vol. 29, pp. 403–408. Masson, Paris (1993)
72. Nochetto, R.H., Paolini, M., Verdi, C.: Optimal interface error estimates for the mean curvature flow. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (4) **21**, 193–212 (1994)
73. Nochetto, R.H., Paolini, M., Verdi, C.: A dynamic mesh method for curvature dependent evolving interfaces. J. Comput. Phys. **123**, 296–310 (1996)
74. Nochetto, R.H., Savaré, G., Verdi, C.: A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations. Commun. Pure Appl. Math. **53**, 525–589 (2000)
75. Oleinik, O.A.: A method of solution of the general Stefan problem. Sov. Math. Dokl. **1**, 1350–1354 (1960)
76. Paolini, M., Verdi, C.: Asymptotic and numerical analyses of the mean curvature flow with a space-dependent relaxation parameter. Asymptot. Anal. **5**, 553–574 (1992)
77. Paolini, M., Sacchi, G., Verdi, C.: Finite element approximations of singular parabolic problems. Int. J. Numer. Methods Eng. **26**, 1989–2007 (1988)
78. Pietra, P., Verdi, C.: Convergence of the approximate free boundary for the multidimensional one-phase Stefan problem. Comput. Mech. **1**, 115–125 (1986)
79. Rubinstein, L.: On the determination of the position of the boundary which separates two phases in the one-dimensional problem of Stefan. Dokl. Acad. Nauk USSR **58**, 217–220 (1947)
80. Rubinstein, L.: The Stefan Problem. Translations of Mathematical Monographs, vol. 27. Am. Math. Soc., Providence (1971)
81. Rulla, J., Walkington, N.J.: Optimal rates of convergence for degenerate parabolic problems in two dimensions. SIAM J. Numer. Anal. **33**, 56–67 (1996)
82. Savaré, G., Visintin, A.: Variational convergence of nonlinear diffusion equations: applications to concentrated capacity problems with change of phase. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (9) **8**, 49–89 (1997)
83. Tarzia, D.A.: A Bibliography on Moving-Free Boundary Problems for the Heat-Diffusion Equation. The Stefan and Related Problems. MAT. Serie A: Conferencias. Seminarios y Trabajos de Matemática, vol. 2, 1–297 (2000)
84. Verdi, C.: On the numerical approach to a two-phase Stefan problem with nonlinear flux. Calcolo **22**, 351–381 (1985)
85. Verdi, C.: Optimal error estimates for an approximation of degenerate parabolic problems. Numer. Funct. Anal. Optim. **9**, 657–670 (1987)
86. Verdi, C.: Numerical methods for phase transition problems. Boll. Unione Mat. Ital. Suppl. **8**, 83–108 (1998)
87. Verdi, C., Visintin, A.: Error estimates for a semi-explicit numerical scheme for Stefan-type problems. Numer. Math. **52**, 165–185 (1988)
88. Visintin, A.: Sur le problème de Stefan avec flux non linéaire. Boll. Unione Mat. Ital., C Anal. Funz. Appl. **18**, 63–86 (1981)
89. Visintin, A.: General free boundary evolutions problems in several space dimension. J. Math. Anal. Appl. **95**, 117–143 (1983)
90. Visintin, A.: Stefan problem with phase relaxation. IMA J. Appl. Math. **34**, 225–245 (1985)
91. Visintin, A.: Models of Phase Transitions. Birkhäuser, Boston (1996)
92. Visintin, A.: Introduction to the models of phase transitions. Boll. Unione Mat. Ital. Suppl. **8**, 1–47 (1998)
93. Visintin, A.: Introduction to Stefan-type problems. In: Dafermos, C.M., Pokorny, M. (eds.) Handbook of Differential Equations, Evolutionary Equations, vol. 4, pp. 377–484. North-Holland, Amsterdam (2008)

Enrico Magenes and the Teaching of Mathematics

Mario Ferrari

Abstract It rarely happens that a mathematician of international value, as it was the case for Enrico Magenes, devotes a large part of his time and his interest to Mathematics Education. That is exactly what Enrico Magenes did, and not occasionally, but systematically. First of all, his commitment was at university level with the courses of Complementary Mathematics, specifically devoted to the education of future teachers, assigning graduation thesis on educational topics, and with an active participation in order to give birth to a Mathematics Education Seminar in the School of Sciences. In second place, Magenes spent his time for the Teaching of Mathematics in middle and high schools, working along three different directions: he put forth specific proposals about the teaching programs; during his term as president of the Italian Mathematical Union, he founded the Research Units for Mathematics Education, which initiated the research activity on this topic in Italy; he committed himself directly in the Project "Mathematics as a Discovery," writing the Mathematical Analysis volume with the late Giovanni Prodi.

1 Introduction

Enrico Magenes is internationally known as a researcher in Mathematical Analysis. His contributions were dealt with in the paper published in the December 2010 issue of the Notices of the Italian Mathematical Union, and are presented in the other chapters of the first part of this volume. I would just like to remember, as it seems to me that nobody else has already done it, that towards the end of the Fifties of the last century, Magenes, Pucci, Stampacchia and others founded the CONARM (Comitato Nazionale Ricercatori Matematici, National Committee of Researchers in Mathematics), with the explicit aim of rejuvenating Italian Mathematics, and reinserting it in the framework of international Mathematics, after the twenty years of Fascism. I mention this, because the CONARM took an interest in Mathematics Education in a broad sense, organizing math contests for high school students in a number of Italian cities.

M. Ferrari (✉)
Dipartimento di Matematica "F. Casorati", Università di Pavia, via Ferrata 1, 27100 Pavia, Italy
e-mail: mario.ferrari@unipv.it

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*,
Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_6,

Here my task is to briefly discuss Magenes's contribution to the research on the Teaching of Mathematics.

2 Magenes and the Teaching of Mathematics: A Constant Commitment

From Genoa, where he was full professor of Mathematical Analysis, Magenes arrived in Pavia in 1959, where he had been called as full professor of Complementary Mathematics: in two years this discipline would become a characteristic subject of the course of studies in Mathematics Education. Since he was a novice in the field, in order to define the contents of the two courses he was going to teach, Complementary Mathematics I and II, Magenes followed the examples of well-known experts.

For the course that had more specific geometric contents, he adopted the book "Lessons of Complementary Mathematics: Geometry" by Francesco Cecioni, who had taught the same subject in Pisa; it was a very rich text, largely based on the Hilbert's "Grundlagen der Geometrie," (Foundations of Geometry). Before long, Magenes added a new chapter, inspired by G. Choquet's ideas, based on geometrical transforms.

For the course with arithmetic-algebraic contents, he used two manuals: "Development of mathematical thesis for qualifying exams in middle and high schools" by Rocco Serini from Pavia, and the second volume of "Elementary Mathematics" by Modesto Dedò; moreover, he added a long chapter devoted to Galois theory, for which he used a booklet by Postnikov.

It was Magenes, who initiated in Pavia the course of studies in Mathematics Education in the context of the Degree in Mathematics: he strongly believed in such a program, and worked very hard for it, even after handing on to me the baton of the Complementary Mathematics course in 1970. Here I limit myself to recall the following:

- His full adhesion to the idea of assigning thesis in Mathematics Education. To be honest, at the beginning they were not of experimental kind, that is, with direct practice in schools, but rather based on the critical exam of manuals or of new types of approach to the Teaching of Mathematics in middle and high schools.
- His regular participation to the activities of the Pavia section of the Mathesis Society: these activities were aimed at the constant education of teachers.
- His involvement in the formulation of the so-called "Programs of Frascati," in 1967 for high schools. Even though at that time these programs were not officially accepted by the Italian Ministry of Education, nevertheless they were the starting point for future programs prepared by the same ministry later on. The group from Pavia was made up by Enrico Magenes and Agostino Savaré. De Finetti, who took part to the discussions and to the preparation of the final report with Prodi and others, wrote a long commentary on the journal *Periodico di Matematiche*.

- His fundamental contribution, so that the School of Sciences started the "Seminar in Mathematics Education" in the framework of the same School, and his active participation to the Conference organized in 1975 in Salice Terme with the title "Sciences in Middle and High Schools and the role of University."
- His attendance to the many conferences organized by CIIM (Commissione Italiana per l'Insegnamento della Matematica, Italian Committee for the Teaching of Mathematics), even when he did not have to be there as President of the Italian Mathematical Union, or as one of the keynote speakers.
- Finally, the fact of being from the mid seventies of the last century a member of the "Center for Research in Education *Ugo Morin*," with a regular subscription to its journal "L'Insegnamento della Matematica e delle Scienze Integrate," "The Teaching of Mathematics and of Integrated Sciences."

3 Magenes and the Research in Mathematics Education

Strictly speaking, Magenes never carried out a specific research activity in Mathematics Education, but I think it is right to say that he originated all, or almost all, the Italian research in the Teaching of Mathematics. The reason for this is simple. These scientific activities were carried out in Italy by the Units of Research in Education, at least starting from 1975. Now, the interesting fact is that all these units were born with a large financial endowment from the Italian Research Council, whose mathematical committee was at that time chaired by C. Pucci, during Magenes's tenure as President of the Italian Mathematical Union. The birth was not painless, as it can be clearly seen from the minutes of the Special Assembly of the Italian Mathematical Union, which took place in Alghero on September 27th 1975, and as all the participants to the Tenth Congress of the Union in Cagliari–Alghero probably still remember.

The birth of these units was the natural product of one of the two main guiding lines, which characterized Magenes's Presidency of the Union, that is, as he himself underlined, "Initiatives for problems of schools" (the other one was "Initiatives for the development and strengthening of research") (see [2]).

Answering the critics put forth by G. Stampacchia during the Alghero Assembly, Magenes recalls the Union position, as it was defined during the Assembly that took place on April 4th 1975: "The Assembly ...

- Notices the strong connection between the issue of the teachers' updating and the problem of rejuvenating the teaching of Mathematics both in contents and methodology;
- Recognizes the urgency that CIIM fosters the birth of groups where experts in Mathematics Education closely collaborate with teachers of middle and high schools on different programs; moreover, these groups should work out detailed plans and present reports, in such a way they can be positively used;
- Thinks that this first experience paves the way to a right statement of the problem of the training and updating of teachers."

For some years the units, with their reports and regular updating of the ongoing activities, were the protagonists of the yearly CIIM meetings. The one that took place in 1976 had the very significative title "Experiments in Mathematics Education, with regards to the ongoing discussion about the reform of High School, and the revision of Middle School in Italy."

Later on some of these units started organizing "domestic meetings", although open to the contributions coming from other units. For example, the units in Pavia, Pisa, Trieste, that were working on the G. Prodi's Project "Mathematics as a Discovery" organized about ten of these meetings.

Then, so-called Inter-Units were born, in order to link the different school levels. Nowadays there are few units, which still maintain this definition, as the research in Education, even in Italy, has taken new roads, at the time unthinkable: it has opened up to the comparison with the international research in Education, it has become an independent discipline, with its own research subject and its own language, it has deepened the study and the use of other disciplines, such as psychology, pedagogy, linguistics. However, we can say that everything had its origin in Magenes's brave and far-sighted decision.

4 Magenes and the Research Unit in Pavia: *The Project "Mathematics as a Discovery"*

I just want to say few things, mainly in order to avoid any form of parochialism.

The Pavia Unit was initiated by Magenes and was founded in 1975, one of the first in Italy. Since its beginning, the unit worked together with the units of Pisa and Trieste, in order to test the project "Mathematics as a Discovery," worked out by Giovanni Prodi. Pavia was the administrative base of the three units. I would like to underline the constant and active participation by Magenes to the weekly meetings of the unit. They were not of purely academic interest, but they were rather working meetings, in which suggestions from the Pisa Unit were discussed (mainly due to Prodi and Checcucci), problems and exercises to be inserted in the manual were proposed, new arguments (at least for many members of the Pavia Unit) were studied. Moreover, Magenes did not miss any single meeting, which the three units organized in turns in Pavia, Pisa, Trieste. This is particularly remarkable, as he had a lot of other things to do!

The Project "Mathematics as a Discovery," meant for *Liceo Scientifico* (an Italian High School program with a specific scientific curriculum), is without any doubt the most organic and well-known, among the ones developed by the Research Units. It is mainly due to Giovanni Prodi, who entirely wrote the first two volumes. Magenes contributed to the third one, *G. Prodi and E. Magenes, Elementi di analisi matematica per il triennio delle scuole secondarie superiori, G. D'Anna, 1982, (Elements of Mathematical Analysis for the Second Level of High Schools)*, see [7].

Even though it is the result of a joint work, the first six chapters are mainly due to G. Prodi, whereas the seventh (Differential Calculus) and the eighth (Integral Calculus) are mostly due to Magenes.

These three volumes are endowed with teachers' guides. The guides to the first (1977) and second (1978) volumes, joint work of the three research units, fit, as it is written in the introduction, "amongst the initiatives of the Italian Mathematical Union for the improvement of the teaching of Mathematics at various levels." The Italian Mathematical Union edited the volumes (see [3, 4]). The third guide, authored by L. Bazzini, A. Pesci and M. Reggiani and published in 1985, is the fruit of the Pavia Unit, with the precious collaboration of Prodi and Magenes (see [1]).

In order to complete the description, I would like to add that part of the project were also a "Workbook # 3," devoted to solid Geometry and some complements of plane Geometry, mainly due to V. Checcucci and M.A. Mariotti, and a "Workbook Probability and Statistics," written by L. Piccinato and N. Pintacuda.

The Project "Mathematics as a Discovery" evolved and later became "Discover Mathematics," with a new publisher (Ghisetti & Corvi), a lot of new material, a number of nimble booklets, and even new authors. Magenes personally contributed to this project too. In 2006 it was published the volume *G. Prodi, E. Magenes, M.R. Magenes, A. Pesci and M. Reggiani, Calcolo Differenziale e Integrale, Ghisetti & Corvi, (Differential and Integral Calculus)*, see [8].

The preface underlines that "the third part of the volume is devoted to complements, which are somewhat unusual, but nonetheless equally stimulating," such as "Algorithms and Numerical Methods" and "Expansion in Fourier Series and Wavelets." In writing them, Magenes benefited from the contributions of V. Comincioli and G. Savaré.

5 Two Talks as Keynote Speaker

I would like to recall, without going into details, two talks, which Magenes held in two conferences devoted to topics in education. I reread them, and I found them still topical, even though they are dated.

The first one dates back to 1976, and it was delivered during the Conference "Experiments in Mathematics Education and the Reform of High School" held in Bologna and organized by the Italian Mathematical Union, of which at the time Magenes was still President. The title of the talk was "Present Problems in the Teaching of Mathematics" and it was intended both for mathematics teachers and for a larger audience, namely the visitors of the "Book Fair." The talk was published in the proceedings of the conference, edited by the Italian Mathematical Union (see [5]).

The second one dates back to 1985, and it was delivered at the "Tenth Conference in Mathematics Education: High School," held in Salsomaggiore. The topic of the conference was "New Contents and New Methodologies for the Teaching of Mathematics in High Schools." Magenes talked about "Mathematical Bases for Everybody," with a detailed program proposal, which maintains its validity even nowadays (see [6]).

6 Conclusion

I am well aware that what I have written is a sheer description of facts. Relying on them, every single reader can make up her own mind about Magenes's commitment for the development of the research in Mathematics Education. The research Unit in Pavia, who had Magenes as its guardian angel, will never forget him.

References

1. Bazzini, L., Pesci, A., Reggiani, M.: Guida al Volume "G. Prodi, E. Magenes, Elementi di Analisi Matematica per il Triennio delle Scuole Secondarie Superiori". G. D'Anna, Firenze, Messina (1985)
2. Italian Mathematical Union (ed.): Il X Congresso dell'UMI. Not. Unione Mat. Ital. **2**(12), 1–12 (1975)
3. Italian Mathematical Union (ed.): Guida al Progetto di Insegnamento della Matematica Nelle Scuole Superiori Proposto da G. Prodi. Volume Primo. G. D'Anna, Firenze, Messina (1977)
4. Italian Mathematical Union (ed.): Guida al Progetto di Insegnamento della Matematica Nelle Scuole Superiori Proposto da G. Prodi. Volume Secondo. G. D'Anna, Firenze, Messina (1978)
5. Magenes, E.: Problemi Attuali dell'Insegnamento della Matematica. Not. Unione Mat. Ital. **6**(Suppl.), 122–133 (1976)
6. Magenes, E.: Le Basi Matematiche per Tutti. Not. Unione Mat. Ital. **7**(Suppl.), 7–33 (1986)
7. Prodi, G., Magenes, E.: Elementi di Analisi Matematica per il Triennio delle Scuole Secondarie Superiori. G. D'Anna, Firenze, Messina (1982)
8. Prodi, G., Magenes, E., Magenes, M.R., Pesci, A., Reggiani, M.: Calcolo Differenziale e Calcolo Integrale. Ghisetti Corvi, Milano (2006)

List of Mathematical Works Authored or Edited by Enrico Magenes

Ugo Gianazza

This is a list of mathematical works authored or edited by Enrico Magenes in his long career; it has been compiled relying on a number of different sources, and it is probably the first one ever prepared.

References

1. Magenes, E.: Sui teoremi di Tonelli per la semicontinuità nei problemi di Mayer e di Lagrange. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (2) **15**, 113–125 (1950). 1946
2. Magenes, E.: Sopra un problema di T. Satô per l'equazione differenziale $y'' = f(x, y, y')$. I. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (8) **2**, 130–136 (1947)
3. Magenes, E.: Sopra un problema di T. Satô per l'equazione differenziale $y'' = f(x, y, y')$. II. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (8) **2**, 258–261 (1947)
4. Magenes, E.: Intorno agli integrali di Fubini-Tonelli. I. Condizioni sufficienti per la semicontinuità. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (3) **2**, 1–38 (1950). 1948
5. Magenes, E.: Problemi di valori al contorno per l'equazione differenziale $y^{(n)} = \lambda f(x, y, y', \ldots, y^{(n-1)})$. Ann. Mat. Pura Appl. (4) **27**, 39–74 (1948)
6. Magenes, E.: Una questione di stabilità relativa ad un problema di moto centrale a massa variabile. Pont. Acad. Sci. Comment. **12**, 229–259 (1948)
7. Magenes, E.: Intorno agli integrali di Fubini-Tonelli. II. Teoremi di esistenza dell'estremo. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (3) **3**, 95–131 (1950). 1949
8. Magenes, E.: Proprietà topologiche di certi insiemi di punti e teoremi di esistenza di punti uniti in trasformazioni plurivalenti di una r-cella in sè. G. Mat. Battaglini (4) **2**(78), 168–181 (1949)
9. Magenes, E.: Un criterio di esistenza di punti uniti in trasformazioni topologiche piane. Rend. Semin. Mat. Univ. Padova **18**, 68–114 (1949)
10. Magenes, E.: Sul minimo relativo degli integrali di Fubini-Tonelli. G. Mat. Battaglini (4) **3**(79), 144–168 (1950)
11. Magenes, E.: Sulle equazioni di Eulero relative ai problemi di calcolo delle variazioni degli integrali di Fubini-Tonelli. Rend. Semin. Mat. Univ. Padova **19**, 62–102 (1950)

U. Gianazza (✉)
Dipartimento di Matematica "F. Casorati", Università di Pavia, via Ferrata 1, 27100 Pavia, Italy
e-mail: gianazza@imati.cnr.it

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*,
Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_7,

12. Magenes, E.: Un'osservazione sui teoremi di esistenza di punti uniti in trasformazioni plurivalenti di una N-cella. Rend. Semin. Mat. Univ. Padova **19**, 108–113 (1950)
13. Magenes, E.: Un'osservazione sulle condizioni necessarie per la semicontinuità degli integrali di Fubini-Tonelli. Rend. Semin. Mat. Univ. Padova **19**, 44–53 (1950)
14. Magenes, E.: Condizioni sufficienti per il minimo relativo in certi problemi di Mayer. Rend. Semin. Mat. Univ. Padova **20**, 78–98 (1951)
15. Magenes, E.: Intorno ad un nuovo tipo di funzionali del calcolo delle variazioni. In: Atti III. Congr. Unione Mat. Ital., Pisa, 23–26 Settembre 1948, pp. 102–104 (1951)
16. Magenes, E.: Sul minimo semi-forte degli integrali di Fubini-Tonelli. Rend. Semin. Mat. Univ. Padova **20**, 401–424 (1951)
17. Magenes, E.: Sulle estremanti dei polinomiali nella sfera di Hilbert. Rend. Semin. Mat. Univ. Padova **20**, 24–47 (1951)
18. Magenes, E.: Una questione di stabilità relativa ad un problema di moto centrale a massa variabile. In: Atti III. Congr. Unione Mat. Ital., Pisa, 23–26 Settembre 1948, pp. 105–106 (1951)
19. Magenes, E.: Sul minimo relativo nei problemi di calcolo delle variazioni d'ordine n. Rend. Semin. Mat. Univ. Padova **21**, 1–24 (1952)
20. Magenes, E.: Sull'equazione del calore: teoremi di unicità e teoremi di completezza connessi col metodo di integrazione di M. Picone. I. Rend. Semin. Mat. Univ. Padova **21**, 99–123 (1952)
21. Magenes, E.: Sull'equazione del calore: teoremi di unicità e teoremi di completezza connessi col metodo d'integrazione di M. Picone. II. Rend. Semin. Mat. Univ. Padova **21**, 136–170 (1952)
22. Magenes, E.: Sui problemi al contorno misti per le equazioni lineari del secondo ordine di tipo ellittico. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (3) **8**, 93–120 (1954)
23. Magenes, E.: Osservazioni su alcuni teoremi di completezza connessi con i problemi misti per le equazioni lineari ellittiche. Boll. Unione Mat. Ital. (3) **10**, 452–459 (1955)
24. Magenes, E.: Problema generalizzato di Dirichlet e teoria del potenziale. Rend. Semin. Mat. Univ. Padova **24**, 220–229 (1955)
25. Magenes, E.: Problemi al contorno misti per l'equazione del calore. Rend. Semin. Mat. Univ. Padova **24**, 1–28 (1955)
26. Magenes, E.: Sui problemi di derivata obliqua regolare per le equazioni lineari del secondo ordine di tipo ellittico. Ann. Mat. Pura Appl. (4) **40**, 143–160 (1955)
27. Magenes, E.: Sul teorema dell'alternativa nei problemi misti per le equazioni lineari ellittiche del secondo ordine. Ann. Sc. Norm. Super. Pisa (3) **9**, 161–200 (1956). 1955
28. Magenes, E.: Sulla teoria del potenziale. Rend. Semin. Mat. Univ. Padova **24**, 510–522 (1955)
29. Magenes, E.: Su alcune recenti impostazioni dei problemi al contorno, in particolare misti, per le equazioni lineari ellittiche del secondo ordine. Ann. Sc. Norm. Super. Pisa (3) **10**, 75–84 (1956)
30. Magenes, E.: Il problema della derivata obliqua regolare per le equazioni lineari ellittico-paraboliche del secondo ordine in m variabili. Rend. Mat. Appl. (5) **16**, 363–414 (1957)
31. Magenes, E.: Recenti sviluppi nella teoria dei problemi misti per le equazioni lineari ellittiche. Rend. Semin. Mat. Fis. Milano **27**, 75–95 (1957)
32. Magenes, E.: Sui problemi al contorno per i sistemi di equazioni differenziali lineari ellittici di ordine qualunque. Univ. e Politec. Torino. Rend. Sem. Mat. **17**, 25–45 (1957/1958)
33. Magenes, E., Stampacchia, G.: I problemi al contorno per le equazioni differenziali di tipo ellittico. Ann. Sc. Norm. Super. Pisa (3) **12**, 247–358 (1958)
34. Lions, J.-L., Magenes, E.: Problemi ai limiti non omogenei. I. Ann. Sc. Norm. Super. Pisa (3) **14**, 269–308 (1960)
35. Lions, J.-L., Magenes, E.: Remarque sur les problèmes aux limites pour opérateurs paraboliques. C. R. Acad. Sci. Paris **251**, 2118–2120 (1960)
36. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes. II. Ann. Inst. Fourier (Grenoble) **11**, 137–178 (1961)

37. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes. IV. Ann. Sc. Norm. Super. Pisa (3) **15**, 311–326 (1961)
38. Lions, J.-L., Magenes, E.: Problemi ai limiti non omogenei. III. Ann. Sc. Norm. Super. Pisa (3) **15**, 41–103 (1961)
39. Lions, J.-L., Magenes, E.: Problemi ai limiti non omogenei. V. Ann. Sc. Norm. Super. Pisa (3) **16**, 1–44 (1962)
40. Lions, J.-L., Magenes, E.: Remarques sur les problèmes aux limites linéaires elliptiques. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (8) **32**, 873–883 (1962)
41. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes. VI. J. Anal. Math. **11**, 165–188 (1963)
42. Lions, J.L., Magenes, E.: Problèmes aux limites non homogènes. VII. Ann. Mat. Pura Appl. (4) **63**, 201–224 (1963)
43. Magenes, E.: Sur les problèmes aux limites pour les équations linéaires elliptiques. In: Les Équations aux Dérivées Partielles, Paris, 1962, pp. 95–111. Éditions du Centre National de la Recherche Scientifique, Paris (1963)
44. Lions, J.L., Magenes, E.: Sur certains aspects des problèmes aux limites non homogènes pour des opérateurs paraboliques. Ann. Sc. Norm. Super. Pisa (3) **18**, 303–344 (1964)
45. Magenes, E.: Problèmes de traces et problèmes aux limites pour équations linéaires elliptiques et paraboliques. In: Deuxième Colloq. l'Anal. Fonct. Centre Belge Recherches Math., pp. 83–95. Librairie Universitaire, Louvain (1964)
46. Lions, J.-L., Magenes, E.: Espaces de fonctions et distributions du type de Gevrey et problèmes aux limites paraboliques. Ann. Mat. Pura Appl. (4) **68**, 341–417 (1965)
47. Magenes, E.: Spazi d'interpolazione ed equazioni a derivate parziali. In: Atti del Settimo Congresso dell'Unione Matematica Italiana, Genova, 1963, pp. 134–197. Edizioni Cremonese, Rome (1965)
48. Lions, J.-L., Magenes, E.: Espaces du type de Gevrey et problèmes aux limites pour diverses classes d'équations d'évolution. Ann. Mat. Pura Appl. (4) **72**, 343–394 (1966)
49. Lions, J.-L., Magenes, E.: Résultats de régularité et problèmes aux limites non homogènes pour opérateurs paraboliques. In: Atti del Convegno su le Equazioni alle Derivate Parziali, Nervi, 1965, pp. 75–80. Edizioni Cremonese, Rome (1966)
50. Madženes, È.: Interpolational spaces and partial differential equations. Usp. Mat. Nauk **21**(2), 169–218 (1966)
51. Magenes, E.: Espaces de fonctions et de distributions vectorielles du type de Gevrey et équations différentielles. In: Séminaire sur les Équations aux Dérivées Partielles (1965–1966). I, pp. 1–34. Collège de France, Paris (1966)
52. Lions, J.-L., Magenes, E.: Quelques remarques sur les problèmes aux limites linéaires elliptiques et paraboliques dans des classes d'ultra-distributions. I, II. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (8) **43**, 469–476 (1967)
53. Geymonat, G., Magenes, E.: La teoria delle distribuzioni. In: Annuario della EST, vol. 68, pp. 413–419. Mondadori, Milan (1968)
54. Lions, J.-L., Magenes, E.: Contrôle optimal et espaces du type de Gevrey. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (8) **44**, 34–39 (1968)
55. Lions, J.-L., Magenes, E.: Contrôle optimal et espaces du type de Gevrey. II. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (8) **44**, 151–157 (1968)
56. Magenes, E.: Alcuni aspetti della teoria delle ultradistribuzioni e delle equazioni a derivate parziali. In: Symposia Mathematica, vol. II, INDAM, Rome, 1968, pp. 235–254. Academic Press, London (1969)
57. Gevrey, M.: Œuvres de Maurice Gevrey (With a Preface by E. Magenes and C. Pucci). Éditions du Centre National de la Recherche Scientifique, Paris (1970), pages xvi+573 (1 plate)

58. Magenes, E.: Sui problemi ai limiti per le equazioni di evoluzione, lineari alle derivate parziali, del secondo ordine nel tempo. In: Symposia Mathematica, vol. VII, Convegno sulle Problemi di Evoluzione, INDAM, Rome, 1970, pp. 165–184. Academic Press, London (1971)
59. Magenes, E.: Su alcuni problemi ellittici di frontiera libera connessi con il comportamento dei fluidi nei mezzi porosi. In: Symposia Mathematica, vol. X, Convegno di Analisi Numerica, INDAM, Rome, 1972, pp. 265–279. Academic Press, London (1972)
60. Baiocchi, C., Comincioli, V., Magenes, E., Pozzi, G.A.: Free boundary problems in the theory of fluid flow through porous media: existence and uniqueness theorems. Ann. Mat. Pura Appl. (4) **97**, 1–82 (1973)
61. Magenes, E.: Problèmes de frontière libre liés à certaines questions d'hydraulique. In: Proceedings of Equadiff III (Third Czechoslovak Conf. Differential Equations and their Applications, Brno, 1972). Folia Fac. Sci. Natur. Univ. Purkynianae Brunensis, Ser. Monograph., vol. 1, pp. 51–58. Purkyně University, Brno (1973)
62. Baĭokki, K., Madženes, È.: Free boundary value problems connected with fluid flow through porous materials. Usp. Mat. Nauk **29**(2(176)), 50–69 (1974). Translated from the Italian by T.D. Ventcel', Collection of Articles Dedicated to the Memory of Ivan Georgievič Petrovskiĭ (1901–1973), I
63. Magenes, E.: Problemi Attuali dell'Insegnamento della Matematica. Not. Unione Mat. Ital., Suppl. **6**, 122–133 (1976)
64. Mathematical aspects of finite element methods (1977). Proceedings of the Conference held at the Consiglio Nazionale delle Ricerche (C.N.R.) in Rome, December 10–12, 1975
65. Magenes, E.: Topics in parabolic equations: some typical free boundary problems. In: Boundary Value Problems for Linear Evolution: Partial Differential Equations, Proc. NATO Advanced Study Inst., Liège, 1976. NATO Advanced Study Inst. Ser., Ser. C: Math. and Phys. Sci., vol. 29, pp. 239–312. Reidel, Dordrecht (1977)
66. Magenes, E., Lions, J.-L.: Guido Stampacchia (1922–1978). Boll. Unione Mat. Ital., A (5) **15**(3), 715–756 (1978)
67. Colli Franzone, P., Magenes, E.: On the inverse potential problem of electrocardiology. Calcolo **16**(4), 459–538 (1980). 1979
68. Magenes, E.: Mauro Pagni (1922–1979). Boll. Unione Mat. Ital., A (5) **17**(2), 357–363 (1980)
69. Magenes, E.: Two-phase Stefan problems in several space variables. Matematiche **36**(1), 65–108 (1983). 1981
70. Guerri, L., Mazhenes, E.: An inverse problem of electrocardiology. In: Current Problems in Numerical and Applied Mathematics, Novosibirsk, 1981, pp. 59–72. Nauka, Sibirsk. Otdel., Novosibirsk (1983)
71. Magenes, E.: Carlo Miranda. Ann. Mat. Pura Appl. (4) **135**, i–iii (1984). 1983
72. Magenes, E.: Mathematical problems in electrocardiologic potential theory. In: Methods of Functional Analysis and Theory of Elliptic Equations, Liguori, Naples, 1982, pp. 199–216. Liguori, Naples (1983)
73. Magenes, E., Verdi, C., Visintin, A.: Semigroup approach to the Stefan problem with nonlinear flux. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (8) **75**(1–2), 24–33 (1984). 1983
74. Numerical solutions of nonlinear problems (1984). Lectures presented at the 6th France-Italy-USSR joint symposium of applied mathematics held at the Institut National de Recherche en Informatique et en Automatique (INRIA), Rocquencourt, December 19–21, 1983
75. Colli Franzone, P., Guerri, L., Magenes, E.: Oblique double layer potentials for the direct and inverse problems of electrocardiology. Math. Biosci. **68**(1), 23–55 (1984)
76. Magenes, A., Verdi, C.: The semigroup approach to the two-phase Stefan problem with nonlinear flux conditions. In: Free Boundary Problems: Applications and Theory, vol. III, Maubuisson, 1984. Res. Notes Math., vol. 120, pp. 28–39. Pitman, Boston (1985)

77. Magenes, E.: An inverse problem in the theory of logarithmic potentials. Calcolo **22**(1), 31–46 (1985)
78. Magenes, E.: Le Basi Matematiche per Tutti. Not. Unione Mat. Ital., Suppl. **7**, 7–33 (1986)
79. Magenes, E., Nochetto, R.H., Verdi, C.: Energy error estimates for a linear scheme to approximate nonlinear parabolic problems. Modél. Math. Anal. Numér. **21**(4), 655–678 (1987)
80. Magenes, E.: Remarques sur l'approximation des problèmes paraboliques non linéaires. In: Analyse Mathématique et Applications, pp. 297–318. Gauthier-Villars, Montrouge (1988)
81. Mazhenes, È.: Linear schemes for the approximation of parabolic problems of Stefan type. In: Numerical Analysis and Mathematical Modeling, pp. 144–165. Akad. Nauk SSSR Otdel Vychisl. Mat., Moscow (1988) (in Russian)
82. Magenes, E.: A time-discretization scheme approximating the nonlinear evolution equation $u_t + ABu = 0$. In: Partial Differential Equations and the Calculus of Variations, vol. II. Progr. Nonlinear Differential Equations Appl., vol. 2, pp. 743–765. Birkhäuser Boston, Boston (1989)
83. Magenes, E., Verdi, C., Visintin, A.: Theoretical and numerical results on the two-phase Stefan problem. SIAM J. Numer. Anal. **26**(6), 1425–1438 (1989)
84. Magenes, E.: On the regular oblique derivative problem for harmonic functions. Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. **1**(3), 195–202 (1990)
85. Magenes, E.: Numerical approximation of nonlinear evolution problems. In: Frontiers in Pure and Applied Mathematics, pp. 193–207. North-Holland, Amsterdam (1991)
86. Magenes, E.: On a Stefan problem in a concentrated capacity. In: Nonlinear Analysis. Sc. Norm. Super. di Pisa Quaderni, pp. 217–229. Scuola Norm. Sup., Pisa (1991)
87. Magenes, E.: On the approximation of some non-linear evolution equations. Ric. Mat. **40**(Suppl.), 215–240 (1991). International Symposium in Honor of Renato Caccioppoli, Naples, 1989
88. Magenes, E.: On a Stefan problem on the boundary of a domain. In: Partial Differential Equations and Related Subjects, Trento, 1990. Pitman Res. Notes Math. Ser., vol. 269, pp. 209–226. Longman Sci. Tech., Harlow (1992)
89. Magenes, E.: Some new results on a Stefan problem in a concentrated capacity. Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. **3**(1), 23–34 (1992)
90. Magenes, E.: The Stefan problem in a concentrated capacity. In: Current Problems of Analysis and Mathematical Physics, Taormina, 1992, pp. 155–182. Univ. Roma "La Sapienza", Rome (1993) (in Italian)
91. Magenes, E., Verdi, C.: Time discretization schemes for the Stefan problem in a concentrated capacity. Meccanica **28**, 121–128 (1993)
92. Magenes, E.: Maria Cibrario Cinquini. Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Suppl. **5**, 35–47 (1995). 1994
93. Magenes, E.: Stefan problems in a concentrated capacity. In: Advanced Mathematics: Computations and Applications, Novosibirsk, 1995, pp. 82–90. NCC Publ., Novosibirsk (1995)
94. Magenes, E.: Giuseppe Scorza-Dragoni. Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Suppl. **7**, 69–87 (1997). 1996
95. Magenes, E.: On a Stefan problem in a concentrated capacity. In: Partial Differential Equations and Applications. Lecture Notes in Pure and Appl. Math., vol. 177, pp. 237–253. Dekker, New York (1996)
96. Magenes, E.: On the scientific work of Olga Oleinik. Rend. Mat. Appl. (7) **16**(3), 347–373 (1996)
97. Magenes, E.: Giuseppe Scorza. Boll. Unione Mat. Ital., A (7) **11**(1), 207–216 (1997)
98. Magenes, E.: Giuseppe Scorza-Dragoni (1908–1996). Ann. Mat. Pura Appl. (4) **172**, 1–3 (1997)
99. Magenes, E.: An account of the third congress of the UMI, Pisa, Sept. 23–26, 1948. Boll. Unione Mat. Ital. Sez. A Mat. Soc. Cult. (8) **1**(1), 1–6 (1998)
100. Magenes, E.: Stefan problems with a concentrated capacity. Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8) **1**(1), 71–81 (1998)

101. Magenes, E.: Sur l'opérateur du type Dirichlet-Neumann pour certaines équations paraboliques non linéaires. In: Équations aux dérivées partielles et applications, pp. 655–665. Gauthier-Villars, Éd. Sci. Méd. Elsevier, Paris (1998)
102. Magenes, E.: The UMI in the first post-war period (1945–1951). Boll. Unione Mat. Ital. Sez. A Mat. Soc. Cult. (8) **1**(2), 145–152 (1999). 1998
103. Magenes, E.: In memory of Lamberto Cattabriga. Ann. Univ. Ferrara, Sez. 7 (N.S.) **45**(suppl.), 1–4 (2000). 1999. Workshop on Partial Differential Equations, Ferrara, 1999
104. Lax, P.D., Magenes, E., Temam, R.: Jacques-Louis Lions (1928–2001). Not. Am. Math. Soc. **48**(11), 1315–1321 (2001)
105. Magenes, E.: Remembering Jacques-Louis Lions. Boll. Unione Mat. Ital. Sez. A Mat. Soc. Cult. (8) **4**(2), 185–198 (2001)
106. Magenes, E.: In memoriam Jacques-Louis Lions. Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Suppl. **13**, 19–23 (2004). 2002
107. Magenes, E.: In memoriam Olga Oleinik. Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Suppl. **13**, 25–27 (2004). 2002
108. Magenes, E.: Opening address of the international conference on nonlinear evolution problems. Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. **15**(3–4), 149–159 (2004). Held at L'Accademia Nazionale dei Lincei and L'Università di Roma "La Sapienza", Rome, January 28–31, 2003
109. Magenes, E.: The collaboration between Guido Stampacchia and Jacques-Louis Lions on variational inequalities. In: Variational Analysis and Applications. Nonconvex Optim. Appl., vol. 79, pp. 33–38. Springer, New York (2005)
110. Magenes, E.: Memories of Renato Caccioppoli. Ric. Mat. **54**(2), 333–338 (2006). 2005
111. Magenes, E.: Regularity of the solution to a boundary value problem for a parabolic integrodifferential equation. Rend. Accad. Naz. Sci. XL Mem. Mat. Appl. (5) **29**(1), 271–279 (2005)
112. Integrali singolari e questioni connesse (2011). Lectures from the Centro Internazionale Matematico Estivo (C.I.M.E.) Summer School Held in Varenna, June 10–19, 1957. Reprint of the 1958 original
113. Teoria delle distribuzioni (2011). Lectures from the Centro Internazionale Matematico Estivo (C.I.M.E.) Summer School Held in Saltino, September 1–9, 1961. Reprint of the 1961 original
114. De Giorgi, E., Magenes, E., Mosco, U. (eds.): Proceedings of the International Meeting on Recent Methods in Nonlinear Analysis. Pitagora Editrice, Bologna (1979)
115. Lions, J.-L., Magenes, E.: Neodnorodnye Granichnye Zadachi i Ikh Prilozheniya, tom 1 [Nonhomogeneous Boundary Value Problems and Their Applications, vol. 1]. Izdat. "Mir", Moscow (1971). Translated from the French by L.S. Frank. Edited by V.V. Grušin
116. Lions, J.-L., Magenes, E.: Non-homogeneous Boundary Value Problems and Applications, vol. I. Die Grundlehren der mathematischen Wissenschaften, vol. 181. Springer, New York (1972). Translated from the French by P. Kenneth
117. Lions, J.-L., Magenes, E.: Non-homogeneous Boundary Value Problems and Applications, vol. II. Die Grundlehren der mathematischen Wissenschaften, vol. 182. Springer, New York (1972). Translated from the French by P. Kenneth
118. Lions, J.-L., Magenes, E.: Non-homogeneous Boundary Value Problems and Applications, vol. III. Die Grundlehren der mathematischen Wissenschaften, vol. 183. Springer, New York (1973). Translated from the French by P. Kenneth
119. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes et applications, vol. 1. Travaux et Recherches Mathématiques, vol. 17. Dunod, Paris (1968)
120. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes et applications, vol. 2. Travaux et Recherches Mathématiques, vol. 18. Dunod, Paris (1968)
121. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes et applications, vol. 3. Travaux et Recherches Mathématiques, vol. 20. Dunod, Paris (1970)
122. Magenes, E. (ed.): Free Boundary Problems, vol. I. Istituto Nazionale di Alta Matematica Francesco Severi, Rome (1980)

123. Magenes, E. (ed.): Free Boundary Problems, vol. II. Istituto Nazionale di Alta Matematica Francesco Severi, Rome (1980)
124. Magenes, E. (ed.): Seminari su la risoluzione numerica delle equazioni differenziali ordinarie di tipo "stiff". Segreteria del Laboratorio di Analisi Numerical del C.N.R., Pavia (1972). Con una prefazione di Enrico Magenes, Laboratorio di Anal. Numer. Consiglio Naz. Ricerche Pubbl. No. 36 (1972)
125. Prodi, G., Magenes, E.: Elementi di Analisi Matematica per il Triennio delle Scuole Secondarie Superiori. Casa Editrice G. D'Anna, Messina–Firenze (1982)
126. Prodi, G., Magenes, E., Magenes, M.R., Pesci, A., Reggiani, M.: Calcolo Differenziale e Calcolo Integrale. Ghisetti Corvi, Milano (2006)

Part II
Recent Developments

Heat Flow and Calculus on Metric Measure Spaces with Ricci Curvature Bounded Below—The Compact Case

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré

Abstract We provide a quick overview of various calculus tools and of the main results concerning the heat flow on compact metric measure spaces, with applications to spaces with lower Ricci curvature bounds. Topics include the Hopf-Lax semigroup and the Hamilton-Jacobi equation in metric spaces, a new approach to differentiation and to the theory of Sobolev spaces over metric measure spaces, the equivalence of the L^2-gradient flow of a suitably defined "Dirichlet energy" and the Wasserstein gradient flow of the relative entropy functional, a metric version of Brenier's Theorem, and a new (stronger) definition of Ricci curvature bound from below for metric measure spaces. This new notion is stable w.r.t. measured Gromov-Hausdorff convergence and it is strictly connected with the linearity of the heat flow.

1 Introduction

Aim of these notes is to provide a quick overview of the main results contained in [4] and [6] in the simplified case of compact metric spaces (X, d) endowed with a reference probability measure $\mathfrak{m}$. The idea is to give the interested reader the possibility to get as quickly as possible the key ideas behind the proofs of our recent results, neglecting all the problems that appear in a more general framework (as a matter of fact, no compactness assumption is made in [4, 6] and finiteness of $\mathfrak{m}$ is assumed only in [6]). Passing from compact spaces to complete and separable ones

To the memory of Enrico Magenes, whose exemplar life, research and teaching shaped generations of mathematicians.

L. Ambrosio
Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
e-mail: l.ambrosio@sns.it

N. Gigli
Université de Nice, Mathématiques, Parc Valrose, 06108 Nice, France
e-mail: nicola.gigli@unice.fr

G. Savaré (✉)
Dipartimento di Matematica "F. Casorati", Università di Pavia, via Ferrata 1, 27100 Pavia, Italy
e-mail: giuseppe.savare@unipv.it

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*,
Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_8,

(and even to a more general framework which includes the so-called Wiener space) is not just a technical problem, meaning that several concepts need to be properly adapted in order to achieve such generality. Hence, in particular, the discussion here is by no means exhaustive, as both the key statements and the auxiliary lemmas are stated in the simplified case of a probability measure in a compact space.

Apart some very basic concepts about optimal transport, Wasserstein distance and gradient flows, this paper pretends to be self-contained. All the concepts that we need are recalled in the preliminary section, whose proofs can be found, for instance, in the first three chapters of [1] (for an overview on the theory of gradient flows, see also [3], and for a much broader discussion on optimal transport, see the monograph by Villani [32]). For completeness reasons, we included in our discussion some results coming from previous contributions which are potentially less known, in particular: the (sketch of the) proof by Lisini [22] of the characterization of absolutely continuous curves w.r.t. the Wasserstein distance (Proposition 4.13), and the proof of uniqueness of the gradient flow of the relative entropy w.r.t. the Wasserstein distance on spaces with Ricci curvature bounded below in the sense of Lott-Sturm-Villani ($CD(K,\infty)$ spaces in short) given by the second author in [12] (Theorem 5.9).

In summary, the main arguments and results that we present here are the following.

(1) The Hopf-Lax formula produces subsolutions of the Hamilton-Jacobi equation, and solutions on geodesic spaces (Theorem 3.2 and Theorem 3.3).
(2) A new approach to the theory of Sobolev spaces over metric measure spaces, which leads in particular to the proof that *Lipschitz functions are always dense in energy in* $W^{1,2}(X,\mathsf{d},\mathfrak{m})$ (Theorem 4.7).
(3) The uniqueness of the gradient flow w.r.t. the Wasserstein distance W_2 of the relative entropy in $CD(K,\infty)$ spaces (Theorem 5.9).
(4) The identification of the L^2-gradient flow of the natural "Dirichlet energy" and the W_2-gradient flow of the relative entropy in $CD(K,\infty)$ spaces (see also [15] for the Alexandrov case, a paper to which our paper [4] owes a lot).
(5) A metric version of Brenier's theorem valid in spaces having Ricci curvature bounded from below in a sense slightly stronger than the one proposed by Lott-Sturm-Villani. If this curvature assumption holds (Definition 7.11) and μ,ν are absolutely continuous w.r.t. $\mathfrak{m}$, then "the distance traveled is uniquely determined by the starting point", i.e. there exists a map $D: X\to\mathbb{R}$ such that for any optimal plan $\boldsymbol{\gamma}$ it holds $\mathsf{d}(x,y)=D(x)$ for $\boldsymbol{\gamma}$-a.e. (x,y). Moreover, the map D is nothing but the weak gradient (according to the theory illustrated in Sect. 4) of any Kantorovich potential. See Theorem 7.11.
(6) A key lemma (Lemma 8.7) concerning "horizontal" and "vertical" differentiation: it allows to compare the derivative of the squared Wasserstein distance along the heat flow with the derivative of the relative entropy along a geodesic.
(7) A new (stronger) definition of Ricci curvature bound from below for metric measure spaces which is stable w.r.t. measured Gromov-Hausdorff convergence and rules out Finsler geometries (Theorem 9.12 and the discussion thereafter).

2 Preliminary Notions

As a general convention, we will always denote by (X, d) a compact metric space and by $\mathfrak{m}$ a Borel probability measure on X; we will always refer to the structure $(X, \mathsf{d}, \mathfrak{m})$ as a compact and normalized metric measure space. We will use the symbol (Y, d_Y) for metric spaces when the compactness is not implicitly assumed.

2.1 Absolutely Continuous Curves and Slopes

Let (Y, d_Y) be a complete and separable metric space, $J \subset \mathbb{R}$ an interval with nonempty interior and $J \ni t \mapsto \gamma_t \in Y$. We say that γ_t is *absolutely continuous* if

$$\mathsf{d}_Y(\gamma_s, \gamma_t) \leq \int_t^s g(r)\mathrm{d}r, \quad \forall s, t \in J,\ t < s$$

for some $g \in L^1(J)$. It turns out that, if γ_t is absolutely continuous, there is a minimal function g with this property, called *metric speed* and given for a.e. $t \in J$ by

$$|\dot{\gamma}_t| = \lim_{s \to t} \frac{\mathsf{d}_Y(\gamma_s, \gamma_t)}{|s - t|}.$$

See [3, Theorem 1.1.2] for the simple proof. Notice that the absolute continuity property of the integral ensures that absolutely continuous functions can be extended by continuity to the closure of their domain.

We will denote by $C([0, 1], Y)$ the space of continuous curves on $[0, 1]$ with values in Y endowed with the sup norm. The set $AC^2([0, 1], Y) \subset C([0, 1], Y)$ consists of all absolutely continuous curves γ such that $\int_0^1 |\dot{\gamma}_t|^2 \mathrm{d}t < \infty$: it is easily seen to be equal to the countable union of the closed sets $\{\gamma : \int_0^1 |\dot{\gamma}_t|^2 \mathrm{d}t \leq n\}$, and thus it is a Borel subset of $C([0, 1], Y)$. The *evaluation maps* $\mathrm{e}_t : C([0, 1], Y) \to Y$ are defined by

$$\mathrm{e}_t(\gamma) := \gamma_t,$$

and are clearly 1-Lipschitz.

We say that a subset D of Y is geodesic if for any $x, y \in D$ there exists a curve $(\gamma_t) \subset D$ on $[0, 1]$ such that $\gamma_0 = x$, $\gamma_1 = y$ and $\mathsf{d}_Y(\gamma_t, \gamma_s) = |t - s|\mathsf{d}_Y(x, y)$ for all $s, t \in [0, 1]$. Such a curve is called constant speed geodesic, or simply geodesic. The space of all geodesics in Y endowed with the sup distance will be denoted by $\mathrm{Geo}(Y)$.

Given $f : Y \to \mathbb{R} \cup \{\pm\infty\}$ we define the *slope* (also called local Lipschitz constant) at points x where $f(x) \in \mathbb{R}$ by

$$|Df|(x) := \overline{\lim_{y \to x}} \frac{|f(y) - f(x)|}{\mathsf{d}_Y(y, x)}.$$

We shall also need the one-sided counterparts of the slope called respectively *descending slope* and *ascending slope*:

$$\left|D^- f\right|(x) := \overline{\lim_{y\to x}} \frac{[f(y)-f(x)]^-}{\mathsf{d}_Y(y,x)}, \qquad \left|D^+ f\right|(x) := \overline{\lim_{y\to x}} \frac{[f(y)-f(x)]^+}{\mathsf{d}_Y(y,x)}, \tag{1}$$

where $[\cdot]^+$ and $[\cdot]^-$ denote respectively the positive and negative part. Notice the change of notation w.r.t. previous works of the authors: the slopes and its one-sided counterparts were denoted by $|\nabla f|$, $|\nabla^\pm f|$. Yet, as remarked in [13], these notions, being defined in duality with the distance, are naturally cotangent notions, rather than tangent ones, whence the notation proposed here.

It is not difficult to see that for f Lipschitz the slopes and the local Lipschitz constant are upper gradients according to [18], namely

$$\left|\int_{\partial\gamma} f\right| \le \int_\gamma \left|D^\pm f\right|$$

for any absolutely continuous curve $\gamma : [0,1] \to Y$; here and in the following we write $\int_{\partial\gamma} f$ for $f(\gamma_1) - f(\gamma_0)$ and $\int_\gamma g$ for $\int_0^1 g(\gamma_s)|\dot\gamma_s|\mathrm{d}s$.

Also, for $f, g : Y \to \mathbb{R}$ Lipschitz it clearly holds

$$\left|D(\alpha f + \beta g)\right| \le |\alpha||Df| + |\beta||Dg|, \quad \forall \alpha, \beta \in \mathbb{R}; \tag{2a}$$

$$\left|D(fg)\right| \le |f||Dg| + |g||Df|. \tag{2b}$$

2.2 The Space $(\mathscr{P}(X), W_2)$

Let (X, d) be a compact metric space. The set $\mathscr{P}(X)$ consists of all Borel probability measures on X. As usual, if $\mu \in \mathscr{P}(X)$ and $T : X \to Y$ is a μ-measurable map with values in the topological space Y, the push-forward measure $T_\sharp\mu \in \mathscr{P}(Y)$ is defined by $T_\sharp\mu(B) := \mu(T^{-1}(B))$ for every set Borel set $B \subset Y$.

Given $\mu, \nu \in \mathscr{P}(X)$, we define the Wasserstein distance $W_2(\mu,\nu)$ between them as

$$W_2^2(\mu,\nu) := \min \int \mathsf{d}^2(x,y)\mathrm{d}\boldsymbol{\gamma}(x,y), \tag{3}$$

where the minimum is taken among all Borel probability measures $\boldsymbol{\gamma}$ on X^2 such that

$$\pi^1_\sharp\boldsymbol{\gamma} = \mu, \qquad \pi^2_\sharp\boldsymbol{\gamma} = \nu; \quad \text{here } \pi^i : X^2 \to X,\ \pi^i(x_1,x_2) := x_i.$$

Such measures are called admissible plans or couplings for the couple (μ,ν); a plan $\boldsymbol{\gamma}$ which realizes the minimum in (3) is called optimal, and we write $\boldsymbol{\gamma} \in \mathrm{OPT}(\mu,\nu)$. From the linearity of the admissibility condition we get that the squared Wasserstein distance is convex, i.e.:

$$W_2^2\big((1-\lambda)\mu_1 + \lambda\mu_2, (1-\lambda)\nu_1 + \lambda\nu_2\big) \le (1-\lambda)W_2^2(\mu_1,\nu_1) + \lambda W_2^2(\mu_2,\nu_2). \tag{4}$$

It is also well known (see e.g. Theorem 2.7 in [1]) that the Wasserstein distance metrizes the weak convergence of measures in $\mathscr{P}(X)$, i.e. the weak convergence with respect to the duality with $C(X)$; in particular $(\mathscr{P}(X), W_2)$ is a compact metric space.

An equivalent definition of W_2 comes from the dual formulation of the transport problem:

$$\frac{1}{2}W_2^2(\mu,\nu) = \sup_{\psi} \int_X \psi \mathrm{d}\mu + \int_X \psi^c \mathrm{d}\nu, \tag{5}$$

the supremum being taken among all Lipschitz functions ψ, where the c-transform in this formula is defined by

$$\psi^c(y) := \inf_{x\in X} \frac{\mathsf{d}^2(x,y)}{2} - \psi(x).$$

A function $\psi : X \to \mathbb{R}$ is said to be c-concave if $\psi = \phi^c$ for some $\phi : X \to \mathbb{R}$. It is possible to prove that the supremum in (5) is always achieved by a c-concave function, and we will call any such function ψ a Kantorovich potential. We shall also use the fact that c-concave functions satisfy

$$\psi^{cc} = \psi. \tag{6}$$

The (graph of the) c-superdifferential $\partial^c\psi$ of a c-concave function ψ is the subset of X^2 defined by

$$\partial^c\psi := \left\{(x,y) \ : \ \psi(x) + \psi^c(y) = \frac{\mathsf{d}^2(x,y)}{2}\right\},$$

and the c-superdifferential $\partial^c\psi(x)$ at x is the set of y's such that $(x,y) \in \partial^c\psi$. A consequence of the compactness of X is that any c-concave function ψ is Lipschitz and that the set $\partial^c\psi(x)$ is non empty for any $x \in X$.

It is not difficult to see that if ψ is a Kantorovich potential for $\mu, \nu \in \mathscr{P}(X)$ and $\boldsymbol{\gamma}$ is a coupling for (μ,ν) then $\boldsymbol{\gamma}$ is optimal if and only if $\mathrm{supp}(\gamma) \subset \partial^c\psi$.

If (X,d) is geodesic, then so is $(\mathscr{P}(X), W_2)$, and in this case a curve (μ_t) is a constant speed geodesic from μ_0 to μ_1 if and only if there exists a measure $\boldsymbol{\pi} \in \mathscr{P}(C([0,1],X))$ concentrated on $\mathrm{Geo}(X)$ such that $(\mathrm{e}_t)_\sharp\boldsymbol{\pi} = \mu_t$ for all $t \in [0,1]$ and $(\mathrm{e}_0,\mathrm{e}_1)_\sharp\boldsymbol{\pi} \in \mathrm{OPT}(\mu_0,\mu_1)$. We will denote the set of such measures, called optimal geodesic plans, by $\mathrm{GeoOpt}(\mu_0,\mu_1)$.

2.3 Geodesically Convex Functionals and Their Gradient Flows

Given a geodesic space (Y,d_Y) (in the following this will always be the Wasserstein space built over a geodesic space (X,d)), a functional $E : Y \to \mathbb{R}\cup\{+\infty\}$ is said

K-geodesically convex (or simply K-convex) if for any $y_0, y_1 \in Y$ there exists a constant speed geodesic $\gamma : [0, 1] \to Y$ such that $\gamma_0 = y_0$, $\gamma_1 = y_1$ and

$$E(\gamma_t) \le (1-t)E(y_0) + tE(y_1) - \frac{K}{2}t(1-t)\mathsf{d}_Y^2(y_0, y_1), \quad \forall t \in [0, 1].$$

We will denote by $D(E)$ the domain of E i.e. $D(E) := \{y : E(y) < \infty\}$: if E is K-geodesically convex, then $D(E)$ is geodesic.

An easy consequence of the K-convexity is the fact that the descending slope defined in (1) can de computed as a sup, rather than as a limsup:

$$|D^-E|(y) = \sup_{z \neq y}\left(\frac{E(y) - E(z)}{\mathsf{d}_Y(y, z)} + \frac{K}{2}\mathsf{d}_Y(y, z)\right)^+. \tag{7}$$

What we want to discuss here is the definition of gradient flow of a K-convex functional. There are essentially two different ways of giving such a notion in a metric setting. The first one, which we call Energy Dissipation Equality (EDE), ensures existence for any K-convex and lower semicontinuous functional (under suitable compactness assumptions), the second one, which we call Evolution Variation Inequality (EVI), ensures uniqueness and K-contractivity of the flow. However, the price we pay for these stronger properties is that existence results for EVI solutions hold under much more restrictive assumptions.

It is important to distinguish the two notions. The EDE one is the "correct one" to be used in a general metric context, because it ensures existence for any initial datum in the domain of the functional. However, typically gradient flows in the EDE sense are not unique: this is the reason of the analysis made in Sect. 5, which ensures that for the special case of the entropy functional uniqueness is indeed true.

EVI gradient flows are in particular gradient flows in the EDE sense (see Proposition 2.2), ensure uniqueness, K-contractivity and provide strong a priori regularizing effects. Heuristically speaking, existence of gradient flows in the EVI sense depends also on properties of the distance, rather than on properties of the functional only. A more or less correct way of thinking at this is: gradient flows in the EVI sense exist if and only if the distance is Hilbertian on small scales. For instance, if the underlying metric space is an Hilbert space, then the two notions coincide.

Now recall that one of our goals here is to study the gradient flow of the relative entropy in spaces with Ricci curvature bounded below (Definition 5.9), and recall that Finsler geometries are included in this setting (see page 926 of [32]). Thus, in general we must deal with the EDE notion of gradient flow. The EVI one will come into play in Sect. 9, where we use it to identify those spaces with Ricci curvature bounded below which are more 'Riemannian like'.

Note: later on we will refer to gradient flows in the EDE sense simply as "gradient flows", keeping the distinguished notation EVI-gradient flows for those in the EVI sense.

Energy Dissipation Equality

An important property of K-geodesically convex and lower semicontinuous functionals (see Corollary 2.4.10 of [3] or Proposition 3.19 of [1]) is that the descending slope is an upper gradient, that is: for any absolutely continuous curve $y_t : J \subset \mathbb{R} \to D(E)$ it holds

$$\left|E(y_t) - E(y_s)\right| \leq \int_t^s |\dot{y}_r| \left|D^- E\right|(y_r)\mathrm{d}r, \quad \forall t \leq s. \tag{8}$$

An application of Young inequality gives that

$$E(y_t) \leq E(y_s) + \frac{1}{2}\int_t^s |\dot{y}_r|^2 \mathrm{d}r + \frac{1}{2}\int_t^s |D^- E|^2(y_r)\mathrm{d}r, \quad \forall t \leq s. \tag{9}$$

This inequality motivates the following definition:

Definition 2.1 (Energy Dissipation Equality definition of gradient flow) Let E be a K-convex and lower semicontinuous functional and let $y_0 \in D(E)$. We say that a continuous curve $[0, \infty) \ni t \mapsto y_t$ is a gradient flow for E in the EDE sense (or simply a gradient flow) if it is locally absolutely continuous in $(0, \infty)$, it takes values in the domain of E and it holds

$$E(y_t) = E(y_s) + \frac{1}{2}\int_t^s |\dot{y}_r|^2 \mathrm{d}r + \frac{1}{2}\int_t^s |D^- E|^2(y_r)\mathrm{d}r, \quad \forall t \leq s. \tag{10}$$

Notice that, due to (9), the equality (10) is equivalent to

$$E(y_0) \geq E(y_s) + \frac{1}{2}\int_0^s |\dot{y}_r|^2 \mathrm{d}r + \frac{1}{2}\int_0^s |D^- E|^2(y_r)\mathrm{d}r, \quad \forall s > 0. \tag{11}$$

Indeed, if (11) holds, then (10) holds with $t = 0$, and then by the additivity of the integral (10) holds in general.

It is not hard to check that if $E : \mathbb{R}^d \to \mathbb{R}$ is a C^1 function, then a curve $y_t : J \to \mathbb{R}^d$ is a gradient flow according to the previous definition if and only if it satisfies

$$y_t' = -\nabla E(y_t), \quad \forall t \in J,$$

so that the metric definition reduces to the classical one when specialized to Euclidean spaces.

The following theorem has been proved in [3] (Corollary 2.4.11):

Theorem 2.1 (Existence of gradient flows in the EDE sense) *Let (Y, d_Y) be a compact metric space and let $E : Y \to \mathbb{R} \cup \{+\infty\}$ be a K-geodesically convex and lower semicontinuous functional. Then every $y_0 \in D(E)$ is the starting point of a gradient flow in the EDE sense of E.*

It is important to stress the fact that in general gradient flows in the EDE sense are *not* unique. A simple example is $Y := \mathbb{R}^2$ endowed with the L^∞ norm, and E defined by $E(x, y) := x$. It is immediate to see that E is 0-convex and that for any point (x_0, y_0) there exist uncountably many gradient flows in the EDE sense starting from it, for instance all curves $(x_0 - t, y(t))$ with $|y'(t)| \le 1$ and $y(0) = y_0$.

Evolution Variational Inequality

To see where the EVI notion comes from, notice that for a K-convex and smooth function f on $\mathbb{R}^d$ it holds $y_t' = -\nabla f(y)$ for any $t \ge 0$ if and only if

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{|y_t - z|^2}{2} + \frac{K}{2}|y_t - z|^2 + f(y_t) \le f(z), \quad \forall z \in \mathbb{R}^d,\ \forall t \ge 0. \tag{12}$$

This equivalence is true because K-convexity ensures that $v = -\nabla f(y)$ if and only

$$\langle v, y - z\rangle + \frac{K}{2}|y - z|^2 + f(y) \le f(z), \quad \forall z \in \mathbb{R}^d.$$

Inequality (12) can be written in a metric context in several ways, which we collect in the following statement (we omit the easy proof).

Proposition 2.1 (Evolution Variational Inequality: equivalent statements) *Let (Y, d_Y) be a complete and separable metric space, $E : Y \to (-\infty, \infty]$ a lower semicontinuous functional, and (y_t) a locally absolutely continuous curve in Y. Then the following properties are equivalent*:

(i) *For any $z \in D(E)$ it holds*

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mathsf{d}_Y^2(y_t, z)}{2} + \frac{K}{2}\mathsf{d}_Y^2(y_t, z) + E(y_t) \le E(z), \quad \text{for a.e. } t \in (0, \infty).$$

(ii) *For any $z \in D(E)$ it holds $\forall 0 < t < s < \infty$*

$$\frac{\mathsf{d}_Y^2(y_s, z) - \mathsf{d}_Y^2(y_t, z)}{2h} + \frac{K}{2}\int_t^s \mathsf{d}_Y^2(y_r, z)\mathrm{d}r + \int_t^s E(y_r)\mathrm{d}r \le (s - t)E(z).$$

(iii) *There exists a set $A \subset D(E)$ dense in energy (i.e., for any $z \in D(E)$ there exists $(z_n) \subset A$ converging to z such that $E(z_n) \to E(z)$) such that for any $z \in A$ it holds*

$$\varlimsup_{h \downarrow 0} \frac{\mathsf{d}_Y^2(y_{t+h}, z) - \mathsf{d}_Y^2(y_t, z)}{2} + \frac{K}{2}\mathsf{d}_Y^2(y_t, z) + E(y_t) \le E(z), \quad \forall t \in (0, \infty).$$

Definition 2.2 (Evolution Variational Inequality definition of gradient flow) We say that a curve (y_t) is a gradient flow of E in the EVI sense relative to $K \in \mathbb{R}$ (in short, EVI_K-gradient flow), if any of the above equivalent properties are true. We say that y_t starts from y_0 if $y_t \to y_0$ as $t \downarrow 0$.

This definition of gradient flow is stronger than the one discussed in the previous section, because of the following result proved by the third author in [29] (see also Proposition 3.6 of [1]), which we state without proof.

Proposition 2.2 (EVI implies EDE) *Let (Y, d_Y) be a complete and separable metric space, $K \in \mathbb{R}$, $E : Y \to (-\infty, \infty]$ a lower semicontinuous functional and $y_t : (0, \infty) \to D(E)$ a locally absolutely continuous curve. Assume that y_t is an EVI_K-gradient flow for E. Then (10) holds for any $0 < t < s$.*

Remark 2.1 (Contractivity) It can be proved that if (y_t) and (z_t) are gradient flows in the EVI_K sense of the l.s.c. functional E, then

$$\mathsf{d}_Y(y_t, z_t) \le e^{-Kt}\mathsf{d}_Y(y_0, z_0), \quad \forall t \ge 0.$$

In particular, gradient flows in the EVI sense are unique. This contractivity property, used in conjunction with (ii) of Proposition 2.1, guarantees that if existence of gradient flows in the EVI sense is known for initial data lying in some subset $S \subset Y$, then it is also known for initial data in the closure $\overline{S}$ of S.

We also point out the following geometric consequence of the EVI, proven in [10].

Proposition 2.3 *Let $E : Y \to (-\infty, \infty]$ be a lower semicontinuous functional on a complete space (Y, d_Y). Assume that every $y_0 \in D(E)$ is the starting point of an EVI_K-gradient flow of E. Then E is K-convex along* all *geodesics contained in $\overline{D(E)}$.*

As we already said, gradient flows in the EVI sense do not necessarily exist, and their existence depends on the properties of the distance d_Y. For instance, it is not hard to see that if we endow $\mathbb{R}^2$ with the L^∞ norm and consider the functional $E(x, y) := x$, then there re is no gradient flow in the EVI_K-sense, regardless of the constant K.

3 Hopf-Lax Formula and Hamilton-Jacobi Equation

Aim of this subsection is to study the properties of the Hopf-Lax formula in a metric setting and its relations with the Hamilton-Jacobi equation. Here we assume that (X, d) is a compact metric space. Notice that there is no reference measure $\mathfrak{m}$ in the discussion.

Let $f : X \to \mathbb{R}$ be a Lipschitz function. For $t > 0$ define

$$F(t, x, y) := f(y) + \frac{\mathsf{d}^2(x, y)}{2t},$$

and the function $Q_t f : X \to \mathbb{R}$ by

$$Q_t f(x) := \inf_{y \in X} F(t, x, y) = \min_{y \in X} F(t, x, y).$$

Also, we introduce the functions $D^+, D^- : X \times (0, \infty) \to \mathbb{R}$ as

$$\begin{aligned} D^+(x, t) &:= \max \mathsf{d}(x, y), \\ D^-(x, t) &:= \min \mathsf{d}(x, y), \end{aligned} \tag{13}$$

where, in both cases, the y's vary among all minima of $F(t, x, \cdot)$. We also set $Q_0 f = f$ and $D^{\pm}(x, 0) = 0$. Thanks to the continuity of F and the compactness of X, it is easy to check that the map $[0, \infty) \times X \ni (t, x) \mapsto Q_t f(x)$ is continuous. Furthermore, the fact that f is Lipschitz easily yields

$$D^-(x, t) \le D^+(x, t) \le 2t \operatorname{Lip}(f), \tag{14}$$

and from the fact that the functions $\{\mathsf{d}^2(\cdot, y)\}_{y \in Y}$ are uniformly Lipschitz (because (X, d) is bounded) we get that $Q_t f$ is Lipschitz for any $t > 0$.

Proposition 3.4 (Monotonicity of $D^{\pm}$) *For all $x \in X$ it holds*

$$D^+(x, t) \le D^-(x, s), \quad 0 \le t < s. \tag{15}$$

As a consequence, $D^+(x, \cdot)$ and $D^-(x, \cdot)$ are both nondecreasing, and they coincide with at most countably many exceptions in $[0, \infty)$.

Proof Fix $x \in X$. For $t = 0$ there is nothing to prove. Now pick $0 < t < s$ and choose x_t and x_s minimizers of $F(t, x, \cdot)$ and $F(s, x, \cdot)$ respectively, such that $\mathsf{d}(x, x_t) = D^+(x, t)$ and $\mathsf{d}(x, x_s) = D^-(x, s)$. The minimality of x_t, x_s gives

$$\begin{aligned} f(x_t) + \frac{\mathsf{d}^2(x_t, x)}{2t} &\le f(x_s) + \frac{\mathsf{d}^2(x_s, x)}{2t}, \\ f(x_s) + \frac{\mathsf{d}^2(x_s, x)}{2s} &\le f(x_t) + \frac{\mathsf{d}^2(x_t, x)}{2s}. \end{aligned}$$

Adding up and using the fact that $\frac{1}{t} \ge \frac{1}{s}$ we deduce

$$D^+(x, t) = \mathsf{d}(x_t, x) \le \mathsf{d}(x_s, x) = D^-(x, s),$$

which is (15).

Combining this with the inequality $D^- \le D^+$ we immediately obtain that both functions are nondecreasing. At a point of right continuity of $D^-(x, \cdot)$ we get

$$D^+(x, t) \le \inf_{s > t} D^-(x, s) = D^-(x, t).$$

This implies that the two functions coincide out of a countable set. □

Next, we examine the semicontinuity properties of $D^{\pm}$. These properties imply that points (x,t) where the equality $D^+(x,t) = D^-(x,t)$ occurs are continuity points for both D^+ and D^-.

Proposition 3.5 (Semicontinuity of $D^{\pm}$) *The map D^+ is upper semicontinuous and the map D^- is lower semicontinuous in $X \times (0,\infty)$.*

Proof We prove lower semicontinuity of D^-, the proof of upper semicontinuity of D^+ being similar. Let (x_i,t_i) be any sequence converging to (x,t) and, for every i, let (y_i) be a minimum of $F(t_i,x_i,\cdot)$ for which $\mathsf{d}(y_i,x_i) = D^-(x_i,t_i)$. For all i we have

$$f(y_i) + \frac{\mathsf{d}^2(y_i,x_i)}{2t_i} = Q_{t_i} f(x_i).$$

Moreover, the continuity of $(x,t) \mapsto Q_t f(x)$ gives that $\lim_i Q_{t_i} f(x_i) = Q_t f(x)$, thus

$$\lim_{i\to\infty} f(y_i) + \frac{\mathsf{d}^2(y_i,x)}{2t} = Q_t f(x).$$

This means that (y_i) is a minimizing sequence for $F(t,x,\cdot)$. Since (X,d) is compact, possibly passing to a subsequence, not relabeled, we may assume that (y_i) converges to y as $i \to \infty$. Therefore

$$D^-(x,t) \le \mathsf{d}(x,y) = \lim_{i\to\infty} \mathsf{d}(x,y_i) = \lim_{i\to\infty} D^-(x_i,t_i). \qquad \square$$

Proposition 3.6 (Time derivative of $Q_t f$) *The map $t \mapsto Q_t f$ is Lipschitz from $[0,\infty)$ to $C(X)$ and, for all $x \in X$, it satisfies*

$$\frac{\mathrm{d}}{\mathrm{d}t} Q_t f(x) = -\frac{[D^{\pm}(x,t)]^2}{2t^2}, \tag{16}$$

for any $t > 0$ with at most countably many exceptions.

Proof Let $t < s$ and x_t, x_s be minima of $F(t,x,\cdot)$ and $F(s,x,\cdot)$. We have

$$Q_s f(x) - Q_t f(x) \le F(s,x,x_t) - F(t,x,x_t) = \frac{\mathsf{d}^2(x,x_t)}{2}\frac{t-s}{ts},$$

$$Q_s f(x) - Q_t f(x) \ge F(s,x,x_s) - F(t,x,x_s) = \frac{\mathsf{d}^2(x,x_s)}{2}\frac{t-s}{ts},$$

which gives that $t \mapsto Q_t f(x)$ is Lipschitz in $(\varepsilon,+\infty)$ for any $\varepsilon > 0$ and $x \in X$. Also, dividing by $(s-t)$ and taking Proposition 3.4 into account, we get (16). Now notice that from (14) we get that $|\frac{\mathrm{d}}{\mathrm{d}t} Q_t f(x)| \le 2\,\mathrm{Lip}^2(f)$ for any x and a.e. t, which, together with the pointwise convergence of $Q_t f$ to f as $t \downarrow 0$, yields that $t \mapsto Q_t f \in C(X)$ is Lipschitz in $[0,\infty)$. $\square$

Proposition 3.7 (Bound on the local Lipschitz constant of $Q_t f$) *For* $(x,t) \in X \times (0,\infty)$ *it holds*:

$$|DQ_t f|(x) \le \frac{D^+(x,t)}{t}. \tag{17}$$

Proof Fix $x \in X$ and $t \in (0,\infty)$, pick a sequence (x_i) converging to x and a corresponding sequence (y_i) of minimizers for $F(t,x_i,\cdot)$ and similarly a minimizer y of $F(t,x,\cdot)$. We start proving that

$$\overline{\lim_{i\to\infty}} \frac{Q_t f(x) - Q_t f(x_i)}{d(x,x_i)} \le \frac{D^+(x,t)}{t}.$$

Since it holds

$$\begin{aligned} Q_t f(x) - Q_t f(x_i) &\le F(t,x,y_i) - F(t,x_i,y_i) \\ &\le f(y_i) + \frac{\mathsf{d}^2(x,y_i)}{2t} - f(y_i) - \frac{\mathsf{d}^2(x_i,y_i)}{2t} \\ &\le \frac{\mathsf{d}(x,x_i)}{2t}\big(\mathsf{d}(x,y_i) + \mathsf{d}(x_i,y_i)\big) \\ &\le \frac{\mathsf{d}(x,x_i)}{2t}\big(\mathsf{d}(x,x_i) + 2D^+(x_i,t)\big), \end{aligned}$$

dividing by $d(x,x_i)$, letting $i \to \infty$ and using the upper semicontinuity of D^+ we get the claim. To conclude, we need to show that

$$\overline{\lim_{i\to\infty}} \frac{Q_t f(x_i) - Q_t f(x)}{d(x,x_i)} \le \frac{D^+(x,t)}{t}.$$

This follows along similar lines starting from the inequality

$$Q_t f(x_i) - Q_t f(x) \le F(t,x_i,y) - F(t,x,y_i). \qquad \square$$

Theorem 3.2 (Subsolution of HJ) *For every* $x \in X$ *it holds*

$$\frac{\mathrm{d}}{\mathrm{d}t} Q_t f(x) + \frac{1}{2}|DQ_t f|^2(x) \le 0 \tag{18}$$

with at most countably many exceptions in $(0,\infty)$.

Proof The claim is a direct consequence of Proposition 3.6 and Proposition 3.7. $\square$

We just proved that in an arbitrary metric space the Hopf-Lax formula produces subsolutions of the Hamilton-Jacobi equation. Our aim now is to prove that if (X,d) is a geodesic space, then the same formula provides also supersolutions.

Theorem 3.3 (Supersolution of HJ) *Assume that* (X, d) *is a geodesic space. Then equality holds in* (17). *In particular, for all* $x \in X$ *it holds*

$$\frac{\mathrm{d}}{\mathrm{d}t} Q_t f(x) + \frac{1}{2} |DQ_t f|^2(x) = 0,$$

with at most countably many exceptions in $(0, \infty)$.

Proof Let y be a minimum of $F(t, x, \cdot)$ such that $\mathsf{d}(x, y) = D^+(x, t)$. Let $\gamma : [0, 1] \to X$ be a constant speed geodesic connecting x to y. We have

$$\begin{aligned} Q_t f(x) - Q_t f(\gamma_s) &\geq f(y) + \frac{\mathsf{d}^2(x, y)}{2t} - f(y) - \frac{\mathsf{d}^2(\gamma_s, y)}{2t} \\ &= \frac{\mathsf{d}^2(x, y) - \mathsf{d}^2(\gamma_s, y)}{2t} = \frac{(D^+(x, t))^2(2s - s^2)}{2t}. \end{aligned}$$

Therefore we obtain

$$\overline{\lim_{s \downarrow 0}} \frac{Q_t f(x) - Q_t f(\gamma_s)}{\mathsf{d}(x, \gamma_s)} = \overline{\lim_{s \downarrow 0}} \frac{Q_t f(x) - Q_t f(\gamma_s)}{s D^+(x, t)} \geq \frac{D^+(x, t)}{t}.$$

Since $s \mapsto \gamma_s$ is a particular family converging to x we deduce

$$\left| D^- Q_t f \right|(x) \geq \frac{D^+(x, t)}{t}.$$

Taking into account Proposition 3.6 and Proposition 3.7 we conclude. □

4 Weak Definitions of Gradient

In this section we introduce two weak notions of 'norm of the differential', one inspired by Cheeger's seminal paper [9], that we call minimal relaxed slope and denote by $|Df|_*$, and one inspired by the papers of Koskela-MacManus [20] and of Shanmugalingam [30], that we call minimal weak upper gradient and denote by $|Df|_w$. Notice that, as for the slopes, the objects that we are going to define are naturally in duality with the distance, thus are cotangent notion: that's why we use the 'D' instead of the '∇' in the notation. Still, we will continue speaking of upper gradients and their weak counterparts to be aligned with the convention used in the literature (see [13] for a broader discussion on this distinction between tangent and cotangent objects and its effects on calculus).

We compare our concepts with those of the original papers in Sect. 4.4, where we show that all these approaches a posteriori coincide. As usual, we will adopt the simplifying assumption that $(X, \mathsf{d}, \mathfrak{m})$ is compact and normalized metric measure space, i.e. (X, d) is compact and $\mathfrak{m} \in \mathscr{P}(X)$.

4.1 The "Vertical" Approach: Minimal Relaxed Slope

Definition 4.3 (Relaxed slopes) We say that $G \in L^2(X, \mathfrak{m})$ is a relaxed slope of $f \in L^2(X, \mathfrak{m})$ if there exist $\tilde{G} \in L^2(X, \mathfrak{m})$ and Lipschitz functions $f_n : X \to \mathbb{R}$ such that:

(a) $f_n \to f$ in $L^2(X, \mathfrak{m})$ and $|Df_n|$ weakly converges to $\tilde{G}$ in $L^2(X, \mathfrak{m})$;
(b) $\tilde{G} \le G$ $\mathfrak{m}$-a.e. in X.

We say that G is the minimal relaxed slope of f if its $L^2(X, \mathfrak{m})$ norm is minimal among relaxed slopes. We shall denote by $|Df|_*$ the minimal relaxed slope.

Using Mazur's lemma and (2a) (see Proposition 4.8) it is possible to show that an equivalent characterization of relaxed slopes can be given by modifying (a) as follows: $\tilde{G}$ is the *strong* limit in $L^2(X, \mathfrak{m})$ of $G_n \ge |Df_n|$. The definition of relaxed slope we gave is useful to show existence of relaxed slopes (as soon as an approximating sequence (f_n) with $|Df_n|$ bounded in $L^2(X, \mathfrak{m})$ exists) while the equivalent characterization is useful to perform diagonal arguments and to show that the class of relaxed slopes is a convex closed set. Therefore the definition of $|Df|_*$ is well posed.

Lemma 4.1 (Locality) *Let G_1, G_2 be relaxed slopes of f. Then $\min\{G_1, G_2\}$ is a relaxed slope as well. In particular, for any relaxed slope G it holds*

$$|Df|_* \le G \quad \mathfrak{m}\text{-a.e. in } X.$$

Proof It is sufficient to prove that if $B \subset X$ is a Borel set, then $\chi_B G_1 + \chi_{X\setminus B} G_2$ is a relaxed slope of f. By approximation, taking into account the closure of the class of relaxed slopes, we can assume with no loss of generality that B is an open set. We fix $r > 0$ and a Lipschitz function $\phi_r : X \to [0, 1]$ equal to 0 on $X \setminus B_r$ and equal to 1 on B_{2r}, where the open sets $B_s \subset B$ are defined by

$$B_s := \big\{x \in X : \mathrm{dist}(x, X \setminus B) > s\big\} \subset B.$$

Let now $f_{n,i}, i = 1, 2$, be Lipschitz and L^2 functions converging to f in $L^2(X, \mathfrak{m})$ as $n \to \infty$, with $|Df_{n,i}|$ weakly convergent to G_i and set $f_n := \phi_r f_{n,1} + (1 - \phi_r) f_{n,2}$. Then, $|Df_n| = |Df_{n,1}|$ on B_{2r} and $|Df_n| = |Df_{n,2}|$ on $X \setminus \overline{B_r}$; in $\overline{B_r} \setminus B_{2r}$, by applying (2a) and (2b), we can estimate

$$|Df_n| \le |Df_{n,2}| + \mathrm{Lip}(\phi_r)|f_{n,1} - f_{n,2}| + \phi_r\big(|Df_{n,1}| + |Df_{n,2}|\big).$$

Since $\overline{B_r} \subset B$, by taking weak limits of a subsequence, it follows that

$$\chi_{B_{2r}} G_1 + \chi_{X\setminus\overline{B_r}} G_2 + \chi_{B\setminus B_{2r}}(G_1 + 2G_2)$$

is a relaxed slope of f. Letting $r \downarrow 0$ gives that $\chi_B G_1 + \chi_{X\setminus B} G_2$ is a relaxed slope as well.

For the second part of the statement argue by contradiction: let G be a relaxed slope of f and assume that $B = \{G < |Df|_*\}$ is such that $\mathfrak{m}(B) > 0$. Consider the relaxed slope $G\chi_B + |Df|_*\chi_{X\setminus B}$: its L^2 norm is strictly less than the L^2 norm of $|Df|_*$, which is a contradiction. □

A trivial consequence of the definition and of the locality principle we just proved is that if $f : X \to \mathbb{R}$ is Lipschitz it holds:

$$|Df|_* \le |Df| \quad \mathfrak{m}\text{-a.e. in } X. \tag{19}$$

We also remark that it is possible to obtain the minimal relaxed slope as strong limit in L^2 of slopes of Lipschitz functions, and not only weak, as shown in the next proposition.

Proposition 4.8 (Strong approximation) *If $f \in L^2(X, \mathfrak{m})$ has a relaxed slope, there exist Lipschitz functions f_n convergent to f in $L^2(X, \mathfrak{m})$ with $|Df_n|$ convergent to $|Df|_*$ in $L^2(X, \mathfrak{m})$.*

Proof If $g_i \to f$ in L^2 and $|Dg_i|$ weakly converges to $|Df|_*$ in L^2, by Mazur's lemma we can find a sequence convex combinations

$$G_h = \sum_{i=N_h+1}^{N_{h+1}} \alpha_{h,i}|Dg_i|, \quad \text{with } \alpha_{i,h} \ge 0, \quad \sum_{i=N_h+1}^{N_{h+1}} \alpha_{h,i} = 1, \quad N_h \to \infty$$

of $|Dg_i|$ strongly convergent to $|Df|_*$ in L^2; the corresponding convex combinations of g_i, that we shall denote by f_h, still converge in L^2 to f and $|Df_h|$ is dominated by G_h. It follows that

$$\overline{\lim_{h\to\infty}} \int_X |Df_h|^2 \mathrm{d}\mathfrak{m} \le \overline{\lim_{h\to\infty}} \int_X G_h^2 \mathrm{d}\mathfrak{m} = \int_X |Df|_*^2 \mathrm{d}\mathfrak{m}.$$

This implies at once that $|Df_h|$ weakly converges to $|Df|_*$ (because any limit point in the weak topology is a relaxed slope with minimal norm) and that the convergence is strong. □

Theorem 4.4 *The Cheeger energy functional*

$$\mathsf{Ch}(f) := \frac{1}{2}\int_X |Df|_*^2 \mathrm{d}\mathfrak{m}, \tag{20}$$

set to $+\infty$ if f has no relaxed slope, is convex and lower semicontinuous in $L^2(X, \mathfrak{m})$.

Proof A simple byproduct of condition (2a) is that $\alpha F + \beta G$ is a relaxed slope of $\alpha f + \beta g$ whenever α, β are nonnegative constants and F, G are relaxed slopes of f, g respectively. Taking $F = |Df|_*$ and $G = |Dg|_*$ yields the convexity of Ch,

while lower semicontinuity follows by a simple diagonal argument based on the strong approximation property stated in Proposition 4.8. □

Proposition 4.9 (Chain rule) *If $f \in L^2(X, \mathfrak{m})$ has a relaxed slope and $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz and C^1, then $|D\phi(f)|_* = |\phi'(f)||Df|_*$ $\mathfrak{m}$-a.e. in X.*

Proof We trivially have $|D\phi(f)| \le |\phi'(f)||Df|$. If we apply this inequality to the "optimal" approximating sequence of Lipschitz functions given by Proposition 4.8 we get that $|\phi'(f)||Df|_*$ is a relaxed slope of $\phi(f)$, so that $|D\phi(f)|_* \le |\phi'(f)||Df|_*$ $\mathfrak{m}$-a.e. in X. Applying twice this inequality with $\phi(r) := -r$ we get $|Df|_* \le |D(-f)|_* \le |Df|_*$ and thus $|Df|_* = |D(-f)|_*$ $\mathfrak{m}$-a.e. in X.

Up to a simple rescaling, we can assume $|\phi'| \le 1$. Let $\psi_1(z) := z - \phi(z)$, notice that $\psi_1' \ge 0$ and thus $\mathfrak{m}$-a.e. on $f^{-1}(\{\phi' \ge 0\})$ it holds

$$|Df|_* \le \big|D\big(\phi(f)\big)\big|_* + \big|D\big(\psi_1(f)\big)\big|_* \le \phi'(f)|Df|_* + \psi_1'(f)|Df|_* = |Df|_*,$$

hence all the inequalities must be equalities, which forces $|D(\phi(f))|_* = \phi'(f)|Df|_*$ $\mathfrak{m}$-a.e. on $f^{-1}(\{\phi' \ge 0\})$. Similarly, let $\psi_2(z) = -z - \phi(z)$ and notice that $\psi_2' \le 0$, so that $\mathfrak{m}$-a.e. on $f^{-1}(\{\phi' \le 0\})$ it holds

$$\begin{aligned} |Df|_* = \big|D(-f)\big|_* &\le \big|D\big(\phi(f)\big)\big|_* + \big|D\big(\psi_2(f)\big)\big|_* \\ &\le -\phi'(f)|Df|_* - \psi_2'(f)|Df|_* = |Df|_*. \end{aligned}$$

As before we can conclude that $|D(\phi(f))|_* = -\phi'(f)|Df|_*$ $\mathfrak{m}$-a.e. on $f^{-1}(\{\phi' \le 0\})$. □

Still by approximation, it is not difficult to show that $\phi(f)$ has a relaxed slope if ϕ is Lipschitz, and that $|D\phi(f)|_* = |\phi'(f)||Df|_*$ $\mathfrak{m}$-a.e. in X. In this case $\phi'(f)$ is undefined at points x such that ϕ is not differentiable at $f(x)$, on the other hand the formula still makes sense because $|Df|_* = 0$ $\mathfrak{m}$-a.e. on $f^{-1}(N)$ for any Lebesgue negligible set $N \subset \mathbb{R}$. Particularly useful is the case when ϕ is a truncation function, for instance $\phi(z) = \min\{z, M\}$. In this case

$$\big|D\min\{f, M\}\big|_* = \begin{cases} |Df|_* & \text{if } f(x) < M \\ 0 & \text{if } f(x) \ge M. \end{cases}$$

Analogous formulas hold for truncations from below.

Laplacian: Definition and Basic Properties

Since the domain of Ch is dense in $L^2(X, \mathfrak{m})$ (it includes Lipschitz functions), the Hilbertian theory of gradient flows (see for instance [3, 8]) can be applied to

Cheeger's functional (20) to provide, for all $f_0 \in L^2(X, \mathfrak{m})$, a locally Lipschitz continuous map $t \mapsto f_t$ from $(0, \infty)$ to $L^2(X, \mathfrak{m})$, with $f_t \to f_0$ as $t \downarrow 0$, whose derivative satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t \in -\partial \mathsf{Ch}(f_t) \quad \text{for a.e. } t. \tag{21}$$

Here $\partial \mathsf{Ch}(g)$ denotes the subdifferential of Ch at $g \in D(\mathsf{Ch})$ in the sense of convex analysis, i.e.

$$\partial \mathsf{Ch}(g) := \left\{ \xi \in L^2(X, \mathfrak{m}) : \mathsf{Ch}(f) \geq \mathsf{Ch}(g) + \int_X \xi(f - g) \mathrm{d}\mathfrak{m} \ \forall f \in L^2(X, \mathfrak{m}) \right\}.$$

Another important regularizing effect of gradient flows of convex l.s.c. functionals lies in the fact that for every $t > 0$ (the opposite of) the right derivative $-\frac{\mathrm{d}}{\mathrm{d}t_+} f_t = \lim_{h \downarrow 0} \frac{1}{h}(f_t - f_{t+h})$ exists and it is actually the element with minimal $L^2(X, \mathfrak{m})$ norm in $\partial \mathsf{Ch}(f_t)$. This motivates the next definition:

Definition 4.4 (Laplacian) The Laplacian Δf of $f \in L^2(X, \mathfrak{m})$ is defined for those f such that $\partial \mathsf{Ch}(f) \neq \emptyset$. For those f, $-\Delta f$ is the element of minimal $L^2(X, \mathfrak{m})$ norm in $\partial \mathsf{Ch}(f)$. The domain of Δ is defined as $D(\Delta)$.

Remark 4.2 (Potential lack of linearity) It should be observed that in general the Laplacian—as we just defined it—is *not* a linear operator: the potential lack of linearity is strictly related to the fact that potentially the space $W^{1,2}(X, \mathsf{d}, \mathfrak{m})$ is not Hilbert, because $f \mapsto \int |Df|_*^2 \mathrm{d}\mathfrak{m}$ need not be quadratic. For instance if $X = \mathbb{R}^2$, $\mathfrak{m}$ is the Lebesgue measure and d is the distance induced by the L^∞ norm, then it is easily seen that

$$|Df|_*^2 = \left(\left| \frac{\partial f}{\partial x} \right| + \left| \frac{\partial f}{\partial y} \right| \right)^2.$$

Even though the Laplacian is not linear, the trivial implication

$$v \in \partial \mathsf{Ch}(f) \quad \Rightarrow \quad \lambda v \in \partial \mathsf{Ch}(\lambda f), \quad \forall \lambda \in \mathbb{R},$$

ensures that the Laplacian (and so the gradient flow of Ch) is 1-homogeneous.

We can now write

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t = \Delta f_t$$

for gradient flows f_t of Ch, the derivative being understood in $L^2(X, \mathfrak{m})$, in accordance with the classical case. The classical Hilbertian theory of gradient flows also ensures that

$$\lim_{t \to \infty} \mathsf{Ch}(f_t) = 0 \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}t} \mathsf{Ch}(f_t) = -\|\Delta f_t\|^2_{L^2(X, \mathfrak{m})}, \quad \text{for a.e. } t \in (0, \infty). \tag{22}$$

Proposition 4.10 (Integration by parts) *For all $f \in D(\Delta)$, $g \in D(\mathsf{Ch})$ it holds*

$$\left| \int_X g \Delta f \, \mathrm{d}\mathfrak{m} \right| \leq \int_X |Dg|_* |Df|_* \mathrm{d}\mathfrak{m}. \tag{23}$$

Also, let $f \in D(\Delta)$ and $\phi \in C^1(\mathbb{R})$ with bounded derivative on an interval containing the image of f. Then

$$\int_X \phi(f) \Delta f \, \mathrm{d}\mathfrak{m} = -\int_X |Df|_*^2 \phi'(f) \mathrm{d}\mathfrak{m}. \tag{24}$$

Proof Since $-\Delta f \in \partial \mathsf{Ch}(f)$ it holds

$$\mathsf{Ch}(f) - \int_X \varepsilon g \Delta f \, \mathrm{d}\mathfrak{m} \leq \mathsf{Ch}(f + \varepsilon g), \quad \forall g \in L^2(X, \mathfrak{m}),\ \varepsilon \in \mathbb{R}.$$

For $\varepsilon > 0$, $|Df|_* + \varepsilon |Dg|_*$ is a relaxed slope of $f + \varepsilon g$ (possibly not minimal). Thus it holds $2\mathsf{Ch}(f + \varepsilon g) \leq \int_X (|Df|_* + \varepsilon |Dg|_*)^2 \mathrm{d}\mathfrak{m}$ and therefore

$$\begin{aligned} -\int_X \varepsilon g \Delta f \, \mathrm{d}\mathfrak{m} &\leq \frac{1}{2} \int_X \big(|Df|_* + \varepsilon |Dg|_*\big)^2 - |Df|_*^2 \mathrm{d}\mathfrak{m} \\ &= \varepsilon \int_X |Df|_* |Dg|_* \mathrm{d}\mathfrak{m} + o(\varepsilon). \end{aligned}$$

Dividing by ε, letting $\varepsilon \downarrow 0$ and then repeating the argument with $-g$ in place of g we get (23).

For the second part we recall that, by the chain rule, $|D(f + \varepsilon\phi(f))|_* = (1 + \varepsilon\phi'(f))|Df|_*$ for $|\varepsilon|$ small enough. Hence

$$\begin{aligned} \mathsf{Ch}\big(f + \varepsilon\phi(f)\big) - \mathsf{Ch}(f) &= \frac{1}{2} \int_X |Df|_*^2 \Big(\big(1 + \varepsilon\phi'(f)\big)^2 - 1\Big) \mathrm{d}\mathfrak{m} \\ &= \varepsilon \int_X |Df|_*^2 \phi'(f) \mathrm{d}\mathfrak{m} + o(\varepsilon), \end{aligned}$$

which implies that for any $v \in \partial \mathsf{Ch}(f)$ it holds $\int_X v\phi(f) \mathrm{d}\mathfrak{m} = \int_X |Df|_*^2 \phi'(f) \mathrm{d}\mathfrak{m}$, and gives the thesis with $v = -\Delta f$. □

Proposition 4.11 (Some properties of the gradient flow of Ch) *Let $f_0 \in L^2(X, \mathfrak{m})$ and let (f_t) be the gradient flow of Ch starting from f_0. Then the following properties hold.*

Mass preservation. $\int f_t \mathrm{d}\mathfrak{m} = \int f_0 \mathrm{d}\mathfrak{m}$ *for any $t \geq 0$.*

Maximum principle. *If $f_0 \leq C$ (resp. $f_0 \geq c$) $\mathfrak{m}$-a.e. in X, then $f_t \leq C$ (resp $f_t \geq c$) $\mathfrak{m}$-a.e. in X for any $t \geq 0$.*

Entropy dissipation. *Suppose* $0 < c \le f_0 \le C < \infty$ $\mathfrak{m}$*-a.e. Then* $t \mapsto \int f_t \log f_t \mathrm{d}\mathfrak{m}$ *is absolutely continuous in* $[0, \infty)$ *and it holds*

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_X f_t \log f_t \mathrm{d}\mathfrak{m} = -\int_X \frac{|Df_t|_*^2}{f_t} \mathrm{d}\mathfrak{m}, \quad \text{for a.e. } t \in (0, \infty).$$

Proof Mass preservation. Just notice that from (23) we get

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} \int_X f_t \mathrm{d}\mathfrak{m} \right| = \left| \int_X \mathbf{1} \cdot \Delta f_t \mathrm{d}\mathfrak{m} \right| \le \int_X |D\mathbf{1}|_* |Df_t|_* \mathrm{d}\mathfrak{m} = 0, \quad \text{for a.e. } t \in (0, \infty),$$

where $\mathbf{1}$ is the function identically equal to 1, which has minimal relaxed gradient equal to 0.

Maximum principle. Fix $f \in L^2(X, \mathfrak{m})$, $\tau > 0$ and, according to the implicit Euler scheme, let f^τ be the unique minimizer of

$$g \mapsto \mathsf{Ch}(g) + \frac{1}{2\tau} \int_X |g - f|^2 \mathrm{d}\mathfrak{m}.$$

Assume that $f \le C$. We claim that in this case $f^\tau \le C$ as well. Indeed, if this is not the case we can consider the competitor $g := \min\{f^\tau, C\}$ in the above minimization problem. By Proposition 4.9 (with the choice $\phi(r) := \min\{r, C\}$) we get $\mathsf{Ch}(g) \le \mathsf{Ch}(f^\tau)$ and the L^2 distance of f and g is strictly smaller than the one of f and f^τ as soon as $\mathfrak{m}(\{f^\tau > C\}) > 0$, which is a contradiction.

Starting from f_0, iterating this procedure, and using the fact that the implicit Euler scheme converges as $\tau \downarrow 0$ (see [3, 8] for details) to the gradient flow we get the conclusion.

The same arguments applies to uniform bounds from below.

Entropy dissipation. The map $z \mapsto z \log z$ is Lipschitz on $[c, C]$ which, together with the maximum principle and the fact that $t \mapsto f_t \in L^2(X, \mathfrak{m})$ is locally absolutely continuous, yields the claimed absolute continuity statement. Now notice that we have $\frac{\mathrm{d}}{\mathrm{d}t} \int f_t \log f_t \mathrm{d}\mathfrak{m} = \int (\log f_t + 1) \Delta f_t \mathrm{d}\mathfrak{m}$ for a.e. t. Since by the maximum principle $f_t \ge c$ $\mathfrak{m}$-a.e., the function $\log z + 1$ is Lipschitz and C^1 on the image of f_t for any $t \ge 0$, thus from (24) we get the conclusion. □

4.2 The "Horizontal" Approach: Weak Upper Gradients

In this subsection, following the approach of [4, 5], we introduce a different notion of "weak norm of gradient" in a compact and normalized metric measure space $(X, \mathsf{d}, \mathfrak{m})$. This notion of gradient is Lagrangian in spirit, it does not require a relaxation procedure, it will provide a new estimate of entropy dissipation along the gradient flow of Ch, and it will also be useful in the analysis of the derivative of the entropy along Wasserstein geodesics.

While the definition of minimal relaxed slope was taken from Cheeger's work [9], the notion we are going to introduce is inspired by the work of Koskela-MacManus

[20] and Shanmugalingam [30], the only difference being that we consider a different notion of null set of curves.

Negligible Sets of Curves and Functions Sobolev Along a.e. Curve

Recall that the evaluation maps $\mathrm{e}_t : C([0,1],X) \to X$ are defined by $\mathrm{e}_t(\gamma) := \gamma_t$. We also introduce the restriction maps $\mathrm{restr}_t^s : C([0,1],X) \to C([0,1],X)$, $0 \le t \le s \le 1$, given by

$$\mathrm{restr}_t^s(\gamma)_r := \gamma_{((1-r)t+rs)}, \tag{25}$$

so that restr_t^s restricts the curve γ to the interval $[t,s]$ and then "stretches" it on the whole of $[0,1]$.

Definition 4.5 (Test plans and negligible sets of curves) We say that a probability measure $\boldsymbol{\pi} \in \mathscr{P}(C([0,1],X))$ is a *test plan* if it is concentrated on $AC^2([0,1];X)$, $\int\int_0^1 |\dot\gamma_t|^2 \mathrm{d}t\mathrm{d}\boldsymbol{\pi} < \infty$, and there exists a constant $C(\boldsymbol{\pi})$ such that

$$(\mathrm{e}_t)_\sharp \boldsymbol{\pi} \le C(\boldsymbol{\pi})\mathfrak{m} \quad \text{for every } t \in [0,1]. \tag{26}$$

A Borel set $A \subset AC^2([0,1],X)$ is said *negligible* if for any test plan $\boldsymbol{\pi}$ there exists a $\boldsymbol{\pi}$-negligible set N such that $A \subset N$. A property which holds for every $\gamma \in AC^2([0,1],X)$, except possibly a negligible set, is said to hold for almost every curve.

Remark 4.3 An easy consequence of condition (26) is that if two $\mathfrak{m}$-measurable functions $f, g : X \to \mathbb{R}$ coincide up to a $\mathfrak{m}$-negligible set and $\mathcal{T}$ is an at most countable subset of $[0,1]$, then the functions $f \circ \gamma$ and $g \circ \gamma$ coincide in $\mathcal{T}$ for almost every curve γ.

Moreover, choosing an arbitrary test plan $\boldsymbol{\pi}$ and applying Fubini's Theorem to the product measure $\mathscr{L}^1 \times \boldsymbol{\pi}$ in $(0,1) \times C([0,1];X)$ we also obtain that $f \circ \gamma = g \circ \gamma$ $\mathscr{L}^1$-a.e. in $(0,1)$ for $\boldsymbol{\pi}$-a.e. curve γ; since $\boldsymbol{\pi}$ is arbitrary, the same property holds for almost every curve.

Coupled with the definition of negligible set of curves, there are the definitions of weak upper gradient and of functions which are Sobolev along a.e. curve.

Definition 4.6 (Weak upper gradients) A Borel function $g : X \to [0,\infty]$ is a weak upper gradient of $f : X \to \mathbb{R}$ if

$$\left| \int_{\partial\gamma} f \right| \le \int_\gamma g < \infty \quad \text{for a.e. } \gamma. \tag{27}$$

Definition 4.7 (Sobolev functions along a.e. curve) A function $f : X \to \mathbb{R}$ is Sobolev along a.e. curve if for a.e. curve γ the function $f \circ \gamma$ coincides a.e. in $[0,1]$ and in $\{0,1\}$ with an absolutely continuous map $f_\gamma : [0,1] \to \mathbb{R}$.

By Remark 4.3 applied to $\mathcal{T} := \{0, 1\}$, (27) does not depend on the particular representative of f in the class of $\mathfrak{m}$-measurable function coinciding with f up to a $\mathfrak{m}$-negligible set. The same Remark also shows that the property of being Sobolev along almost every curve γ is independent of the representative in the class of $\mathfrak{m}$-measurable functions coinciding with f $\mathfrak{m}$-a.e. in X.

In the following remarks we will make use of this basic calculus lemma:

Lemma 4.2 *Let $f : (0, 1) \to \mathbb{R}$ Lebesgue measurable, $q \in [1, \infty]$, $g \in L^q(0, 1)$ nonnegative be satisfying*

$$\big|f(s) - f(t)\big| \le \left| \int_s^t g(r)\mathrm{d}r \right| \quad \text{for } \mathscr{L}^2\text{-a.e. } (s, t) \in (0, 1)^2.$$

Then $f \in W^{1,q}(0, 1)$ and $|f'| \le g$ a.e. in $(0, 1)$.

Proof We start by proving the Lemma in the case $q = 1$. It is immediate to check that $f \in L^\infty(0, 1)$. Let $N \subset (0, 1)^2$ be the $\mathscr{L}^2$-negligible subset where the above inequality fails. By Fubini's theorem, also the set $\{(t, h) \in (0, 1)^2 : (t, t + h) \in N \cap (0, 1)^2\}$ is $\mathscr{L}^2$-negligible. In particular, by Fubini's theorem, for a.e. h we have $(t, t + h) \notin N$ for a.e. $t \in (0, 1)$. Let $h_i \downarrow 0$ with this property and use the identities

$$\int_0^1 f(t)\frac{\phi(t + h) - \phi(t)}{h}\mathrm{d}t = -\int_0^1 \frac{f(t - h) - f(t)}{-h}\phi(t)\mathrm{d}t$$

with $\phi \in C_c^1(0, 1)$ and $h = h_i$ sufficiently small to get

$$\left| \int_0^1 f(t)\phi'(t)\mathrm{d}t \right| \le \int_0^1 g(t)\big|\phi(t)\big|\mathrm{d}t.$$

It follows that the distributional derivative of f is a signed measure η with finite total variation which satisfies

$$-\int_0^1 f\phi'\mathrm{d}t = \int_0^1 \phi\mathrm{d}\eta, \qquad \left| \int_0^1 \phi\mathrm{d}\eta \right| \le \int_0^1 g|\phi|\mathrm{d}t \quad \text{for every } \phi \in C_c^1(0, 1);$$

therefore η is absolutely continuous with respect to the Lebesgue measure with $|\eta| \le g\mathscr{L}^1$. This gives the $W^{1,1}(0, 1)$ regularity and, at the same time, the inequality $|f'| \le g$ a.e. in $(0, 1)$. The case $q > 1$ immediately follows by applying this inequality when $g \in L^q(0, 1)$. □

With the aid of this lemma, we can prove that the existence of a weak upper gradient implies Sobolev regularity along a.e. curve.

Remark 4.4 (Restriction and equivalent formulation) Notice that if $\boldsymbol{\pi}$ is a test plan, so is $(\mathrm{restr}_t^s)_\sharp\boldsymbol{\pi}$. Hence if g is a weak upper gradient of f then for every $t < s$ in

[0, 1] it holds

$$\big|f(\gamma_s) - f(\gamma_t)\big| \le \int_t^s g(\gamma_r)|\dot\gamma_r|\mathrm{d}r \quad \text{for a.e. } \gamma.$$

Let $\boldsymbol{\pi}$ be a test plan: by Fubini's theorem applied to the product measure $\mathscr{L}^2 \times \boldsymbol{\pi}$ in $(0,1)^2 \times C([0,1]; X)$, it follows that for $\boldsymbol{\pi}$-a.e. γ the function f satisfies

$$\big|f(\gamma_s) - f(\gamma_t)\big| \le \left|\int_t^s g(\gamma_r)|\dot\gamma_r|\mathrm{d}r\right| \quad \text{for } \mathscr{L}^2\text{-a.e. } (t,s) \in (0,1)^2.$$

An analogous argument shows that

$$\begin{cases} |f(\gamma_s) - f(\gamma_0)| \le \int_0^s g(\gamma_r)|\dot\gamma_r|\mathrm{d}r \\ |f(\gamma_1) - f(\gamma_s)| \le \int_s^1 g(\gamma_r)|\dot\gamma_r|\mathrm{d}r \end{cases} \quad \text{for } \mathscr{L}^1\text{-a.e. } s \in (0,1). \tag{28}$$

Since $g \circ \gamma|\dot\gamma| \in L^1(0,1)$ for $\boldsymbol{\pi}$-a.e. γ, by Lemma 4.2 it follows that $f \circ \gamma \in W^{1,1}(0,1)$ for $\boldsymbol{\pi}$-a.e. γ, and

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)\right| \le g \circ \gamma|\dot\gamma| \quad \text{a.e. in } (0,1), \text{ for } \boldsymbol{\pi}\text{-a.e. } \gamma. \tag{29}$$

Since $\boldsymbol{\pi}$ is arbitrary, we conclude that $f \circ \gamma \in W^{1,1}(0,1)$ for a.e. γ, and therefore it admits an absolutely continuous representative f_γ; moreover, by (28), it is immediate to check that $f(\gamma(t)) = f_\gamma(t)$ for $t \in \{0,1\}$ and a.e. γ.

Remark 4.5 (An approach with a non explicit use of negligible set of curves) The previous remark could be used to introduce the notion of weak upper gradients without speaking (explicitly) of Borel sets at all. One can simply say that $g \in L^2(X, \mathfrak{m})$ is a weak upper gradient of $f : X \to \mathbb{R}$ provided for every test plan $\boldsymbol{\pi}$ it holds

$$\int \big|f(\gamma_1) - f(\gamma_0)\big|\mathrm{d}\boldsymbol{\pi}(\gamma) \le \iint_0^1 g(\gamma_s)|\dot\gamma_s|\mathrm{d}s\mathrm{d}\boldsymbol{\pi}(\gamma)$$

(this has been the approach followed in [13]).

Proposition 4.12 (Locality) *Let $f : X \to \mathbb{R}$ be Sobolev along almost all absolutely continuous curves, and let G_1, G_2 be weak upper gradients of f. Then* $\min\{G_1, G_2\}$ *is a weak upper gradient of f.*

Proof It is a direct consequence of (29). □

Definition 4.8 (Minimal weak upper gradient) Let $f : X \to \mathbb{R}$ be Sobolev along almost all curves. The minimal weak upper gradient $|Df|_w$ of f is the weak upper gradient characterized, up to $\mathfrak{m}$-negligible sets, by the property

$$|Df|_w \le G \quad \mathfrak{m}\text{-a.e. in } X, \text{ for every weak upper gradient } G \text{ of } f. \tag{30}$$

Uniqueness of the minimal weak upper gradient is obvious. For existence, we take $|Df|_w := \inf_n G_n$, where G_n are weak upper gradients which provide a minimizing sequence in

$$\inf\left\{ \int_X \tan^{-1} G \mathrm{d}\mathfrak{m} : G \text{ is a weak upper gradient of } f \right\}.$$

We immediately see, thanks to Proposition 4.12, that we can assume with no loss of generality that $G_{n+1} \le G_n$. Hence, by monotone convergence, the function $|Df|_w$ is a weak upper gradient of f and $\int_X \tan^{-1} G \mathrm{d}\mathfrak{m}$ is minimal at $G = |Df|_w$. This minimality, in conjunction with Proposition 4.12, gives (30).

Theorem 4.5 (Stability w.r.t. $\mathfrak{m}$-a.e. convergence) *Assume that f_n are $\mathfrak{m}$-measurable, Sobolev along almost all curves and that G_n are weak upper gradients of f_n. Assume furthermore that $f_n(x) \to f(x) \in \mathbb{R}$ for $\mathfrak{m}$-a.e. $x \in X$ and that (G_n) weakly converges to G in $L^2(X, \mathfrak{m})$. Then G is a weak upper gradient of f.*

Proof Fix a test plan $\boldsymbol{\pi}$. By Mazur's theorem we can find convex combinations

$$H_h := \sum_{i=N_h+1}^{N_{h+1}} \alpha_{h,i} G_i \quad \text{with } \alpha_{h,i} \ge 0, \quad \sum_{i=N_h+1}^{N_{h+1}} \alpha_{h,i} = 1, \quad N_h \to \infty$$

converging strongly to G in $L^2(X, \mathfrak{m})$. Denoting by $\tilde{f}_h$ the corresponding convex combinations of f_h, H_h are weak upper gradients of $\tilde{f}_h$ and still $\tilde{f}_h \to f$ $\mathfrak{m}$-a.e. in X.

Since for every nonnegative Borel function $\varphi : X \to [0, \infty]$ it holds (with $C = C(\boldsymbol{\pi})$)

$$\begin{aligned}
\int \left(\int_\gamma \varphi \right) \mathrm{d}\boldsymbol{\pi} &= \int \left(\int_0^1 \varphi(\gamma_t) |\dot{\gamma}_t| \mathrm{d}t \right) \mathrm{d}\boldsymbol{\pi} \\
&\le \int \left(\int_0^1 \varphi^2(\gamma_t) \mathrm{d}t \right)^{1/2} \left(\int_0^1 |\dot{\gamma}_t|^2 \mathrm{d}t \right)^{1/2} \mathrm{d}\boldsymbol{\pi} \\
&\le \left(\int_0^1 \int \varphi^2 \mathrm{d}(\mathrm{e}_t)_\sharp \boldsymbol{\pi} \mathrm{d}t \right)^{1/2} \left(\iint_0^1 |\dot{\gamma}_t|^2 \mathrm{d}t \mathrm{d}\boldsymbol{\pi} \right)^{1/2} \\
&\le \left(C \int \varphi^2 \mathrm{d}\mathfrak{m} \right)^{1/2} \left(\iint_0^1 |\dot{\gamma}_t|^2 \mathrm{d}t \mathrm{d}\boldsymbol{\pi} \right)^{1/2}, \qquad (31)
\end{aligned}$$

we obtain, for $\bar{C} := \sqrt{C} (\iint_0^1 |\dot{\gamma}_t|^2 \mathrm{d}t \mathrm{d}\boldsymbol{\pi})^{1/2}$,

$$\begin{aligned}
&\int \left(\int_\gamma |H_h - G| + \min\{|\tilde{f}_h - f|, 1\} \right) \mathrm{d}\boldsymbol{\pi} \\
&\quad \le \bar{C} \left(\|H_h - G\|_{L^2} + \left\| \min\{|\tilde{f}_h - f|, 1\} \right\|_{L^2} \right) \to 0.
\end{aligned}$$

By a diagonal argument we can find a subsequence $h(n)$ such that $\int_\gamma |H_{h(n)} - G| + \min\{|\tilde f_{h(n)} - f|, 1\} \to 0$ as $n \to \infty$ for $\boldsymbol{\pi}$-a.e. γ. Since $\tilde f_h$ converge $\mathfrak{m}$-a.e. to f and the marginals of $\boldsymbol{\pi}$ are absolutely continuous w.r.t. $\mathfrak{m}$ we have also that for $\boldsymbol{\pi}$-a.e. γ it holds $\tilde f_h(\gamma_0) \to f(\gamma_0)$ and $\tilde f_h(\gamma_1) \to f(\gamma_1)$.

If we fix a curve γ satisfying these convergence properties, since $(\tilde f_{h(n)})_\gamma$ are equi-absolutely continuous (being their derivatives bounded by $H_{h(n)} \circ \gamma |\dot\gamma|$) and a further subsequence of $\tilde f_{h(n)}$ converges a.e. in $[0,1]$ and in $\{0,1\}$ to $f(\gamma_s)$, we can pass to the limit to obtain an absolutely continuous function f_γ equal to $f(\gamma_s)$ a.e. in $[0,1]$ and in $\{0,1\}$ with derivative bounded by $G(\gamma_s)|\dot\gamma_s|$. Since $\boldsymbol{\pi}$ is arbitrary we conclude that f is Sobolev along almost all curves and that G is a weak upper gradient of f. □

Remark 4.6 ($|Df|_w \le |Df|_*$) An immediate consequence of the previous proposition is that any $f \in D(\mathsf{Ch})$ is Sobolev along a.e. curve and satisfies $|Df|_w \le |Df|_*$. Indeed, for such f just pick a sequence of Lipschitz functions converging to f in $L^2(X,\mathfrak{m})$ such that $|Df_n| \to |Df|_*$ in $L^2(X,\mathfrak{m})$ (as in Proposition 4.8) and recall that for Lipschitz functions the local Lipschitz constant is an upper gradient.

A Bound from Below on Weak Gradients

In this short subsection we show how, using test plans and the very definition of minimal weak gradients, it is possible to use $|Df|_w$ to bound from below the increments of the relative entropy. We start with the following result, proved—in a more general setting—by Lisini in [22]: it shows how to associate to a curve $\mu \in AC^2([0,1];(\mathscr{P}(X),W_2))$ a plan $\boldsymbol{\pi} \in \mathscr{P}(C([0,1],X))$ concentrated on $AC^2([0,1];X)$ representing the curve itself (see also Theorem 8.2.1 of [3] for the Euclidean case). We will only sketch the proof.

Proposition 4.13 (Superposition principle) *Let (X,d) be a compact space and let $\mu \in AC^2([0,1];(\mathscr{P}(X),W_2))$. Then there exists $\boldsymbol{\pi} \in \mathscr{P}(C([0,1],X))$ concentrated on $AC^2([0,1];X)$ such that $(\mathrm{e}_t)_\sharp\boldsymbol{\pi} = \mu_t$ for any $t \in [0,1]$ and $\int |\dot\gamma_t|^2 \mathrm{d}\boldsymbol{\pi}(\gamma) = |\dot\mu_t|^2$ for a.e. $t \in [0,1]$.*

Proof If $\boldsymbol{\pi} \in \mathscr{P}(C([0,1],X))$ is any plan concentrated on $AC^2([0,1],X)$ such that $(\mathrm{e}_t)_\sharp\boldsymbol{\pi} = \mu_t$ for any $t \in [0,1]$, since $(\mathrm{e}_t,\mathrm{e}_s)_\sharp\boldsymbol{\pi} \in \mathrm{ADM}(\mu_t,\mu_s)$, for any $t < s$ it holds

$$W_2^2(\mu_t,\mu_s) \le \int \mathsf{d}^2(\gamma_t,\gamma_s)\mathrm{d}\boldsymbol{\pi}(\gamma) \le \int \left(\int_t^s |\dot\gamma_r|\mathrm{d}r\right)^2 \mathrm{d}\boldsymbol{\pi}(\gamma)$$

$$\le (s-t)\iint_t^s |\dot\gamma_r|^2 \mathrm{d}r\mathrm{d}\boldsymbol{\pi}(\gamma),$$

which shows that $|\dot\mu_t|^2 \le \int |\dot\gamma_t|^2 \mathrm{d}\boldsymbol{\pi}(\gamma)$ for a.e. t. Hence, to conclude it is sufficient to find a plan $\boldsymbol{\pi} \in \mathscr{P}(C([0,1],X))$, concentrated on $AC^2([0,1],X)$, with $(\mathrm{e}_t)_\sharp\boldsymbol{\pi} = \mu_t$ for any $t \in [0,1]$ such that $\int |\dot\mu_t|^2 \mathrm{d}t \ge \iint_0^1 |\dot\gamma_t|^2 \mathrm{d}t\mathrm{d}\boldsymbol{\pi}(\gamma)$.

To build such a $\boldsymbol{\pi}$ we make the simplifying assumption that (X, d) is geodesic (the proof for the general case is similar, but rather than interpolating with piecewise geodesic curves one uses piecewise constant ones, this leads to some technical complications that we want to avoid here—see [22] for the complete argument). Fix $n \in \mathbb{N}$ and use a gluing argument to find $\boldsymbol{\gamma}^n \in \mathscr{P}(X^{n+1})$ such that $(\pi^i, \pi^{i+1})_\sharp \boldsymbol{\gamma}^n \in \mathrm{OPT}(\mu_{\frac{i}{n}}, \mu_{\frac{i+1}{n}})$ for $i = 0, \dots, n-1$. By standard measurable selection arguments, there exists a Borel map $T^n : X^{n+1} \to C([0,1], X)$ such that $\gamma := T^n(x_0, \dots, x_n)$ is a constant speed geodesic on each of the intervals $[i/n, (i+1)/n]$ and $\gamma_{i/n} = x_i$, $i = 0, \dots, n$. Define $\boldsymbol{\pi}^n := T^n_\sharp \boldsymbol{\gamma}^n$. It holds

$$\begin{aligned}\iint_0^1 |\dot\gamma_t|^2 \mathrm{d}t \mathrm{d}\boldsymbol{\pi}^n(\gamma) &= \frac{1}{n} \int \sum_{i=0}^{n-1} \mathsf{d}^2(\gamma_{\frac{i}{n}}, \gamma_{\frac{i+1}{n}}) \mathrm{d}\boldsymbol{\pi}(\gamma) = \frac{1}{n} \sum_{i=0}^{n-1} W_2^2(\mu_{\frac{i}{n}}, \mu_{\frac{i+1}{n}}) \\ &\le \int_0^1 |\dot\mu_t|^2 \mathrm{d}t. \end{aligned} \tag{32}$$

Now notice that the map $E : C([0,1], X) \to [0, \infty]$ given by $E(\gamma) := \int_0^1 |\dot\gamma_t|^2 \mathrm{d}t$ if $\gamma \in AC^2([0,1], X)$ and $+\infty$ otherwise, is lower semicontinuous and, via a simple equicontinuity argument, with compact sublevels. Therefore by Prokorov's theorem we get that $(\boldsymbol{\pi}^n) \subset \mathscr{P}(C([0,1], X))$ is a tight sequence, hence for any limit measure $\boldsymbol{\pi}$ the uniform bound (32) gives the thesis. □

Proposition 4.14 *Let $[0,1] \ni t \mapsto \mu_t = f_t \mathfrak{m}$ be a curve in $AC^2([0,1], (\mathscr{P}(X), W_2))$. Assume that for some $0 < c < C < \infty$ it holds $c \le f_t \le C$ $\mathfrak{m}$-a.e. for any $t \in [0,1]$, and that f_0 is Sobolev along a.e. curve with $|Df_0|_w \in L^2(X, \mathfrak{m})$. Then*

$$\int_X f_0 \log f_0 \mathrm{d}\mathfrak{m} - \int_X f_t \log f_t \mathrm{d}\mathfrak{m} \le \frac{1}{2} \int_0^t \int_X \frac{|Df_0|_w^2}{f_0^2} f_s \mathrm{d}s \mathrm{d}\mathfrak{m} + \frac{1}{2} \int_0^t |\dot\mu_s|^2 \mathrm{d}s,$$

$\forall t > 0.$

Proof Let $\boldsymbol{\pi} \in \mathscr{P}(C([0,1], X))$ be a plan associated to the curve (μ_t) as in Proposition 4.13. The assumption $f_t \le C$ $\mathfrak{m}$-a.e. and the fact that $\iint_0^1 |\dot\gamma_t|^2 \mathrm{d}t \mathrm{d}\boldsymbol{\pi}(\gamma) = \int |\dot\mu_t|^2 \mathrm{d}t < \infty$ guarantee that $\boldsymbol{\pi}$ is a test plan. Now notice that it holds $|D \log f_t|_w = |Df_t|_w / f_t$ (because $z \mapsto \log z$ is C^1 in $[c, C]$), thus we get

$$\begin{aligned} &\int_X f_0 \log f_0 \mathrm{d}\mathfrak{m} - \int_X f_t \log f_t \mathrm{d}\mathfrak{m} \\ &\quad \le \int_X \log f_0 (f_0 - f_t) \mathrm{d}\mathfrak{m} \\ &\quad = \int (\log f_0 \circ \mathrm{e}_0 - \log f_0 \circ \mathrm{e}_t) \mathrm{d}\boldsymbol{\pi} \\ &\quad \le \iint_0^t \frac{|Df_0|_w(\gamma_s)}{f_0(\gamma_s)} |\dot\gamma_s| \mathrm{d}s \mathrm{d}\boldsymbol{\pi}(\gamma) \end{aligned}$$

$$\leq \frac{1}{2}\int\int_0^t \frac{|Df_0|_w^2(\gamma_s)}{f_0^2(\gamma_s)}\mathrm{d}s\mathrm{d}\boldsymbol{\pi}(\gamma) + \frac{1}{2}\int\int_0^t |\dot{\gamma}_s|^2\mathrm{d}s\mathrm{d}\boldsymbol{\pi}(\gamma)$$

$$= \frac{1}{2}\int_0^t \int_X \frac{|Df_0|_w^2}{f_0^2} f_s\mathrm{d}s\mathrm{d}\mathfrak{m} + \frac{1}{2}\int_0^t |\dot{\mu}_s|^2\mathrm{d}s. \qquad \square$$

4.3 The Two Notions of Gradient Coincide

Here we prove that the two notions of "norm of weak gradient" we introduced coincide. We already noticed in Remark 4.6 that $|Df|_w \leq |Df|_*$, so that to conclude we need to show that $|Df|_w \geq |Df|_*$.

The key argument to achieve this is the following lemma, which gives a sharp bound on the W_2-speed of the L^2-gradient flow of Ch. This lemma has been introduced in [15] to study the heat flow on Alexandrov spaces, see also Sect. 6.

Lemma 4.3 (Kuwada's lemma) *Let $f_0 \in L^2(X,\mathfrak{m})$ and let (f_t) be the L_2-gradient flow of* Ch *starting from f_0. Assume that for some $0 < c \leq C < \infty$ it holds $c \leq f_0 \leq C$ $\mathfrak{m}$-a.e. in X, and that $\int_X f_0\mathrm{d}\mathfrak{m} = 1$. Then the curve $t \mapsto \mu_t := f_t\mathfrak{m}$ is absolutely continuous w.r.t. W_2 and it holds*

$$|\dot{\mu}_t|^2 \leq \int_X \frac{|Df_t|_*^2}{f_t}\mathrm{d}\mathfrak{m}, \quad \text{for a.e. } t \in (0,\infty).$$

Proof We start from the duality formula (5) with $\varphi = -\psi$: taking into account the factor 2 and using the identity $Q_1(-\psi) = \psi^c$ we get

$$\frac{W_2^2(\mu,\nu)}{2} = \sup_{\varphi} \int_X Q_1\varphi d\nu - \int_X \varphi d\mu \tag{33}$$

where the supremum runs among all Lipschitz functions φ.

Fix such a φ and recall (Proposition 3.6) that the map $t \mapsto Q_t\varphi$ is Lipschitz with values in $L^\infty(X,\mathfrak{m})$, and a fortiori in $L^2(X,\mathfrak{m})$.

Fix also $0 \leq t < s$, set $\ell = (s-t)$ and recall that since (f_t) is the Gradient Flow of Ch in L^2, the map $[0,\ell] \ni \tau \mapsto f_{t+\tau}$ is absolutely continuous with values in L^2. Therefore the map $[0,\ell] \ni \tau \mapsto Q_{\frac{\tau}{\ell}}\varphi f_{t+\tau}$ is absolutely continuous with values in L^2. The equality

$$\frac{Q_{\frac{\tau+h}{\ell}}\varphi f_{t+\tau+h} - Q_{\frac{\tau}{\ell}}\varphi f_{t+\tau}}{h} = f_{t+\tau}\frac{Q_{\frac{\tau+h}{\ell}}\varphi - Q_{\frac{\tau}{\ell}}\varphi}{h} + Q_{\frac{\tau+h}{\ell}}\varphi\frac{f_{t+\tau+h} - f_{t+\tau}}{h},$$

together with the uniform continuity of $(x,\tau) \mapsto Q_{\frac{\tau}{\ell}}\varphi(x)$ shows that the derivative of $\tau \mapsto Q_{\frac{\tau}{\ell}}\varphi f_{t+\tau}$ can be computed via the Leibniz rule.

We have:

$$\begin{aligned}\int_X Q_1\varphi \mathrm{d}\mu_s - \int_X \varphi \mathrm{d}\mu_t &= \int_X Q_1\varphi f_{t+\ell}\mathrm{d}\mathfrak{m} - \int_X \varphi f_t \mathrm{d}\mathfrak{m}\\ &= \int_X \int_0^\ell \frac{\mathrm{d}}{\mathrm{d}\tau}(Q_{\frac{\tau}{\ell}}\varphi f_{t+\tau})\mathrm{d}\tau\mathrm{d}\mathfrak{m}\\ &\le \int_X \int_0^\ell \left(-\frac{|DQ_{\frac{\tau}{\ell}}\varphi|^2}{2\ell} f_{t+\tau} + Q_{\frac{\tau}{\ell}}\varphi \Delta f_{t+\tau}\right)\mathrm{d}\tau\mathrm{d}\mathfrak{m}, \end{aligned} \tag{34}$$

having used Theorem 3.2. Observe that by inequalities (23) and (19) we have

$$\begin{aligned}\int_X Q_{\frac{\tau}{\ell}}\varphi \Delta f_{t+\tau}\mathrm{d}\mathfrak{m} &\le \int_X |DQ_{\frac{\tau}{\ell}}\varphi|_* |Df_{t+\tau}|_* \mathrm{d}\mathfrak{m} \le \int_X |DQ_{\frac{\tau}{\ell}}\varphi||Df_{t+\tau}|_*\mathrm{d}\mathfrak{m}\\ &\le \frac{1}{2\ell}\int_X |DQ_{\frac{\tau}{\ell}}\varphi|^2 f_{t+\tau}\mathrm{d}\mathfrak{m} + \frac{\ell}{2}\int_X \frac{|Df_{t+\tau}|_*^2}{f_{t+\tau}}\mathrm{d}\mathfrak{m}. \end{aligned} \tag{35}$$

Plugging this inequality in (34), we obtain

$$\int_X Q_1\varphi \mathrm{d}\mu_s - \int_X \varphi \mathrm{d}\mu_t \le \frac{\ell}{2}\int_0^\ell \int_X \frac{|Df_{t+\tau}|_*^2}{f_{t+\tau}}\mathrm{d}\mathfrak{m}.$$

This latter bound does not depend on φ, so from (33) we deduce

$$W_2^2(\mu_t,\mu_s) \le \ell \int_0^\ell \int_X \frac{|Df_{t+\tau}|_*^2}{f_{t+\tau}}\mathrm{d}\mathfrak{m}.$$

Since $f_r \ge c$ for any $r \ge 0$ and $r \mapsto \mathsf{Ch}(f_r)$ is nonincreasing and finite for every $r > 0$, we immediately get that $t \mapsto \mu_t$ is locally Lipschitz in $(0,\infty)$. At Lebesgue points of $t \mapsto \int_X |Df_t|_*^2/f_t \mathrm{d}\mathfrak{m}$ we obtain the stated pointwise bound on the metric speed. □

Theorem 4.6 *Let $f \in L^2(X,\mathfrak{m})$. Assume that f is Sobolev along a.e. curve and that $|Df|_w \in L^2(X,\mathfrak{m})$. Then $f \in D(\mathsf{Ch})$ and $|Df|_* = |Df|_w$ $\mathfrak{m}$-a.e. in X.*

Proof Up to a truncation argument and addition of a constant, we can assume that $0 < c \le f \le C < \infty$ $\mathfrak{m}$-a.e. in X for some c, C. Let (f_t) be the L_2-gradient flow of Ch starting from f and recall that from Proposition 4.11 we have

$$\int_X f\log f \mathrm{d}\mathfrak{m} - \int_X f_t \log f_t \mathrm{d}\mathfrak{m} = \int_0^t \int_X \frac{|Df_s|_*^2}{f_s}\mathrm{d}s\mathrm{d}\mathfrak{m} < \infty \quad \text{for every } t > 0.$$

On the other hand, from Proposition 4.14 and Lemma 4.3 we have

$$\int_X f\log f \mathrm{d}\mathfrak{m} - \int_X f_t \log f_t \mathrm{d}\mathfrak{m} \le \frac{1}{2}\int_0^t \int_X \frac{|Df|_w^2}{f^2} f_s \mathrm{d}s\mathrm{d}\mathfrak{m} + \frac{1}{2}\int_0^t \int_X \frac{|Df_s|_*^2}{f_s}\mathrm{d}s\mathrm{d}\mathfrak{m}. \tag{36}$$

Hence we deduce

$$\int_0^t 4\mathsf{Ch}(\sqrt{f_s})\mathrm{d}s = \frac{1}{2}\int_0^t \int_X \frac{|Df_s|_*^2}{f_s}\mathrm{d}s\mathrm{d}\mathfrak{m} \le \frac{1}{2}\int_0^t \int_X \frac{|Df|_w^2}{f^2} f_s \mathrm{d}s\mathrm{d}\mathfrak{m}.$$

Letting $t \downarrow 0$, taking into account the L^2-lower semicontinuity of Ch and the fact—easy to check from the maximum principle—that $\sqrt{f_s} \to \sqrt{f}$ as $s \downarrow 0$ in $L^2(X, \mathfrak{m})$, we get $\mathsf{Ch}(\sqrt{f}) \le \underline{\lim}_{t\downarrow 0} \frac{1}{t}\int_0^t \mathsf{Ch}(\sqrt{f_s})\mathrm{d}s$. On the other hand, the bound $f \ge c > 0$ ensures $\frac{|Df|_w^2}{f^2} \in L^1(X, \mathfrak{m})$ and the maximum principle again together with the convergence of f_s to f in $L^2(X, \mathfrak{m})$ when $s \downarrow 0$ grants that the convergence is also weak* in $L^\infty(X, \mathfrak{m})$, therefore $\int_X \frac{|Df|_w^2}{f}\mathrm{d}\mathfrak{m} = \lim_{t\downarrow 0} \frac{1}{t}\int_0^t \int_X \frac{|Df|_w^2}{f^2} f_s \mathrm{d}\mathfrak{m}\mathrm{d}s$.

In summary, we proved

$$\frac{1}{2}\int_X \frac{|Df|_*^2}{f}\mathrm{d}\mathfrak{m} \le \frac{1}{2}\int_X \frac{|Df|_w^2}{f}\mathrm{d}\mathfrak{m},$$

which, together with the inequality $|Df|_w \le |Df|_*$ $\mathfrak{m}$-a.e. in X, gives the conclusion. □

We are now in the position of defining the Sobolev space $W^{1,2}(X, \mathsf{d}, \mathfrak{m})$. We start with the following simple and general lemma.

Lemma 4.4 *Let $(B, \|\cdot\|)$ be a Banach space and let $E : B \to [0, \infty]$ be a 1-homogeneous, convex and lower semicontinuous map. Then the vector space $\{E < \infty\}$ endowed with the norm*

$$\|v\|_E := \sqrt{\|v\|^2 + E^2(v)},$$

is a Banach space.

Proof It is clear that $(D(E), \|\cdot\|_E)$ is a normed space, so we only need to prove completeness. Pick a sequence $(v_n) \subset D(E)$ which is Cauchy w.r.t. $\|\cdot\|_E$. Then, since $\|\cdot\| \le \|\cdot\|_E$ we also get that (v_n) is Cauchy w.r.t. $\|\cdot\|$, and hence there exists $v \in B$ such that $\|v_n - v\| \to 0$. The lower semicontinuity of E grants that $E(v) \le \underline{\lim}_n E(v_n) < \infty$ and also that it holds

$$\overline{\lim_{n\to\infty}} \|v_n - v\|_E \le \overline{\lim_{n,m\to\infty}} \|v_n - v_m\|_E = 0,$$

which is the thesis. □

Therefore, if we want to build the space $W^{1,2}(X, \mathsf{d}, \mathfrak{m}) \subset L^2(X, \mathfrak{m})$, the only thing that we need is an L^2-lower semicontinuous functional playing the role which on $\mathbb{R}^d$ is played by the L^2-norm of the distributional gradient of Sobolev functions. We certainly have this functional, namely the map $f \mapsto$

$\||Df|_*\|_{L^2(X,\mathfrak{m})} = \||Df|_w\|_{L^2(X,\mathfrak{m})}$. Hence the lemma above provides the Banach space $W^{1,2}(X, \mathsf{d}, \mathfrak{m})$. Notice that in general $W^{1,2}(X, \mathsf{d}, \mathfrak{m})$ is not Hilbert: this is not surprising, as already the Sobolev space $W^{1,2}$ built over $(\mathbb{R}^d, \|\cdot\|, \mathcal{L}^d)$ is not Hilbert if the underlying norm $\|\cdot\|$ does not come from a scalar product.

4.4 Comparison with Previous Approaches

It is now time to underline that the one proposed here is certainly not the first definition of Sobolev space over a metric measure space (we refer to [17] for a much broader overview on the subject). Here we confine the discussion only to weak notions of (modulus of) gradient, and in particular to [9] and [20, 30]. Also, we discuss only the quadratic case, referring to [5] for general power functions p and the independence (in a suitable sense) of p of minimal gradients.

In [9] Cheeger proposed a relaxation procedure similar to the one used in Sect. 4.1, but rather than relaxing the local Lipschitz constant of Lipschitz functions, he relaxed upper gradients of arbitrary functions. More precisely, he defined

$$E(f) := \inf \varliminf_{n\to\infty} \|G_n\|_{L^2(X,\mathfrak{m})},$$

where the infimum is taken among all sequences (f_n) converging to f in $L^2(X, \mathfrak{m})$ such that G_n is an upper gradient for f_n. Then, with the same computations done in Sect. 4.1 (actually and obviously, the story goes the other way around: we closely followed his arguments) he showed that for $f \in D(E)$ there is an underlying notion of weak gradient $|Df|_C$, called minimal generalized upper gradient, such that $E(f) = \||Df|_C\|_{L^2(X,\mathfrak{m})}$ and

$$|Df|_C \le G \quad \mathfrak{m}\text{-a.e. in } X,$$

for any G weak limit of a sequence (G_n) as in the definition of $E(f)$.

Notice that since the local Lipschitz constant is always an upper gradient for Lipschitz functions, one certainly has

$$|Df|_C \le |Df|_* \quad \mathfrak{m}\text{-a.e. in } X, \text{ for any } f \in D(\mathsf{Ch}). \tag{37}$$

Koskela and MacManus [20] introduced and Shanmugalingam [30] further studied a procedure close to ours (again: actually we have been inspired by them) to produce a notion of "norm of weak gradient" which does not require a relaxation procedure. Recall that for $\Gamma \subset AC([0,1], X)$ the 2-Modulus $\mathrm{Mod}_2(\Gamma)$ is defined by

$$\mathrm{Mod}_2(\Gamma) := \inf\left\{ \|\rho\|^2_{L^2(X,\mathfrak{m})} : \int_\gamma \rho \ge 1 \ \forall \gamma \in \Gamma \right\} \quad \text{for every } \Gamma \subset AC\big([0,1], X\big). \tag{38}$$

It is possible to show that the 2-Modulus is an outer measure on $AC([0,1],X)$. Building on this notion, Koskela and MacManus [20] considered the class of functions f which satisfy the upper gradient inequality not necessarily along all curves, but only out of a Mod_2-negligible set of curves. In order to compare more properly this concept to Sobolev classes, Shanmugalingam said that $G : X \to [0,\infty]$ is a weak upper gradient for f if there exists $\tilde{f} = f$ $\mathfrak{m}$-a.e. such that

$$\left|\tilde{f}(\gamma_0) - \tilde{f}(\gamma_1)\right| \le \int_\gamma G \quad \text{for every } \gamma \in AC\big([0,1],X\big) \setminus \mathcal{N} \quad \text{with } \mathrm{Mod}_2(\mathcal{N}) = 0.$$

Then, she defined the energy $\tilde{E} : L^2(X,\mathfrak{m}) \to [0,\infty]$ by putting

$$\tilde{E}(f) := \inf \|G\|^2_{L^2(X,\mathfrak{m})},$$

where the infimum is taken among all weak upper gradient G of f according to the previous condition. Thanks to the properties of the 2-modulus (a stability property of weak upper gradients analogous to ours), it is possible to show that $\tilde{E}$ is indeed L^2-lower semicontinuous, so that it leads to a good definition of the Sobolev space. Also, using a key lemma due to Fuglede, Shanmugalingam proved that $E = \tilde{E}$ on $L^2(X,\mathfrak{m})$, so that they produce the same definition of Sobolev space $W^{1,2}(X,\mathsf{d},\mathfrak{m})$ and the underlying gradient $|Df|_S$ which gives a pointwise representation to $\tilde{E}(f)$ is the same $|Df|_C$ behind the energy E.

Observe now that for a Borel set $\Gamma \subset AC^2([0,1],X)$ and a test plan $\boldsymbol{\pi}$, integrating w.r.t. $\boldsymbol{\pi}$ the inequality $\int_\gamma \rho \ge 1$ $\forall \gamma \in \Gamma$ and then minimizing over ρ, we get

$$\big[\boldsymbol{\pi}(\Gamma)\big]^2 \le C(\boldsymbol{\pi})\,\mathrm{Mod}_2(\Gamma) \iint_0^1 |\dot{\gamma}|^2 \mathrm{d}s\mathrm{d}\boldsymbol{\pi}(\gamma),$$

which shows that any Mod_2-negligible set of curves is also negligible according to Definition 4.5. This fact easily yields that any $f \in D(\tilde{E})$ is Sobolev along a.e. curve and satisfies

$$|Df|_w \le |Df|_C, \quad \mathfrak{m}\text{-a.e. in } X. \tag{39}$$

Given that we proved in Theorem 4.6 that $|Df|_* = |Df|_w$, inequalities (37) and (39) also give that $|Df|_* = |Df|_w = |Df|_C = |Df|_S$ (the smallest one among the four notions coincides with the largest one).

What we get by the new approach to Sobolev spaces on metric measure spaces is the following result.

Theorem 4.7 (Density in energy of Lipschitz functions) *Let $(X,\mathsf{d},\mathfrak{m})$ be a compact normalized metric measure space. Then for any $f \in L^2(X,\mathfrak{m})$ with weak upper gradient in $L^2(X,\mathfrak{m})$ there exists a sequence (f_n) of Lipschitz functions converging to f in $L^2(X,\mathfrak{m})$ such that both $|Df_n|$ and $|Df_n|_w$ converge to $|Df|_w$ in $L^2(X,\mathfrak{m})$ as $n \to \infty$.*

Proof Straightforward consequence of the identity of weak and relaxed gradients and of Proposition 4.8. □

Let us point out a few aspects behind the strategy of the proof of Theorem 4.7, which of course strongly relies on Lemma 4.3 and Proposition 4.14. First of all, let us notice that the stated existence of a sequence of Lipschitz function f_n converging to f with $|Df_n| \to |Df|_w$ in $L^2(X, \mathfrak{m})$ is equivalent to show that

$$\lim_{n\to\infty} Y_{1/n}(f) \le \int_X |Df|_w^2 \mathrm{d}\mathfrak{m}, \tag{40}$$

where, for $\tau > 0$, Y_τ denotes the Yosida regularization

$$Y_\tau(f) := \inf_{h\in \mathrm{Lip}(X)} \left\{ \frac{1}{2}\int_X |Dh|^2 \mathrm{d}\mathfrak{m} + \frac{1}{2\tau}\int_X |h-f|^2 \mathrm{d}\mathfrak{m} \right\}.$$

In fact, the sequence f_n can be chosen by a simple diagonal argument among the approximate minimizers of $Y_{1/n}(f)$. On the other hand, it is well known that the relaxation procedure we used to define the Cheeger energy yields

$$Y_{1/n}(f) = \min_{h\in D(\mathsf{Ch})} \left\{ \mathsf{Ch}(h) + \frac{n}{2}\int_X |h-f|^2 \mathrm{d}\mathfrak{m} \right\}, \tag{41}$$

and therefore (40) could be achieved by trying to estimate the Cheeger energy of the unique minimizer $\tilde{f}_n$ of (41) in terms of $|Df|_w$.

Instead of using the Yosida regularization $Y_{1/n}$, in the proof of Theorem 4.6 we obtained a better approximation of f by flowing it (for a small time step, say $t_n \downarrow 0$) through the L^2-gradient flow f_t of the Cheeger energy. This flow is strictly related to Y_τ, since it can be obtained as the limit of suitably rescaled iterated minimizers of Y_τ (the so called Minimizing Movement scheme, see e.g. [3]), but has the great advantage to provide a continuous curve of probability densities f_t, which can be represented as the image of a test plan, through Lisini's Theorem. Thanks to this representation and Kuwada's Lemma, we were allowed to use the weak upper gradient $|Df|_w$ instead of $|Df|_*$ to estimate the Entropy dissipation along f_t (see (36)) and to obtain the desired sharp bound of $|Df_s|_*$ at least for some time $s \in (0, t_n)$. In any case, *a posteriori* we recovered the validity of (40).

This density result was previously known (via the use of maximal functions and covering arguments) under the assumption that the space was doubling and supported a local Poincaré inequality for weak upper gradients, see [9, Theorem 4.14, Theorem 4.24]. Actually, Cheeger proved more, namely that under these hypotheses Lipschitz functions are dense in the $W^{1,2}$ norm, a result which is still unknown in the general case. Also, notice that another byproduct of our density in energy result is the equivalence of local Poincaré inequality stated for Lipschitz functions on the left hand side and slope on the right hand side, and local Poincaré inequality stated for general functions on the left hand side and upper gradients on the right hand side; this result was previously known [19] under much more restrictive assumptions on the metric measure structure.

5 The Relative Entropy and Its W_2-Gradient Flow

In this section we study the W_2-gradient flow of the relative entropy on spaces with Ricci curvature bounded below (in short: $CD(K,\infty)$ spaces). The content is essentially extracted from [12]. As before the space $(X,\mathsf{d},\mathfrak{m})$ is compact and normalized (i.e. $\mathfrak{m}(X)=1$).

Recall that the relative entropy functional $\mathrm{Ent}_{\mathfrak{m}}:\mathscr{P}(X)\to[0,\infty]$ is defined by

$$\mathrm{Ent}_{\mathfrak{m}}(\mu):=\begin{cases}\int_X f\log f\,\mathrm{d}\mathfrak{m} & \text{if } \mu=f\mathfrak{m},\\ +\infty & \text{otherwise.}\end{cases}$$

Definition 5.9 (Weak bound from below on the Ricci curvature) We say that $(X,\mathsf{d},\mathfrak{m})$ has Ricci curvature bounded from below by K for some $K\in\mathbb{R}$ if the Relative Entropy functional $\mathrm{Ent}_{\mathfrak{m}}$ is K-convex along geodesics in $(\mathscr{P}(X),W_2)$. More precisely, if for any $\mu_0,\mu_1\in D(\mathrm{Ent}_{\mathfrak{m}})$ there exists a constant speed geodesic $\mu_t:[0,1]\to\mathscr{P}(X)$ between μ_0 and μ_1 satisfying

$$\mathrm{Ent}_{\mathfrak{m}}(\mu_t)\le(1-t)\,\mathrm{Ent}_{\mathfrak{m}}(\mu_0)+t\,\mathrm{Ent}_{\mathfrak{m}}(\mu_1)-\frac{K}{2}t(1-t)W_2^2(\mu_0,\mu_1)\quad\forall t\in[0,1].$$

This definition was introduced in [23] and [31]. Its two basic features are: **compatibility** with the Riemannian case (i.e. a compact Riemannian manifold endowed with the normalized volume measure has Ricci curvature bounded below by K in the classical pointwise sense if and only if $\mathrm{Ent}_{\mathfrak{m}}$ is K-geodesically convex in $(\mathscr{P}(X),W_2)$) and **stability** w.r.t. measured Gromov-Hausdorff convergence.

We also recall that Finsler geometries are included in the class of metric measure spaces with Ricci curvature bounded below. This means that if we have a smooth compact Finsler manifold (that is: a differentiable manifold endowed with a norm—possibly not coming from an inner product—on each tangent space which varies smoothly on the base point) endowed with an arbitrary positive C^∞ measure, then this space has Ricci curvature bounded below by some $K\in\mathbb{R}$ (see the theorem stated at page 926 of [32] for the flat case and [24] for the general one).

The goal now is to study the W_2-gradient flow of $\mathrm{Ent}_{\mathfrak{m}}$. Notice that the general theory of gradient flows of K-convex functionals ensures the following existence result (see the representation formula for the slope (7) and Theorem 2.1).

Theorem 5.8 (Consequences of the general theory of gradient flows) *Let $(X,\mathsf{d},\mathfrak{m})$ be a $CD(K,\infty)$ space. Then the slope $|D^-\mathrm{Ent}_{\mathfrak{m}}|$ is lower semicontinuous w.r.t. weak convergence and for any $\mu\in D(\mathrm{Ent}_{\mathfrak{m}})$ there exists a gradient flow (in the EDE sense of Definition 2.1) of $\mathrm{Ent}_{\mathfrak{m}}$ starting from μ.*

Thus, existence is granted. The problem is then to show uniqueness of the gradient flow. To this aim, we need to introduce the concept of *push forward via a plan*.

Definition 5.10 (Push forward via a plan) Let $\mu \in \mathscr{P}(X)$ and let $\boldsymbol{\gamma} \in \mathscr{P}(X^2)$ be such that $\mu \ll \pi^1_\sharp \boldsymbol{\gamma}$. The measures $\boldsymbol{\gamma}_\mu \in \mathscr{P}(X^2)$ and $\boldsymbol{\gamma}_\sharp \mu \in \mathscr{P}(X)$ are defined as:

$$\mathrm{d}\boldsymbol{\gamma}_\mu(x,y) := \frac{\mathrm{d}\mu}{\mathrm{d}\pi^1_\sharp \boldsymbol{\gamma}}(x)\mathrm{d}\boldsymbol{\gamma}(x,y), \qquad \boldsymbol{\gamma}_\sharp \mu := \pi^2_\sharp \boldsymbol{\gamma}_\mu.$$

Observe that, since $\boldsymbol{\gamma}_\mu \ll \boldsymbol{\gamma}$, we have $\boldsymbol{\gamma}_\sharp \mu \ll \pi^2_\sharp \boldsymbol{\gamma}$. We will say that $\boldsymbol{\gamma}$ has bounded deformation if there exist $0 < c \le C < \infty$ such that $c\mathfrak{m} \le \pi^i_\sharp \boldsymbol{\gamma} \le C\mathfrak{m}$, $i = 1, 2$. Writing $\mu = f\pi^1_\sharp \boldsymbol{\gamma}$, the definition gives that

$$\boldsymbol{\gamma}_\sharp \mu = \eta \pi^2_\sharp \boldsymbol{\gamma} \quad \text{with } \eta \text{ given by} \quad \eta(y) = \int f(x)\mathrm{d}\boldsymbol{\gamma}_y(x), \tag{42}$$

where $\{\boldsymbol{\gamma}_y\}_{y\in X}$ is the disintegration of $\boldsymbol{\gamma}$ w.r.t. its second marginal.

The operation of push forward via a plan has interesting properties in connection with the relative entropy functional.

Proposition 5.15 *The following properties hold*:

(i) *For any $\mu, \nu \in \mathscr{P}(X)$, $\boldsymbol{\gamma} \in \mathscr{P}(X^2)$ such that $\mu, \nu \ll \pi^1_\sharp \boldsymbol{\gamma}$ it holds*

$$\mathrm{Ent}_{\boldsymbol{\gamma}_\sharp \nu}(\boldsymbol{\gamma}_\sharp \mu) \le \mathrm{Ent}_\nu(\mu).$$

(ii) *For $\mu \in D(\mathrm{Ent}_\mathfrak{m})$ and $\boldsymbol{\gamma} \in \mathscr{P}(X^2)$ with bounded deformation, it holds $\boldsymbol{\gamma}_\sharp \mu \in D(\mathrm{Ent}_\mathfrak{m})$.*

(iii) *Given $\boldsymbol{\gamma} \in \mathscr{P}(X^2)$ with bounded deformation, the map*

$$D(\mathrm{Ent}_\mathfrak{m}) \ni \mu \quad \mapsto \quad \mathrm{Ent}_\mathfrak{m}(\mu) - \mathrm{Ent}_\mathfrak{m}(\boldsymbol{\gamma}_\sharp \mu),$$

is convex (*w.r.t. linear interpolation of measures*).

Proof (i). We can assume $\mu \ll \nu$, otherwise there is nothing to prove. Then it is immediate to check from the definition that $\boldsymbol{\gamma}_\sharp \mu \ll \boldsymbol{\gamma}_\sharp \nu$. Let $\mu = f\nu$, $\nu = \theta \pi^1_\sharp \gamma$, $\boldsymbol{\gamma}_\sharp \mu = \eta \boldsymbol{\gamma}_\sharp \nu$, and $u(z) := z \log z$. By disintegrating $\boldsymbol{\gamma}$ as in (42), we have that

$$\eta(y) = \int f(x)\mathrm{d}\tilde{\boldsymbol{\gamma}}_y(x), \quad \tilde{\boldsymbol{\gamma}}_y = \left(\int \theta(x)\mathrm{d}\boldsymbol{\gamma}_y(x)\right)^{-1} \theta \boldsymbol{\gamma}_y.$$

The convexity of u and Jensen's inequality with the probability measures $\tilde{\boldsymbol{\gamma}}_y$ yield

$$u\big(\eta(y)\big) \le \int u\big(f(x)\big)\mathrm{d}\tilde{\boldsymbol{\gamma}}_y(x).$$

Since $\{\tilde{\boldsymbol{\gamma}}_y\}_{y\in X}$ is the disintegration of $\tilde{\boldsymbol{\gamma}} = (\theta \circ \pi^1)\boldsymbol{\gamma}$ with respect to its second marginal $\boldsymbol{\gamma}_\sharp \nu$ and the first marginal of $\tilde{\boldsymbol{\gamma}}$ is ν, by integration of both sides with

respect to $\boldsymbol{\gamma}_\sharp\nu$ we get

$$\begin{aligned}\mathrm{Ent}_{\boldsymbol{\gamma}_\sharp\nu}(\boldsymbol{\gamma}_\sharp\mu) &= \int u\big(\eta(y)\big)\mathrm{d}\boldsymbol{\gamma}_\sharp\nu(y) \le \int\left(\int u\big(f(x)\big)\mathrm{d}\tilde{\boldsymbol{\gamma}}_y(x)\right)\mathrm{d}\boldsymbol{\gamma}_\sharp\nu(y)\\ &\le \int u\big(f(x)\big)\mathrm{d}\tilde{\boldsymbol{\gamma}}(x,y) = \int u\big(f(x)\big)\mathrm{d}\nu(x) = \mathrm{Ent}_\nu(\mu).\end{aligned}$$

(ii). Taking into account the identity

$$\mathrm{Ent}_\nu(\mu) = \mathrm{Ent}_\sigma(\mu) + \int \log\left(\frac{\mathrm{d}\sigma}{\mathrm{d}\nu}\right)\mathrm{d}\mu, \tag{43}$$

valid for any $\mu,\nu,\sigma\in\mathscr{P}(X)$ with σ having bounded density w.r.t. ν, the fact that $\boldsymbol{\gamma}_\sharp(\pi^1_\sharp\boldsymbol{\gamma}) = \pi^2_\sharp\boldsymbol{\gamma}$ and the fact that $c\mathfrak{m}\le\pi^1_\sharp\boldsymbol{\gamma},\pi^2_\sharp\boldsymbol{\gamma}\le C\mathfrak{m}$, the conclusion follows from

$$\begin{aligned}\mathrm{Ent}_{\mathfrak{m}}(\boldsymbol{\gamma}_\sharp\mu) &\le \mathrm{Ent}_{\pi^2_\sharp\boldsymbol{\gamma}}(\boldsymbol{\gamma}_\sharp\mu) + \log C\\ &\le \mathrm{Ent}_{\pi^1_\sharp\boldsymbol{\gamma}}(\mu) + \log C \le \mathrm{Ent}_{\mathfrak{m}}(\mu) + \log C - \log c.\end{aligned}$$

(iii). Let $\mu_0,\mu_1\in D(\mathrm{Ent}_{\mathfrak{m}})$ and define $\mu_t := (1-t)\mu_0 + t\mu_1$ and $\nu_t := \boldsymbol{\gamma}_\sharp\mu_t$. A direct computation shows that

$$\begin{aligned}(1-t)\,\mathrm{Ent}_{\mathfrak{m}}(\mu_0) + t\,\mathrm{Ent}_{\mathfrak{m}}(\mu_1) - \mathrm{Ent}_{\mathfrak{m}}(\mu_t) &= (1-t)\,\mathrm{Ent}_{\mu_t}(\mu_0) + t\,\mathrm{Ent}_{\mu_t}(\mu_1),\\ (1-t)\,\mathrm{Ent}_{\mathfrak{m}}(\nu_0) + t\,\mathrm{Ent}_{\mathfrak{m}}(\nu_1) - \mathrm{Ent}_{\mathfrak{m}}(\nu_t) &= (1-t)\,\mathrm{Ent}_{\nu_t}(\nu_0) + t\,\mathrm{Ent}_{\nu_t}(\nu_1),\end{aligned}$$

and from (i) we have that

$$\mathrm{Ent}_{\mu_t}(\mu_i) \ge \mathrm{Ent}_{\boldsymbol{\gamma}_\sharp\mu_t}(\boldsymbol{\gamma}_\sharp\mu_i) = \mathrm{Ent}_{\nu_t}(\nu_i),\quad \forall t\in[0,1],\ i=0,1,$$

which gives the conclusion. □

In the next lemma and in the sequel we use the short notation

$$C(\boldsymbol{\gamma}) := \int_{X\times X}\mathsf{d}^2(x,y)\mathrm{d}\boldsymbol{\gamma}(x,y).$$

Lemma 5.5 (Approximability in Entropy and distance) *Let $\mu,\nu\in D(\mathrm{Ent}_{\mathfrak{m}})$. Then there exists a sequence $(\boldsymbol{\gamma}^n)$ of plans with bounded deformation such that $\mathrm{Ent}_{\mathfrak{m}}(\boldsymbol{\gamma}^n_\sharp\mu)\to\mathrm{Ent}_{\mathfrak{m}}(\nu)$ and $C(\boldsymbol{\gamma}^n_\mu)\to W_2^2(\mu,\nu)$ as $n\to\infty$.*

Proof Let f and g respectively be the densities of μ and ν w.r.t. $\mathfrak{m}$; pick $\boldsymbol{\gamma}\in\mathrm{OPT}(\mu,\nu)$ and, for every $n\in\mathbb{N}$, let $A_n := \{(x,y): f(x)+g(y)\le n\}$ and

$$\boldsymbol{\gamma}_n := c_n\left(\boldsymbol{\gamma}|_{A_n} + \frac{1}{n}(\mathrm{Id},\mathrm{Id})_\sharp\mathfrak{m}\right),$$

where $c_n \to 1$ is the normalization constant. It is immediate to check that $\boldsymbol{\gamma}_n$ is of bounded deformation and that this sequence satisfies the thesis (see [12] for further details). □

Proposition 5.16 (Convexity of the squared slope) *Let* $(X, \mathsf{d}, \mathfrak{m})$ *be a* $CD(K, \infty)$ *space. Then the map*

$$D(\mathrm{Ent}_{\mathfrak{m}}) \ni \mu \quad \mapsto \quad |D^- \mathrm{Ent}_{\mathfrak{m}}|^2(\mu)$$

is convex (*w.r.t. linear interpolation of measures*).

Notice that the only assumption that we make is the K-convexity of the entropy w.r.t. W_2, and from this we deduce the convexity w.r.t. the classical linear interpolation of measures of the squared slope.

Proof Recall that from (7) we know that

$$\left|D^- \mathrm{Ent}_{\mathfrak{m}}\right|(\mu) = \sup_{\substack{\nu \in \mathscr{P}_2(X) \\ \nu \neq \mu}} \frac{[\mathrm{Ent}_{\mathfrak{m}}(\mu) - \mathrm{Ent}_{\mathfrak{m}}(\nu) - \frac{K^-}{2} W_2^2(\mu, \nu)]^+}{W_2(\mu, \nu)}.$$

We claim that it also holds

$$\left|D^- \mathrm{Ent}_{\mathfrak{m}}\right|(\mu) = \sup_{\boldsymbol{\gamma}} \frac{[\mathrm{Ent}_{\mathfrak{m}}(\mu) - \mathrm{Ent}_{\mathfrak{m}}(\boldsymbol{\gamma}_\sharp \mu) - \frac{K^-}{2} C(\boldsymbol{\gamma}_\mu)]^+}{\sqrt{C(\boldsymbol{\gamma}_\mu)}},$$

where the supremum is taken among all plans with bounded deformation (where the right hand side is taken 0 by definition if $C(\boldsymbol{\gamma}_\mu) > 0$).

Indeed, Lemma 5.5 gives that the first expression is not larger than the second. For the converse inequality we can assume $C(\boldsymbol{\gamma}_\mu) > 0$, $\nu = \boldsymbol{\gamma}_\sharp \mu \neq \mu$, and $K < 0$. Then it is sufficient to apply the simple inequality

$$a, b, c \in \mathbb{R}, \quad 0 < b \leq c \quad \Rightarrow \quad \frac{(a-b)^+}{\sqrt{b}} \geq \frac{(a-c)^+}{\sqrt{c}},$$

with $a := \mathrm{Ent}_{\mathfrak{m}}(\mu) - \mathrm{Ent}_{\mathfrak{m}}(\boldsymbol{\gamma}_\sharp \mu)$, $b := \frac{K^-}{2} W_2^2(\mu, \boldsymbol{\gamma}_\sharp \mu)$ and $c := \frac{K^-}{2} C(\boldsymbol{\gamma}_\mu)$.

Thus, to prove the thesis it is enough to show that for every $\boldsymbol{\gamma}$ with bounded deformation the map

$$D(\mathrm{Ent}_{\mathfrak{m}}) \ni \mu \quad \mapsto \quad \frac{[(\mathrm{Ent}_{\mathfrak{m}}(\mu) - \mathrm{Ent}_{\mathfrak{m}}(\boldsymbol{\gamma}_\sharp \mu) - \frac{K^-}{2} C(\boldsymbol{\gamma}_\mu))^+]^2}{C(\boldsymbol{\gamma}_\mu)},$$

is convex w.r.t. linear interpolation of measures.

Clearly the map

$$D(\mathrm{Ent}_{\mathfrak{m}}) \ni \mu \quad \mapsto \quad C(\boldsymbol{\gamma}_{\mu}) = \int \left(\int \mathsf{d}^2(x,y) \mathrm{d}\boldsymbol{\gamma}_x(y) \right) \mathrm{d}\mu(x),$$

where $\{\boldsymbol{\gamma}_x\}$ is the disintegration of $\boldsymbol{\gamma}$ w.r.t. its first marginal, is linear. Thus, from (iii) of Proposition 5.15 we know that the map

$$\mu \quad \mapsto \quad \mathrm{Ent}_{\mathfrak{m}}(\mu) - \mathrm{Ent}_{\mathfrak{m}}(\boldsymbol{\gamma}_{\sharp}\mu) - \frac{K^-}{2} C(\boldsymbol{\gamma}_{\mu}),$$

is convex w.r.t. linear interpolation of measures. Hence the same is true for its positive part. The conclusion follows from the fact that the function $\Psi : [0,\infty)^2 \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\Psi(a,b) := \begin{cases} \frac{a^2}{b} & \text{if } b > 0, \\ +\infty & \text{if } b = 0, a > 0 \\ 0 & \text{if } a = b = 0, \end{cases}$$

is convex and it is nondecreasing w.r.t. a. □

The convexity of the squared slope allows to prove uniqueness of the gradient flow of the entropy:

Theorem 5.9 (Uniqueness of the gradient flow of $\mathrm{Ent}_{\mathfrak{m}}$) *Let $(X, \mathsf{d}, \mathfrak{m})$ be a $CD(K,\infty)$ space and let $\mu \in D(\mathrm{Ent}_{\mathfrak{m}})$. Then there exists a unique gradient flow of $\mathrm{Ent}_{\mathfrak{m}}$ starting from μ in $(\mathscr{P}(X), W_2)$.*

Proof We recall (inequality (4)) that the squared Wasserstein distance is convex w.r.t. linear interpolation of measures. Therefore, given two absolutely continuous curves (μ_t^1) and (μ_t^2), the curve $t \mapsto \mu_t := \frac{\mu_t^1 + \mu_t^2}{2}$ is absolutely continuous as well and its metric speed can be bounded from above by

$$|\dot{\mu}_t|^2 \le \frac{|\dot{\mu}_t^1|^2 + |\dot{\mu}_t^2|^2}{2}, \quad \text{for a.e. } t \in (0,\infty). \tag{44}$$

Let (μ_t^1) and (μ_t^2) be gradient flows of $\mathrm{Ent}_{\mathfrak{m}}$ starting from $\mu \in D(\mathrm{Ent}_{\mathfrak{m}})$. Then we have

$$\mathrm{Ent}_{\mathfrak{m}}(\mu) = \mathrm{Ent}_{\mathfrak{m}}\big(\mu_T^1\big) + \frac{1}{2}\int_0^T \big|\dot{\mu}_t^1\big|^2 \mathrm{d}t + \frac{1}{2}\int_0^T |D^- \mathrm{Ent}_{\mathfrak{m}}|^2\big(\mu_t^1\big)\mathrm{d}t, \quad \forall T \ge 0,$$

$$\mathrm{Ent}_{\mathfrak{m}}(\mu) = \mathrm{Ent}_{\mathfrak{m}}\big(\mu_T^2\big) + \frac{1}{2}\int_0^T \big|\dot{\mu}_t^2\big|^2 \mathrm{d}t + \frac{1}{2}\int_0^T |D^- \mathrm{Ent}_{\mathfrak{m}}|^2\big(\mu_t^2\big)\mathrm{d}t, \quad \forall T \ge 0.$$

Adding up these two equalities, using the convexity of the squared slope guaranteed by Proposition 5.16, the convexity of the squared metric speed given by (44) and the

strict convexity of the relative entropy, we deduce that for the curve $t \mapsto \mu_t$ it holds

$$\mathrm{Ent}_{\mathfrak{m}}(\mu) > \mathrm{Ent}_{\mathfrak{m}}(\mu_T) + \frac{1}{2}\int_0^T |\dot{\mu}_t|^2 \mathrm{d}t + \frac{1}{2}\int_0^T \left|D^-\mathrm{Ent}_{\mathfrak{m}}\right|^2(\mu_t)\mathrm{d}t,$$

for every T such that $\mu_T^1 \neq \mu_T^2$. This contradicts inequality (9). □

6 The Heat Flow as Gradient Flow

It is well known that on $\mathbb{R}^d$ the heat flow can be seen both as gradient flow of the Dirichlet energy in L^2 and as gradient flow of the relative entropy in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$. It is therefore natural to ask whether this identification between the two a priori different gradient flows persists or not in a general compact and normalized metric measure space $(X, \mathsf{d}, \mathfrak{m})$.

The strategy consists in considering a gradient flow (f_t) of Ch with nonnegative initial data and in proving that the curve $t \mapsto \mu_t := f_t\mathfrak{m}$ is a gradient flow of $\mathrm{Ent}_{\mathfrak{m}}(\cdot)$ in $(\mathscr{P}(X), W_2)$: by the uniqueness result of Theorem 5.9 this will be sufficient to conclude.

We already built most of the ingredients needed for the proof to work, the only thing that we should add is the following lemma, where the slope of $\mathrm{Ent}_{\mathfrak{m}}$ is bounded from above in terms of the notions of "norm of weak gradient" that we discussed in Chap. 4. Notice that the bound (47) for Lipschitz functions was already known to Lott-Villani [23], so that our added value here is the use of the density in energy of Lipschitz functions to get the correct, sharp inequality (45) (sharpness will be seen in (48)).

Lemma 6.6 (Fisher bounds slope) *Let $(X, \mathsf{d}, \mathfrak{m})$ be a compact and normalized $CD(K, \infty)$ metric-measure space and let f be a probability density which is Sobolev along a.e. curve. Then*

$$\left|D^-\mathrm{Ent}_{\mathfrak{m}}\right|^2(f\mathfrak{m}) \leq \int_X \frac{|Df|_w^2}{f}\mathrm{d}\mathfrak{m} = 4\int_X |D\sqrt{f}|_w^2\mathrm{d}\mathfrak{m}. \tag{45}$$

Proof Assume at first that f is Lipschitz with $0 < c \leq f$, and let (f_n) be a sequence of probability densities such that $W_2(f_n\mathfrak{m}, f\mathfrak{m}) \to 0$ and where the slope of $\mathrm{Ent}_{\mathfrak{m}}$ at $f\mathfrak{m}$ is attained. Choose $\boldsymbol{\gamma}_n \in \mathrm{OPT}(f\mathfrak{m}, f_n\mathfrak{m})$ and notice that

$$\begin{aligned}
&\int_X f\log f\mathrm{d}\mathfrak{m} - \int_X f_n \log f_n \mathrm{d}\mathfrak{m}\\
&\leq \int_X (f - f_n)\log f\mathrm{d}\mathfrak{m}\\
&= \int \big(\log f(x) - \log f(y)\big)\mathrm{d}\boldsymbol{\gamma}_n(x, y)
\end{aligned}$$

$$\leq \sqrt{\int \frac{(\log f(x) - \log f(y))^2}{\mathsf{d}^2(x,y)} \mathrm{d}\boldsymbol{\gamma}_n(x,y)} \sqrt{\int \mathsf{d}^2(x,y) \mathrm{d}\boldsymbol{\gamma}_n(x,y)}$$

$$= \left(\int \left(\int L^2(x,y) \mathrm{d}\boldsymbol{\gamma}_{n,x}(y) \right) f(x) \mathrm{d}\mathfrak{m}(x) \right)^{1/2} W_2(f\mathfrak{m}, f_n\mathfrak{m}), \tag{46}$$

where $\boldsymbol{\gamma}_{n,x}$ is the disintegration of $\boldsymbol{\gamma}_n$ with respect to $f\mathfrak{m}$, and L is the bounded Borel function

$$L(x,y) := \begin{cases} \frac{|\log f(x) - \log f(y)|}{\mathsf{d}(x,y)} & \text{if } x \neq y, \\ |D \log f|(x) = \frac{|Df|(x)}{f(x)} & \text{if} x = y. \end{cases}$$

Notice that for every $x \in X$ the map $y \mapsto L(x,y)$ is upper-semicontinuous; since $\int (\int \mathsf{d}^2(x,y) \mathrm{d}\boldsymbol{\gamma}_{n,x}) f(x) \mathrm{d}\mathfrak{m} \to 0$ as $n \to \infty$, we can assume without loss of generality that

$$\lim_{n\to\infty} \int \mathsf{d}^2(x,y) \mathrm{d}\boldsymbol{\gamma}_{n,x}(y) = 0 \quad \text{for } f\mathfrak{m}\text{-a.e. } x \in X.$$

Fatou's Lemma then yields

$$\overline{\lim_{n\to\infty}} \int L^2(x,y) \mathrm{d}\boldsymbol{\gamma}_n(x,y) \leq \int_X L^2(x,x) f(x) \mathrm{d}\mathfrak{m}(x) = \int_X \frac{|Df|^2}{f} \mathrm{d}\mathfrak{m},$$

hence (46) gives

$$\left| D^- \mathrm{Ent}_\mathfrak{m} \right|(f\mathfrak{m}) = \overline{\lim_{n\to\infty}} \frac{(\mathrm{Ent}_\mathfrak{m}(f\mathfrak{m}) - \mathrm{Ent}_\mathfrak{m}(f_n\mathfrak{m}))^+}{W_2(f\mathfrak{m}, f_n\mathfrak{m})} \leq \sqrt{\int_X \frac{|Df|^2}{f} \mathrm{d}\mathfrak{m}}. \tag{47}$$

We now turn to the general case. Let f be any probability density Sobolev along a.e. curve such that $\sqrt{f} \in D(\mathsf{Ch})$ (otherwise is nothing to prove). We use Theorem 4.7 to find a sequence of Lipschitz functions $(\sqrt{f_n})$ converging to $\sqrt{f}$ in $L^2(X,\mathfrak{m})$ and such that $|D\sqrt{f_n}| \to |D\sqrt{f}|_w$ in $L^2(X,\mathfrak{m})$ and $\mathfrak{m}$-a.e. Up to summing up positive and vanishing constants and multiplying for suitable normalization factors, we can assume that $0 < c_n \leq f_n$ and $\int_X f_n \mathrm{d}\mathfrak{m} = 1$, for any $n \in \mathbb{N}$. The conclusion follows passing to the limit in (47) by taking into account the weak lower semicontinuity of $|D^- \mathrm{Ent}_\mathfrak{m}|$ (formula (7) and discussion thereafter). □

Theorem 6.10 (The heat flow as gradient flow) *Let $f_0 \in L^2(X,\mathfrak{m})$ be such that $\mu_0 = f_0\mathfrak{m} \in \mathscr{P}(X)$ and denote by (f_t) the gradient flow of Ch in $L^2(X,\mathfrak{m})$ starting from f_0 and by (μ_t) the gradient flow of $\mathrm{Ent}_\mathfrak{m}$ in $(\mathscr{P}(X), W_2)$ starting from μ_0. Then $\mu_t = f_t\mathfrak{m}$ for any $t \geq 0$.*

Proof Thanks to the uniqueness result of Theorem 5.9, it is sufficient to prove that $(f_t\mathfrak{m})$ satisfies the Energy Dissipation Equality for $\mathrm{Ent}_\mathfrak{m}$ in $(\mathscr{P}(X), W_2)$. We assume first that $0 < c \leq f_0 \leq C < \infty$ $\mathfrak{m}$-a.e. in X, so that the maximum principle (Proposition 4.11) ensures $0 < c \leq f_t \leq C < \infty$ for any $t > 0$. By Proposition 4.11

we know that $t \mapsto \mathrm{Ent}_{\mathfrak{m}}(f_t\mathfrak{m})$ is absolutely continuous with derivative equal to $-\int_X \frac{|Df_t|_w^2}{f_t}\mathrm{d}\mathfrak{m}$. Lemma 4.3 ensures that $t \mapsto f_t\mathfrak{m}$ is absolutely continuous w.r.t. W_2 with squared metric speed bounded by $\int_X \frac{|Df_t|_w^2}{f_t}\mathrm{d}\mathfrak{m}$, so that taking into account Lemma 6.6 we get

$$\mathrm{Ent}_{\mathfrak{m}}(f_0\mathfrak{m}) \geq \mathrm{Ent}_{\mathfrak{m}}(f_t\mathfrak{m}) + \frac{1}{2}\int_0^t |\dot{f_s\mathfrak{m}}|^2\mathrm{d}s + \frac{1}{2}\int_0^t \left|D^-\mathrm{Ent}_{\mathfrak{m}}\right|^2(f_s\mathfrak{m})\mathrm{d}s,$$

which, together with (9), ensures the thesis.

For the general case we argue by approximation, considering

$$f_0^n := c_n \min\big\{n, \max\{f_0, 1/n\}\big\},$$

c_n being the normalizing constant, and the corresponding gradient flow (f_t^n) of Ch. The fact that $f_0^n \to f_0$ in $L^2(X,\mathfrak{m})$ and the convexity of Ch implies that $f_t^n \to f_t$ in $L^2(X,\mathfrak{m})$ for any $t > 0$. In particular, $W_2(f_t^n\mathfrak{m}, f_t\mathfrak{m}) \to 0$ as $n \to \infty$ for every t (because convergence w.r.t. W_2 is equivalent to weak convergence of measures).

Now notice that we know that

$$\mathrm{Ent}_{\mathfrak{m}}\big(f_0^n\mathfrak{m}\big) = \mathrm{Ent}_{\mathfrak{m}}\big(f_t^n\big) + \frac{1}{2}\int_0^t \big|\dot{f_s^n\mathfrak{m}}\big|^2\mathrm{d}s + \frac{1}{2}\int_0^t \left|D^-\mathrm{Ent}_{\mathfrak{m}}\right|^2\big(f_s^n\big)\mathrm{d}s, \quad \forall t > 0.$$

Furthermore, it is immediate to check that $\mathrm{Ent}_{\mathfrak{m}}(f_0^n\mathfrak{m}) \to \mathrm{Ent}_{\mathfrak{m}}(f_0\mathfrak{m})$ as $n \to \infty$. The pointwise convergence of $f_t^n\mathfrak{m}$ to $f_t\mathfrak{m}$ w.r.t. W_2 easily yields that the terms on the right hand side of the last equation are lower semicontinuous when $n \to \infty$ (recall Theorem 5.8 for the slope). Thus it holds

$$\mathrm{Ent}_{\mathfrak{m}}(f_0\mathfrak{m}) \geq \mathrm{Ent}_{\mathfrak{m}}(f_t) + \frac{1}{2}\int_0^t |\dot{f_s\mathfrak{m}}|^2\mathrm{d}s + \frac{1}{2}\int_0^t \left|D^-\mathrm{Ent}_{\mathfrak{m}}\right|^2(f_s)\mathrm{d}s, \quad \forall t > 0,$$

which, by (11), is the thesis.

We know, by Theorem 5.9, that there is at most a gradient flow starting from μ_0. We also know that a gradient flow f_t' of Ch starting from f_0 exists, and part (i) gives that $\mu_t' := f_t'\mathfrak{m}$ is a gradient flow of $\mathrm{Ent}_{\mathfrak{m}}$. The uniqueness of gradient flows gives $\mu_t = \mu_t'$ for all $t \geq 0$. □

As a consequence of the previous Theorem 6.10 it would not be difficult to prove that the inequality (45) is in fact an identity: if $(X, \mathsf{d}, \mathfrak{m})$ is a compact and normalized $CD(K,\infty)$ space, then $|D^-\mathrm{Ent}_{\mathfrak{m}}|(f\mathfrak{m}) < \infty$ if and only if the probability density f is Sobolev along a.e. curve and $\sqrt{f} \in D(\mathsf{Ch})$; in this case

$$\left|D^-\mathrm{Ent}_{\mathfrak{m}}\right|^2(f\mathfrak{m}) = \int_X \frac{|Df|_w^2}{f}\mathrm{d}\mathfrak{m} = 4\int_X |D\sqrt{f}|_w^2\mathrm{d}\mathfrak{m}. \tag{48}$$

7 A Metric Brenier Theorem

In this section we state and prove the metric Brenier theorem in $CD(K,\infty)$ spaces we announced in the introduction. It was recently proven in [14] that under an additional non-branching assumption one can really recover an optimal transport map, see also [7] for related results, obtained under stronger non-branching assumptions and weaker convexity assumptions.

Definition 7.11 (Strict $CD(K,\infty)$ spaces) We say that a compact normalized metric measure space $(X,\mathsf{d},\mathfrak{m})$ is a strict $CD(K,\infty)$ space if for any $\mu_0,\mu_1\in D(\mathrm{Ent}_\mathfrak{m})$ there exists $\boldsymbol{\pi}\in\mathrm{GeoOpt}(\mu_0,\mu_1)$ with the following property. For any bounded Borel function $F:\mathrm{Geo}(X)\to[0,\infty)$ such that $\int F\,\mathrm{d}\boldsymbol{\pi}=1$, it holds

$$\mathrm{Ent}_\mathfrak{m}\big(\mu_t^F\big)\le(1-t)\,\mathrm{Ent}_\mathfrak{m}\big(\mu_0^F\big)+t\,\mathrm{Ent}_\mathfrak{m}\big(\mu_1^F\big)-\frac{K}{2}t(1-t)W_2^2\big(\mu_0^F,\mu_1^F\big),$$

where $\mu_t^F:=(\mathrm{e}_t)_\sharp(F\boldsymbol{\pi})$, for any $t\in[0,1]$.

Thus, the difference between strict $CD(K,\infty)$ spaces and standard $CD(K,\infty)$ ones is the fact that geodesic convexity is required along *all* geodesics induced by the weighted plans $F\boldsymbol{\pi}$, rather than the one induced by $\boldsymbol{\pi}$ only. Notice that the necessary and sufficient optimality conditions ensure that $(\mathrm{e}_0,\mathrm{e}_1)_\sharp\boldsymbol{\pi}$ is concentrated on a c-monotone set, hence $(\mathrm{e}_0,\mathrm{e}_1)_\sharp(F\boldsymbol{\pi})$ has the same property and it is optimal, relative to its marginals. (We remark that recent results of Rajala [28] suggest that it is not necessary to assume this stronger convexity to get the metric Brenier theorem—and hence not even a treatable notion of spaces with Riemannian Ricci curvature bounded from below—see [2] for progresses in this direction.)

It is not clear to us whether the notion of being strict $CD(K,\infty)$ is stable or not w.r.t. measured Gromov-Hausdorff convergence and, as such, it should be handled with care. The importance of strict $CD(K,\infty)$ bounds relies on the fact that on these spaces geodesic interpolation between bounded probability densities is made of bounded densities as well, thus granting the existence of many test plans.

Notice that non-branching $CD(K,\infty)$ spaces are always strict $CD(K,\infty)$ spaces, indeed let $\mu_0,\mu_1\in D(\mathrm{Ent}_\mathfrak{m})$ and pick $\boldsymbol{\pi}\in\mathrm{GeoOpt}(\mu_0,\mu_1)$ such that $\mathrm{Ent}_\mathfrak{m}$ is K-convex along $((\mathrm{e}_t)_\sharp\boldsymbol{\pi})$. From the non-branching hypothesis it follows that for F as in Definition 7.11 there exists a unique element in $\mathrm{GeoOpt}(\mu_t^F,\mu_1^F)$ (resp. in $\mathrm{GeoOpt}(\mu_t^F,\mu_0^F)$). Also, since F is bounded, from $\mu_t\in D(\mathrm{Ent}_\mathfrak{m})$ we deduce $\mu_t^F\in D(\mathrm{Ent}_\mathfrak{m})$. Hence the map $t\mapsto\mathrm{Ent}_\mathfrak{m}(\mu_t^F)$ is K-convex and bounded on $[\varepsilon,1]$ and on $[0,1-\varepsilon]$ for all $\varepsilon\in(0,1)$, and therefore it is K-convex on $[0,1]$.

Proposition 7.17 (Bound on geodesic interpolant) *Let $(X,\mathsf{d},\mathfrak{m})$ be a strict $CD(K,\infty)$ space and let $\mu_0,\mu_1\in\mathscr{P}(X)$ be with bounded densities. Then there exists a test plan $\boldsymbol{\pi}\in\mathrm{GeoOpt}(\mu_0,\mu_1)$ so that the induced geodesic $\mu_t=(\mathrm{e}_t)_\sharp\boldsymbol{\pi}$ connecting μ_0 to μ_1 is made of measures with uniformly bounded densities.*

Proof Let M be an upper bound on the densities of μ_0, μ_1, $\boldsymbol{\pi} \in \mathrm{GeoOpt}(\mu_0, \mu_1)$ be a plan which satisfies the assumptions of Definition 7.11 and $\mu_t := (\mathrm{e}_t)_\sharp \boldsymbol{\pi}$. We claim that the measures μ_t have uniformly bounded densities. The fact that $\mu_t \ll \mathfrak{m}$ is obvious by geodesic convexity, so let f_t be the density of μ_t and assume by contradiction that for some $t_0 \in [0, 1]$ it holds

$$f_{t_0}(x) > M e^{K^- \mathrm{D}^2/8}, \quad \forall x \in A, \tag{49}$$

where $\mathfrak{m}(A) > 0$ and D is the diameter of X. Define $\tilde{\boldsymbol{\pi}} := c\boldsymbol{\pi}|_{\mathrm{e}_{t_0}^{-1}(A)}$, where c is the normalizing constant (notice that $\tilde{\boldsymbol{\pi}}$ is well defined, because $\boldsymbol{\pi}(\mathrm{e}_{t_0}^{-1}(A)) = \mu_{t_0}(A) > 0$) and observe that the density of $\tilde{\boldsymbol{\pi}}$ w.r.t. $\boldsymbol{\pi}$ is bounded. Let $\tilde{\mu}_t := (\mathrm{e}_t)_\sharp \tilde{\boldsymbol{\pi}}$ and $\tilde{f}_t$ its density w.r.t. $\mathfrak{m}$. From (49) we get $\tilde{f}_{t_0} = c f_{t_0}$ on A and $\tilde{f}_{t_0} = 0$ on $X \setminus A$, hence

$$\mathrm{Ent}_{\mathfrak{m}}(\tilde{\mu}_{t_0}) = \int \log(\tilde{f}_{t_0} \circ \mathrm{e}_{t_0}) \mathrm{d}\boldsymbol{\pi} > \log c + \log M + \frac{K^-}{8}\mathrm{D}^2. \tag{50}$$

On the other hand, we have $\tilde{f}_0 \le c f_0 \le cM$ and $\tilde{f}_1 \le c f_1 \le cM$ and thus

$$\mathrm{Ent}_{\mathfrak{m}}(\tilde{\mu}_i) = \int \log(\tilde{f}_i \circ \mathrm{e}_i) \mathrm{d}\tilde{\boldsymbol{\pi}} \le \log c + \log M, \quad i = 0, 1. \tag{51}$$

Finally, it certainly holds $W_2^2(\tilde{\mu}_0, \tilde{\mu}_1) \le \mathrm{D}^2$, so that (50) and (51) contradict the K-convexity of $\mathrm{Ent}_{\mathfrak{m}}$ along $(\tilde{\mu}_t)$. Hence (49) is false and the f_t's are uniformly bounded. □

An important consequence of this uniform bound is the following metric version of Brenier's theorem.

Theorem 7.11 (A metric Brenier theorem) *Let $(X, \mathsf{d}, \mathfrak{m})$ be a strict $CD(K, \infty)$ space, let f_0, f_1 be probability densities and φ any Kantorovich potential for the couple $(f_0\mathfrak{m}, f_1\mathfrak{m})$. Then for every $\boldsymbol{\pi} \in \mathrm{GeoOpt}(f_0\mathfrak{m}, f_1\mathfrak{m})$ it holds*

$$\mathsf{d}(\gamma_0, \gamma_1) = |D\varphi|_w(\gamma_0) = \left|D^+\varphi\right|(\gamma_0), \quad \text{for } \boldsymbol{\pi}\text{-a.e. } \gamma. \tag{52}$$

In particular,

$$W_2^2(f_0\mathfrak{m}, f_1\mathfrak{m}) = \int_X |D\varphi|_*^2 f_0 \mathrm{d}\mathfrak{m}.$$

If moreover $f_0, f_1 \in L^\infty(X, \mathfrak{m})$ and $\boldsymbol{\pi}$ is a test plan (such a plan exists thanks to Proposition 7.17) then

$$\lim_{t \downarrow 0} \frac{\varphi(\gamma_0) - \varphi(\gamma_t)}{\mathsf{d}(\gamma_0, \gamma_t)} = \left|D^+\varphi\right|(\gamma_0) \quad \text{in } L^2\big(\mathrm{Geo}(X), \boldsymbol{\pi}\big). \tag{53}$$

Proof φ is Lipschitz, therefore $|D^+\varphi|$ is an upper gradient of φ, and hence $|D\varphi|_w \le |D^+\varphi|$ $\mathfrak{m}$-a.e. Now fix $x \in X$ and pick any $y \in \partial^c\varphi(x)$. From the c-concavity of φ

we get

$$\varphi(x) = \frac{\mathsf{d}^2(x,y)}{2} - \varphi^c(y),$$
$$\varphi(z) \le \frac{\mathsf{d}^2(z,y)}{2} - \varphi^c(y) \quad \forall z \in X.$$

Therefore

$$\varphi(z) - \varphi(x) \le \frac{\mathsf{d}^2(z,y)}{2} - \frac{\mathsf{d}^2(x,y)}{2} \le \mathsf{d}(z,x)\frac{\mathsf{d}(z,y) + \mathsf{d}(x,y)}{2}.$$

Dividing by $\mathsf{d}(x,z)$ and letting $z \to x$, by the arbitrariness of $y \in \partial^c\varphi(x)$ and the fact that $\mathrm{supp}((\mathrm{e}_0,\mathrm{e}_1)_\sharp\boldsymbol{\pi}) \subset \partial^c\varphi$ we get

$$\left|D^+\varphi\right|(\gamma_0) \le \min_{y\in\partial^c\varphi(\gamma_0)} \mathsf{d}(\gamma_0,y) \le \mathsf{d}(\gamma_0,\gamma_1) \quad \text{for } \boldsymbol{\pi}\text{-a.e. } \gamma.$$

Since

$$\int_X |D\varphi|_w^2 f_0 \mathrm{d}\mathfrak{m} \le \int \left|D^+\varphi\right|^2(\gamma_0)\mathrm{d}\boldsymbol{\pi} \quad \text{and}$$
$$\int \mathsf{d}^2(\gamma_0,\gamma_1)\mathrm{d}\boldsymbol{\pi}(\gamma) = W_2^2(f_0\mathfrak{m}, f_1\mathfrak{m}),$$

to conclude it is sufficient to prove that

$$W_2^2(f_0\mathfrak{m}, f_1\mathfrak{m}) \le \int_X |D\varphi|_w^2 f_0 \mathrm{d}\mathfrak{m}. \tag{54}$$

Now assume that f_0 and f_1 are bounded from above and let $\tilde{\boldsymbol{\pi}} \in \mathrm{GeoOpt}(f_0\mathfrak{m}, f_1\mathfrak{m})$ be a test plan (such $\tilde{\boldsymbol{\pi}}$ exists thanks to Proposition 7.17). Since φ is a Kantorovich potential and $(\mathrm{e}_0,\mathrm{e}_1)_\sharp\tilde{\boldsymbol{\pi}}$ is optimal, it holds $\gamma_1 \in \partial^c\varphi(\gamma_0)$ for any $\gamma \in \mathrm{supp}(\tilde{\boldsymbol{\pi}})$. Hence arguing as before we get

$$\varphi(\gamma_0) - \varphi(\gamma_t) \ge \frac{\mathsf{d}^2(\gamma_0,\gamma_1)}{2} - \frac{\mathsf{d}^2(\gamma_t,\gamma_1)}{2} = \mathsf{d}^2(\gamma_0,\gamma_1)\big(t - t^2/2\big). \tag{55}$$

Dividing by $\mathsf{d}(\gamma_0,\gamma_t) = t\mathsf{d}(\gamma_0,\gamma_1)$, squaring and integrating w.r.t. $\tilde{\boldsymbol{\pi}}$ we obtain

$$\varliminf_{t\downarrow 0} \int \left(\frac{\varphi(\gamma_0) - \varphi(\gamma_t)}{\mathsf{d}(\gamma_0,\gamma_t)}\right)^2 \mathrm{d}\tilde{\boldsymbol{\pi}}(\gamma) \ge \int \mathsf{d}^2(\gamma_0,\gamma_1)\mathrm{d}\tilde{\boldsymbol{\pi}}(\gamma) = W_2^2(f_0\mathfrak{m}, f_1\mathfrak{m}). \tag{56}$$

Using Remark 4.4 and the fact that $\tilde{\boldsymbol{\pi}}$ is a test plan we have

$$\begin{aligned}\int\left(\frac{\varphi(\gamma_0)-\varphi(\gamma_t)}{\mathsf{d}(\gamma_0,\gamma_t)}\right)^2\mathrm{d}\tilde{\boldsymbol{\pi}}(\gamma) &\le \int\frac{1}{t^2}\left(\int_0^t|D\varphi|_w(\gamma_s)\mathrm{d}s\right)^2\mathrm{d}\tilde{\boldsymbol{\pi}}(\gamma)\\ &\le \frac{1}{t}\iint_0^t|D\varphi|_w^2(\gamma_s)\mathrm{d}s\mathrm{d}\tilde{\boldsymbol{\pi}}(\gamma)\\ &= \frac{1}{t}\iint_0^t|D\varphi|_w^2\mathrm{d}s\mathrm{d}(\mathrm{e}_t)_\sharp\tilde{\boldsymbol{\pi}}\\ &= \frac{1}{t}\iint_0^t|D\varphi|_w^2 f_s\mathrm{d}s\mathrm{d}\mathfrak{m},\end{aligned}\tag{57}$$

where f_s is the density of $(\mathrm{e}_s)_\sharp\tilde{\boldsymbol{\pi}}$. Since $(\mathrm{e}_t)_\sharp\tilde{\boldsymbol{\pi}}$ weakly converges to $(\mathrm{e}_0)_\sharp\tilde{\boldsymbol{\pi}}$ as $t\downarrow 0$ and $\mathrm{Ent}_{\mathfrak{m}}((\mathrm{e}_t)_\sharp\tilde{\boldsymbol{\pi}})$ is uniformly bounded (by the K-geodesic convexity), we conclude that $f_t\to f_0$ weakly in $L^1(X,\mathfrak{m})$ and since $|D\varphi|_w\in L^\infty(X,\mathfrak{m})$ we have

$$\lim_{t\downarrow 0}\frac{1}{t}\iint_0^t|D\varphi|_w^2 f_s\mathrm{d}s\mathrm{d}\mathfrak{m}=\int_X|D\varphi|_w^2 f_0\mathrm{d}\mathfrak{m}.\tag{58}$$

Equations (56), (57) and (58) yield (54).

In order to prove (54) in the general case of possibly unbounded densities, let us fix a Kantorovich potential φ, $\boldsymbol{\pi}\in\mathrm{GeoOpt}(f_0\mathfrak{m},f_1\mathfrak{m})$ and for $n\in\mathbb{N}$ define $\boldsymbol{\pi}^n:=c_n\boldsymbol{\pi}|_{\{\gamma:f_0(\gamma_0)+f_1(\gamma_1)\le n\}}$, $c_n\to 1$ being the normalization constant. Then $\boldsymbol{\pi}^n\in\mathrm{GeoOpt}(f_0^n\mathfrak{m},f_1^n\mathfrak{m})$, where $f_i^n:=(\mathrm{e}_i)_\sharp\boldsymbol{\pi}^n$, φ is a Kantorovich potential for $(f_0^n\mathfrak{m},f_1^n\mathfrak{m})$ and $f_0^n,f_1^n\in L^\infty(X,\mathfrak{m})$. Thus from what we just proved we know that it holds

$$\mathsf{d}(\gamma_0,\gamma_1)=|D\varphi|_w(\gamma_0)=\left|D^+\varphi\right|(\gamma_0),\quad\text{for }\boldsymbol{\pi}^n\text{-a.e. }\gamma.$$

Letting $n\to\infty$ we conclude.

Concerning (53), we can choose $\tilde{\boldsymbol{\pi}}=\boldsymbol{\pi}$ and obtain by (55) and (52)

$$\frac{\varphi(\gamma_0)-\varphi(\gamma_t)}{\mathsf{d}(\gamma_0,\gamma_t)}\ge 0,\qquad\liminf_{t\downarrow 0}\frac{\varphi(\gamma_0)-\varphi(\gamma_t)}{\mathsf{d}(\gamma_0,\gamma_t)}\ge\left|D^+\varphi\right|(\gamma_0)\quad\text{for }\boldsymbol{\pi}\text{-a.e. }\gamma.$$

On the other hand (57) and (58) yield

$$\limsup_{t\downarrow 0}\int\left(\frac{\varphi(\gamma_0)-\varphi(\gamma_t)}{\mathsf{d}(\gamma_0,\gamma_t)}\right)^2\mathrm{d}\boldsymbol{\pi}(\gamma)\le\int\left|D^+\varphi\right|^2(\gamma_0)\mathrm{d}\boldsymbol{\pi}(\gamma),$$

so that, by expanding the square and applying Fatou's Lemma, we obtain

$$\limsup_{t\downarrow 0}\int\left(\frac{\varphi(\gamma_0)-\varphi(\gamma_t)}{\mathsf{d}(\gamma_0,\gamma_t)}-\left|D^+\varphi\right|(\gamma_0)\right)^2\mathrm{d}\boldsymbol{\pi}(\gamma)\le 0.$$

□

8 More on Calculus on Compact $CD(K,\infty)$ Spaces

8.1 On Horizontal and Vertical Derivatives Again

Aim of this subsection is to prove another deep relation between "horizontal" and "vertical" derivation, which will allow to compare the derivative of the squared Wasserstein distance along the heat flow with the derivative of the relative entropy along a geodesic (see the next subsection). This will be key in order to understand the properties of spaces with *Riemannian* Ricci curvature bounded from below, illustrated in the last section.

In order to understand the geometric point, consider the following simple example.

Example 8.1 Let $\|\cdot\|$ be a smooth, strictly convex norm on $\mathbb{R}^d$ and let $\|\cdot\|_*$ be the dual norm. Denoting by $\langle\cdot,\cdot\rangle$ the canonical duality from $(\mathbb{R}^d)^*\times\mathbb{R}^d$ into $\mathbb{R}$, let $\mathcal{L}$ be the duality map from $(\mathbb{R}^d,\|\cdot\|)$ to $((\mathbb{R}^d)^*,\|\cdot\|_*)$, characterized by

$$\big\langle\mathcal{L}(u),u\big\rangle=\big\|\mathcal{L}(u)\big\|_*\|u\|\quad\text{and}\quad\big\|\mathcal{L}(u)\big\|_*=\|u\|\quad\forall u\in\mathbb{R}^d,$$

and let $\mathcal{L}^*$ be its inverse, equally characterized by

$$\big\langle v,\mathcal{L}^*(v)\big\rangle=\|v\|_*\big\|\mathcal{L}^*(v)\big\|\quad\text{and}\quad\big\|\mathcal{L}^*(v)\big\|=\|v\|_*\quad\forall v\in\big(\mathbb{R}^d\big)^*.$$

Using the fact that $\varepsilon\mapsto\|u\|\|u+\varepsilon u'\|-\langle\mathcal{L}u,u+\varepsilon u'\rangle$ attains its minimum at $\varepsilon=0$ and the analogous relation for $\mathcal{L}^*$, one obtains the useful relations

$$\big\langle\mathcal{L}(u),u'\big\rangle=\frac{1}{2}\mathrm{d}_u\|\cdot\|^2\big(u'\big),\qquad\big\langle v',\mathcal{L}^*(v)\big\rangle=\frac{1}{2}\mathrm{d}_v\|\cdot\|_*^2\big(v'\big).\tag{59}$$

For a smooth map $f:\mathbb{R}^d\to\mathbb{R}$ its differential $\mathrm{d}_x f$ at any point x is intrinsically defined as cotangent vector, namely as an element of $(\mathbb{R}^d)^*$. To define the gradient $\nabla f(x)\in\mathbb{R}^d$ (which is a tangent vector), the norm comes into play via the formula $\nabla f(x):=\mathcal{L}^*(\mathrm{d}_x f)$. Now, given two smooth functions f,g, the real number $\mathrm{d}_x f(\nabla g(x))$ is well defined as the application of the cotangent vector $\mathrm{d}_x f$ to the tangent vector $\nabla g(x)$.

What we want to point out, is that there are two very different ways of obtaining $\mathrm{d}_x f(\nabla g(x))$ from a derivation. The first one, which is usually taken as the definition of $\mathrm{d}_x f(\nabla g(x))$, is the "horizontal derivative":

$$\langle\mathrm{d}_x f,\nabla g\rangle=\mathrm{d}_x f\big(\nabla g(x)\big)=\lim_{t\to0}\frac{f(x+t\nabla g(x))-f(x)}{t}.\tag{60}$$

The second one is the "vertical derivative":

$$Df(\nabla g)(x)=\lim_{\varepsilon\to0}\frac{\frac{1}{2}\|\mathrm{d}_x(g+\varepsilon f)\|_*^2-\frac{1}{2}\|\mathrm{d}_x g\|_*^2(x)}{\varepsilon}.\tag{61}$$

It is not difficult to check that (61) is consistent with (60): indeed (omitting the x dependence), recalling the second identity of (59), we have

$$\|\mathrm{d}g+\varepsilon\mathrm{d}f\|_*^2=\|\mathrm{d}g\|_*^2+2\varepsilon\big\langle\mathcal{L}^*(\mathrm{d}g),\mathrm{d}f\big\rangle+o(\varepsilon)=\|\nabla g\|^2+2\varepsilon\langle\nabla g,\mathrm{d}f\rangle+o(\varepsilon).$$

The point is that the equality between the right hand sides of formulas (61) and (60) extends to a genuine metric setting. In the following lemma (where the plan $\boldsymbol{\pi}$ plays the role of $-\nabla g$) we prove one inequality, but we remark that "playing with signs" it is possible to obtain an analogous inequality with $\leq$ in place of $\geq$.

Lemma 8.7 (Horizontal and vertical derivatives) *Let f be a Sobolev function along a.e. curve with $|Df|_w\in L^2(X,\mathfrak{m})$, let $g:X\to\mathbb{R}$ be Lipschitz and let $\boldsymbol{\pi}$ be a test plan concentrated on* $\mathrm{Geo}(X)$. *Assume that*

$$\lim_{t\downarrow 0}\frac{g(\gamma_0)-g(\gamma_t)}{\mathsf{d}(\gamma_0,\gamma_t)}=|Dg|_w(\gamma_0)\quad\text{in } L^2\big(\mathrm{Geo}(X),\boldsymbol{\pi}\big).\tag{62}$$

Then

$$\begin{aligned}&\lim_{t\downarrow 0}\int\frac{f(\gamma_t)-f(\gamma_0)}{t}\mathrm{d}\boldsymbol{\pi}(\gamma)\\&\quad\geq\frac{1}{2}\int\frac{|Dg|_w^2(\gamma_0)-|D(g+\varepsilon f)|_w^2(\gamma_0)}{\varepsilon}\mathrm{d}\boldsymbol{\pi}(\gamma)\quad\forall\varepsilon>0.\end{aligned}\tag{63}$$

Proof Define the functions $F_t,G_t:\mathrm{Geo}(X)\to\mathbb{R}\cup\{\pm\infty\}$ by

$$F_t(\gamma):=\frac{f(\gamma_0)-f(\gamma_t)}{\mathsf{d}(\gamma_0,\gamma_t)},$$
$$G_t(\gamma):=\frac{g(\gamma_0)-g(\gamma_t)}{\mathsf{d}(\gamma_0,\gamma_t)}.$$

By (62) it holds

$$\int|Dg|_w^2\circ\mathrm{e}_0\mathrm{d}\boldsymbol{\pi}(\gamma)=\lim_{t\downarrow 0}\int G_t^2\mathrm{d}\boldsymbol{\pi}.\tag{64}$$

Since the measures $(\mathrm{e}_t)_\sharp\boldsymbol{\pi}\to(\mathrm{e}_0)_\sharp\boldsymbol{\pi}$ weakly in duality with $C(X)$ as $t\downarrow 0$ and their densities with respect to $\mathfrak{m}$ are uniformly bounded, we obtain that the densities are weakly* convergent in $L^\infty(X,\mathfrak{m})$. Therefore, using the fact that $|D(g+\varepsilon f)|_w^2\in L^1(X,\mathfrak{m})$ and taking into account Remark 4.4 we obtain

$$\begin{aligned}\int\big|D(g+\varepsilon f)\big|_w^2\circ\mathrm{e}_0\mathrm{d}\boldsymbol{\pi}(\gamma)&=\int\big|D(g+\varepsilon f)\big|_w^2\mathrm{d}(\mathrm{e}_0)_\sharp\boldsymbol{\pi}\\&=\lim_{t\downarrow 0}\frac{1}{t}\int_0^t\int_X\big|D(g+\varepsilon f)\big|_w^2\mathrm{d}(\mathrm{e}_s)_\sharp\boldsymbol{\pi}\mathrm{d}s\end{aligned}$$

$$
\begin{aligned}
&= \lim_{t\downarrow 0} \frac{1}{t} \iint_0^t \left|D(g+\varepsilon f)\right|_w^2(\gamma_s)\mathrm{d}s\mathrm{d}\boldsymbol{\pi}(\gamma) \\
&\geq \overline{\lim_{t\downarrow 0}} \int \left| \frac{(g+\varepsilon f)(\gamma_0) - (g+\varepsilon f)(\gamma_t)}{t\mathsf{d}(\gamma_0,\gamma_1)} \right|^2 \mathrm{d}\boldsymbol{\pi}(\gamma) \\
&\geq \overline{\lim_{t\downarrow 0}} \int G_t^2 + 2\varepsilon G_t F_t \mathrm{d}\boldsymbol{\pi}.
\end{aligned}
$$

Subtracting this inequality from (64) and dividing by 2ε we get

$$
\frac{1}{2}\int \frac{|Dg|_w^2(\gamma_0) - |D(g+\varepsilon f)|_w^2(\gamma_0)}{\varepsilon}\mathrm{d}\boldsymbol{\pi}(\gamma) \leq \underline{\lim_{t\downarrow 0}} -\int G_t(\gamma)F_t(\gamma)\mathrm{d}\boldsymbol{\pi}(\gamma).
$$

We know that $G_t \to |Dg|_w \circ \mathrm{e}_0$ in $L^2(\mathrm{Geo}(X),\boldsymbol{\pi})$ and that $|Dg|_w(\gamma_0) = \mathsf{d}(\gamma_0,\gamma_1)$ for $\boldsymbol{\pi}$-a.e. γ. Also, by Remark 4.4 and the fact that $\boldsymbol{\pi}$ is a test plan we easily get $\sup_{t\in[0,1]} \|F_t\|_{L^2(\boldsymbol{\pi})} < \infty$. Thus it holds

$$
\begin{aligned}
\underline{\lim_{t\downarrow 0}} -\int G_t(\gamma)F_t(\gamma)\mathrm{d}\boldsymbol{\pi}(\gamma) &= \underline{\lim_{t\downarrow 0}} -\int \mathsf{d}(\gamma_0,\gamma_1)F_t(\gamma)\mathrm{d}\boldsymbol{\pi}(\gamma) \\
&= \underline{\lim_{t\downarrow 0}} \int \frac{f(\gamma_t)-f(\gamma_0)}{t}\mathrm{d}\boldsymbol{\pi}(\gamma),
\end{aligned}
$$

which is the thesis. □

8.2 Two Important Formulas

Proposition 8.18 (Derivative of $\frac{1}{2}W_2^2$ along the heat flow) *Let $(f_t) \subset L^2(X,\mathfrak{m})$ be a heat flow made of probability densities. Then for every $\sigma \in \mathscr{P}(X)$, for a.e. $t \in (0,\infty)$ it holds*:

$$
\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}W_2^2(f_t\mathfrak{m},\sigma) = \int_X \varphi_t \Delta f_t \mathrm{d}\mathfrak{m}, \quad \text{for any Kantorovich potential } \varphi \text{ from } f_t \text{ to } \sigma. \tag{65}
$$

Proof Since $t \mapsto f_t\mathfrak{m}$ is an absolutely continuous curve w.r.t. W_2 (recall Theorem 6.10), the derivative at the left hand side of (65) exists for a.e. $t \in (0,\infty)$. Also, for a.e. $t \in (0,\infty)$ it holds $\lim_{h\to 0}\frac{1}{h}(f_{t+h} - f_t) = \Delta f_t$, the limit being understood in $L^2(X,\mathfrak{m})$.

Fix t_0 such that the derivative of the Wasserstein distance exists and the above limit holds and choose any Kantorovich potential φ_{t_0} for $(f_{t_0}\mathfrak{m},\sigma)$. We have

$$\frac{W_2^2(f_{t_0}\mathfrak{m},\sigma)}{2} = \int_X \varphi_{t_0} f_{t_0} \mathrm{d}\mathfrak{m} + \int \varphi_{t_0}^c \mathrm{d}\sigma$$
$$\frac{W_2^2(f_{t_0+h}\mathfrak{m},\sigma)}{2} \geq \int_X \varphi_{t_0} f_{t_0+h} \mathrm{d}\mathfrak{m} + \int \varphi_{t_0}^c \mathrm{d}\sigma.$$

Therefore, since $\varphi_{t_0} \in L^\infty(X,\mathfrak{m})$ we get

$$\frac{W_2^2(f_{t_0+h}\mathfrak{m},\sigma)}{2} - \frac{W_2^2(f_{t_0}\mathfrak{m},\sigma)}{2} \geq \int_X \varphi_{t_0}(f_{t_0+h} - f_{t_0})\mathrm{d}\mathfrak{m} = h\int_X \varphi_{t_0}\Delta f_{t_0} + o(h).$$

Dividing by $h<0$ and $h>0$ and letting $h\to 0$ we get the thesis. □

Proposition 8.19 (Derivative of the Entropy along a geodesic) *Let $(X,\mathsf{d},\mathfrak{m})$ be a strict $CD(K,\infty)$ space. Let $\mu_0,\mu_1\in\mathscr{P}(X)$, $\boldsymbol{\pi}\in\mathrm{GeoOpt}(\mu_0,\mu_1)$ and φ a Kantorovich potential for (μ_0,μ_1). Assume that $\boldsymbol{\pi}$ is a test plan and that $\mu_0\geq c\mathfrak{m}$ from some $c>0$ and denote by h_t the density of $\mu_t := (\mathrm{e}_t)_\sharp\boldsymbol{\pi}$. Then*

$$\varliminf_{t\downarrow 0}\frac{\mathrm{Ent}_\mathfrak{m}(\mu_t)-\mathrm{Ent}_\mathfrak{m}(\mu_0)}{t} \geq \varlimsup_{\varepsilon\downarrow 0}\frac{\mathsf{Ch}(\varphi)-\mathsf{Ch}(\varphi+\varepsilon h_0)}{\varepsilon}. \tag{66}$$

Proof The convexity of Ch ensures that the limit at the right hand side exists. From the fact that φ is Lipschitz, it is not hard to see that $h_0\notin D(\mathsf{Ch})$ implies $\mathsf{Ch}(\varphi+\varepsilon h_0)=+\infty$ for any $\varepsilon>0$ and in this case there is nothing to prove. Thus, we assume that $h_0\in D(\mathsf{Ch})$.

The convexity of $z\mapsto z\log z$ gives

$$\frac{\mathrm{Ent}_\mathfrak{m}(\mu_t)-\mathrm{Ent}_\mathfrak{m}(\mu_0)}{t} \geq \int_X \log h_0 \frac{h_t-h_0}{t}\mathrm{d}\mathfrak{m} = \int \frac{\log(h_0\circ e_t)-\log(h_0\circ e_0)}{t}\mathrm{d}\boldsymbol{\pi}. \tag{67}$$

Using the trivial inequality given by Taylor's formula

$$\log b-\log a \geq \frac{b-a}{a} - \frac{|b-a|^2}{2c^2},$$

valid for any $a,b\in[c,\infty)$, we obtain

$$\int \frac{\log(h_0\circ \mathrm{e}_t)-\log(h_0\circ \mathrm{e}_0)}{t}\mathrm{d}\boldsymbol{\pi} \geq \int \frac{h_0\circ e_t - h_0\circ \mathrm{e}_0}{t h_0\circ \mathrm{e}_0}\mathrm{d}\boldsymbol{\pi} - \frac{1}{2tc^2}\int |h_0\circ \mathrm{e}_t - h_0\circ \mathrm{e}_0|^2\mathrm{d}\boldsymbol{\pi}. \tag{68}$$

Taking into account Remark 4.4 and the fact that $|\dot\gamma_t| = \mathsf{d}(\gamma_0,\gamma_1)\leq \mathrm{diam}(X)$ for a.e. $t\in(0,1)$ and $\boldsymbol{\pi}$-a.e. γ, the last term in this expression can be bounded from above

by

$$\frac{1}{2tc^2}\int\left(\int_0^t \mathrm{diam}(X)|Dh_0|_w\circ \mathrm{e}_s\right)^2 \mathrm{d}s\mathrm{d}\boldsymbol{\pi} \le \frac{\mathrm{diam}(X)^2}{2c^2}\int\int_0^t |Dh_0|_w^2\circ \mathrm{e}_s \mathrm{d}s\mathrm{d}\boldsymbol{\pi}, \tag{69}$$

which goes to 0 as $t\to 0$.

Now let $S:\mathrm{Geo}(X)\to\mathbb{R}$ be the Borel function defined by $S(\gamma):=h_0\circ\gamma_0$ and define $\tilde{\boldsymbol{\pi}}:=\frac{1}{S}\boldsymbol{\pi}$. It is easy to check that $(\mathrm{e}_0)_\sharp\tilde{\boldsymbol{\pi}}=\mathfrak{m}$, so that in particular $\tilde{\boldsymbol{\pi}}$ is a probability measure. Also, the bound $h_0\ge c>0$ ensures that $\tilde{\boldsymbol{\pi}}$ is a test plan. By definition we have

$$\int \frac{h_0\circ e_t - h_0\circ \mathrm{e}_0}{t h_0\circ \mathrm{e}_0}\mathrm{d}\boldsymbol{\pi} = \int \frac{h_0\circ e_t - h_0\circ \mathrm{e}_0}{t}\mathrm{d}\tilde{\boldsymbol{\pi}}.$$

The latter equality and inequalities (67), (68) and (69) ensure that to conclude it is sufficient to show that

$$\varlimsup_{t\downarrow 0}\int \frac{h_0\circ e_t - h_0\circ \mathrm{e}_0}{t}\mathrm{d}\tilde{\boldsymbol{\pi}} \ge \varlimsup_{\varepsilon\downarrow 0}\frac{\mathsf{Ch}(\varphi)-\mathsf{Ch}(\varphi+\varepsilon h_0)}{\varepsilon}. \tag{70}$$

Here we apply the key Lemma 8.7. Observe that Theorem 7.11 ensures that

$$|D\varphi|_w(\gamma_0)=\lim_{t\downarrow 0}\frac{\varphi(\gamma_0)-\varphi(\gamma_t)}{t}=\mathsf{d}(\gamma_0,\gamma_1)$$

where the convergence is understood in $L^2(\boldsymbol{\pi})$. Thus the same holds for $L^2(\tilde{\boldsymbol{\pi}})$ and the hypotheses of Lemma 8.7 are satisfied with $\tilde{\boldsymbol{\pi}}$ as test plan, $g:=\varphi$ and $f:=h_0$. Equation (63) then gives

$$\begin{aligned}\varlimsup_{t\downarrow 0}\int \frac{h_0\circ e_t - h_0\circ \mathrm{e}_0}{t}\mathrm{d}\tilde{\boldsymbol{\pi}} &\ge \varlimsup_{\varepsilon\downarrow 0}\frac{1}{2}\int \frac{|D\varphi|_w^2(\gamma_0)-|D(\varphi+\varepsilon h_0)|_w^2(\gamma_0)}{\varepsilon}\mathrm{d}\tilde{\boldsymbol{\pi}}(\gamma)\\ &= \varlimsup_{\varepsilon\downarrow 0}\frac{1}{2}\int_X \frac{|D\varphi|_w^2(x)-|D(\varphi+\varepsilon h_0)|_w^2(x)}{\varepsilon}\mathrm{d}\mathfrak{m}(x),\end{aligned}$$

which concludes the proof. □

9 Riemannian Ricci Bounds

We say that $(X,\mathsf{d},\mathfrak{m})$ has *Riemannian Ricci curvature* bounded below by $K\in\mathbb{R}$ (in short, it is a $RCD(K,\infty)$ space) if any of the 3 equivalent conditions stated in the following theorem is true.

Theorem 9.12 *Let $(X,\mathsf{d},\mathfrak{m})$ be a compact and normalized metric measure space and $K\in\mathbb{R}$. The following three properties are equivalent.*

(i) $(X, \mathsf{d}, \mathfrak{m})$ *is a strict* $CD(K, \infty)$ *space* (*Definition* 7.11) *and the* L^2*-gradient flow of* Ch *is linear.*

(ii) $(X, \mathsf{d}, \mathfrak{m})$ *is a strict* $CD(K, \infty)$ *space* (*Definition* 7.11) *and Cheeger's energy is quadratic*, *i.e.*

$$2\big(\mathsf{Ch}(f) + \mathsf{Ch}(g)\big) = \mathsf{Ch}(f+g) + \mathsf{Ch}(f-g), \quad \forall f, g \in L^2(X, \mathfrak{m}). \tag{71}$$

(iii) $\mathrm{supp}(\mathfrak{m})$ *is geodesic and for any* $\mu \in D(\mathrm{Ent}_{\mathfrak{m}}) \subset \mathscr{P}(X)$ *there exists an* EVI_K*-gradient flow for* $\mathrm{Ent}_{\mathfrak{m}}$ *starting from* μ.

Proof (i) $\Rightarrow$ (ii). Since the heat semigroup P_t in $L^2(X, \mathfrak{m})$ is linear we obtain that Δ is a linear operator (i.e. its domain $D(\Delta)$ is a subspace of $L^2(X, \mathfrak{m})$ and $\Delta : D(\Delta) \to L^2(X, \mathfrak{m})$ is linear). Since $t \mapsto \mathsf{Ch}(P_t(f))$ is locally Lipschitz, tends to 0 as $t \to \infty$ and $\partial_t \mathsf{Ch}(P_t(f)) = -\|\Delta P_t(f)\|^2_{L^2}$ for a.e. $t > 0$ (see (22)), we have

$$\mathsf{Ch}(f) = \int_0^\infty \big\|\Delta P_t(f)\big\|^2_{L^2(X,\mathfrak{m})} \mathrm{d}t.$$

Therefore Ch, being an integral of quadratic forms, is a quadratic form. Specifically, for any $f, g \in L^2(X, \mathfrak{m})$ it holds

$$\begin{aligned}
&\mathsf{Ch}(f+g) + \mathsf{Ch}(f-g) \\
&= \int_0^\infty \big\|\Delta P_t(f+g)\big\|^2_{L^2(X,\mathfrak{m})} + \big\|\Delta P_t(f-g)\big\|^2_{L^2(X,\mathfrak{m})} \mathrm{d}t \\
&= \int_0^\infty \big\|\Delta P_t(f) + \Delta P_t(g)\big\|^2_{L^2(X,\mathfrak{m})} + \big\|\Delta P_t(f) - \Delta P_t(g)\big\|^2_{L^2(X,\mathfrak{m})} \mathrm{d}t \\
&= \int_0^\infty 2\big\|\Delta P_t(f)\big\|^2_{L^2(X,\mathfrak{m})} + 2\big\|\Delta P_t(g)\big\|^2_{L^2(X,\mathfrak{m})} \mathrm{d}t \\
&= 2\mathsf{Ch}(f) + 2\mathsf{Ch}(g).
\end{aligned}$$

(ii) $\Rightarrow$ (iii). By [31, Remark 4.6(iii)] $(\mathrm{supp}(\mathfrak{m}), \mathsf{d})$ is a length space and therefore it is also geodesic, since X is compact.

Thanks to Remark 2.1 it is sufficient to prove that a gradient flow in the EVI_K sense exists for an initial datum $\mu_0 \ll \mathfrak{m}$ with density bounded away from 0 and infinity. Let f_0 be this density, (f_t) the heat flow starting from it and recall that from the maximum principle 4.11 we know that the f_t's are far from 0 and infinity as well for any $t > 0$. Fix a reference probability measure σ with density bounded away from 0 and infinity as well. For any $t \geq 0$ pick a test plan $\boldsymbol{\pi}_t$ optimal for $(f_t\mathfrak{m}, \sigma)$. Define $\sigma_t^s := (\mathrm{e}_s)_\sharp \pi_t$.

We claim that for a.e. $t \in (0, \infty)$ it holds

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}W_2^2(f_t\mathfrak{m}, \sigma\mathfrak{m}) \leq \varliminf_{s\downarrow 0} \frac{\mathrm{Ent}_{\mathfrak{m}}(\sigma_t^s) - \mathrm{Ent}_{\mathfrak{m}}(\sigma_t^0)}{s}. \tag{72}$$

Let φ_t be a Kantorovich potential for $f_t\mathfrak{m}, \sigma\mathfrak{m}$. By Proposition 8.18 we know that for a.e. $t \in (0,\infty)$ it holds

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}W_2^2(f_t\mathfrak{m},\sigma\mathfrak{m}) = \int_X \varphi \Delta f_t \mathrm{d}\mathfrak{m} \le \lim_{\varepsilon\downarrow 0}\frac{\mathsf{Ch}(f_t-\varepsilon\varphi_t)-\mathsf{Ch}(f_t)}{\varepsilon},$$

while from Proposition 8.19 we have that for any $t>0$ it holds

$$\lim_{s\downarrow 0}\frac{\mathrm{Ent}_{\mathfrak{m}}(\sigma_t^s)-\mathrm{Ent}_{\mathfrak{m}}(\sigma_t^0)}{s} \ge \lim_{\varepsilon\downarrow 0}\frac{\mathsf{Ch}(\varphi_t)-\mathsf{Ch}(\varphi_t+\varepsilon f_t)}{\varepsilon}.$$

Here we use the fact that Ch is quadratic. Indeed in this case simple algebraic manipulations show that

$$\frac{\mathsf{Ch}(f_t-\varepsilon\varphi_t)-\mathsf{Ch}(f_t)}{\varepsilon} = \frac{\mathsf{Ch}(\varphi_t)-\mathsf{Ch}(\varphi_t+\varepsilon f_t)}{\varepsilon} + O(\varepsilon), \quad \forall t>0,$$

and therefore (72) is proved.

Now notice that the K-convexity of the entropy yields

$$\lim_{s\downarrow 0}\frac{\mathrm{Ent}_{\mathfrak{m}}(\sigma_t^s)-\mathrm{Ent}_{\mathfrak{m}}(\sigma_t^0)}{s} \le \mathrm{Ent}_{\mathfrak{m}}(\sigma)-\mathrm{Ent}_{\mathfrak{m}}(f_t\mathfrak{m})-\frac{K}{2}W_2^2(f_t\mathfrak{m},\sigma),$$

and therefore we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}W_2^2(f_t\mathfrak{m},\sigma\mathfrak{m})+\mathrm{Ent}_{\mathfrak{m}}(f_t\mathfrak{m})+\frac{K}{2}W_2^2(f_t\mathfrak{m},\sigma)\le \mathrm{Ent}_{\mathfrak{m}}(\sigma),$$
$$\text{for a.e. } t\in(0,\infty).$$

By Proposition 2.1 we conclude.

(iii) $\Rightarrow$ (i). Since $(\mathrm{supp}(\mathfrak{m}),\mathsf{d})$ is geodesic, so is $(\overline{D(\mathrm{Ent}_{\mathfrak{m}})}, W_2)$, which together with existence of EVI_K-gradient flows for $\mathrm{Ent}_{\mathfrak{m}}$ yields, via Proposition 2.3, K-geodesic convexity of $\mathrm{Ent}_{\mathfrak{m}}$ along all geodesics in $\overline{D(\mathrm{Ent}_{\mathfrak{m}})}$. In particular, $(X,\mathsf{d},\mathfrak{m})$ is a strict $CD(K,\infty)$ space.

We turn to the linearity. Let (μ_t^0), (μ_t^1) be two EVI_K-gradient flows of the relative entropy and, for $\lambda\in(0,1)$ fixed, define $\mu_t^\lambda := (1-\lambda)\mu_t^0+\lambda\mu_t^1$.

We claim that (μ_t) is an EVI_K-gradient flow of $\mathrm{Ent}_{\mathfrak{m}}$. To prove this, fix $\nu\in\mathscr{P}(X)$, $t>0$ and an optimal plan $\boldsymbol{\gamma}\in\mathrm{Opt}(\mu_t^\lambda,\nu)$. Since $\mu_t^i\ll\mu_t^\lambda=\pi^1_\sharp\boldsymbol{\gamma}$ for $i=0,1$ we can define, as in Definition 5.10, the plans $\boldsymbol{\gamma}_{\mu_t^i}\in\mathscr{P}(X^2)$ and the measures $\nu^i:=\boldsymbol{\gamma}_\sharp\mu_t^i$, $i=0,1$. Since $\mathrm{supp}(\gamma_{\mu_t^i})\subset\mathrm{supp}(\boldsymbol{\gamma})$, we have that $\gamma_{\mu_t^i}\in\mathrm{Opt}(\mu_t^i,\nu^i)$, therefore from $\boldsymbol{\gamma}=(1-\lambda)\boldsymbol{\gamma}_{\mu_t^0}+\lambda\boldsymbol{\gamma}_{\mu_t^1}$ we deduce

$$W_2^2\big(\mu_t^\lambda,\nu\big)=(1-\lambda)W_2^2\big(\mu_t^0,\nu^0\big)+\lambda W_2^2\big(\mu_t^1,\nu^1\big). \tag{73}$$

On the other hand, from the convexity of the squared Wasserstein distance we immediately get that

$$W_2^2\big(\mu_{t+h}^\lambda,\nu\big)\le(1-\lambda)W_2^2\big(\mu_{t+h}^0,\nu^0\big)+\lambda W_2^2\big(\mu_{t+h}^1,\nu^1\big), \quad \forall h>0. \tag{74}$$

Furthermore, recalling (iii) of Proposition 5.15, we get

$$\begin{aligned}\mathrm{Ent}_{\mathfrak{m}}\big(\mu_t^\lambda\big)\mathfrak{m} - \mathrm{Ent}_{\mathfrak{m}}(\nu) \le (1-\lambda)\big(\mathrm{Ent}_{\mathfrak{m}}\big(\mu_t^0\big) - \mathrm{Ent}_{\mathfrak{m}}\big(\nu^0\big)\big)\\ + \lambda\big(\mathrm{Ent}_{\mathfrak{m}}\big(\mu_t^1\big) - \mathrm{Ent}_{\mathfrak{m}}\big(\nu^1\big)\big).\end{aligned} \tag{75}$$

The fact that (μ_t^0) and (μ_t^1) are EVI_K-gradient flows for $\mathrm{Ent}_{\mathfrak{m}}$ (see in particular the characterization (iii) given in Proposition 2.1) in conjunction with (73), (74) and (75) yield

$$\varlimsup_{h\downarrow 0} \frac{W_2^2(\mu_{t+h}^\lambda, \nu) - W_2^2(\mu_t^\lambda, \nu)}{2} + \frac{K}{2} W_2^2\big(\mu_t^\lambda, \nu\big) + \mathrm{Ent}_{\mathfrak{m}}\big(\mu_t^\lambda\big) \le \mathrm{Ent}_{\mathfrak{m}}(\nu). \tag{76}$$

Since $t > 0$ and $\nu \in \mathscr{P}(X)$ were arbitrary, we proved that (μ_t^λ) is a EVI_K-gradient flow of $\mathrm{Ent}_{\mathfrak{m}}$ (see again (iii) of Proposition 2.1).

Thus, recalling the identification of gradient flows, we proved that the L^2-heat flow is additive in $D(\mathrm{Ent}_{\mathfrak{m}})$. Since the heat flow in $L^2(X, \mathfrak{m})$ commutes with additive and multiplicative constants, it is easy to get from this linearity in the class of bounded functions. By L^2 contractivity, linearity extends to the whole of $L^2(X, \mathfrak{m})$. □

We conclude by discussing some basic properties of the spaces with Riemannian Ricci curvature bounded from below.

We start observing that Riemannian manifolds with Ricci curvature bounded below by K are $RCD(K, \infty)$ spaces, as they are non branching $CD(K, \infty)$ spaces and the heat flow is linear on them. Also, from the studies made in [25, 27, 33] and [16] we also know that finite dimensional Alexandrov spaces with curvature bounded from below are $RCD(K, \infty)$ spaces as well. On the other side, Finsler manifolds are ruled out, as it is known (see for instance [26]) that the heat flow is linear on a Finsler manifold if and only if the manifold is Riemannian.

The stability of the $RCD(K, \infty)$ notion can be deduced by the stability of EVI_K-gradient flows w.r.t. Γ-convergence of functionals, which is an easy consequence of the integral formulation in (ii) of Proposition 2.1.

Hence $RCD(K, \infty)$ spaces have the same basic properties of $CD(K, \infty)$ spaces, which gives to this notion the right of being called a synthetic (or weak) notion of Ricci curvature bound.

The point is then to understand the additional analytic/geometric properties of these spaces, which come mainly by the addition of linearity condition. A first consequence is that the heat flow contracts, up to an exponential factor, the distance W_2, i.e.

$$W_2(\mu_t, \nu_t) \le e^{-Kt} W_2(\mu_0, \nu_0), \quad \forall t \ge 0,$$

whenever $(\mu_t), (\nu_t) \subset \mathscr{P}_2(X)$ are gradient flows of the entropy.

By a duality argument (see [6, 15, 21]), this property implies the Bakry-Emery gradient estimate

$$\big|D\mathsf{h}_t(f)\big|_w^2(x) \le e^{-2Kt}\mathsf{h}_t\big(|Df|_w^2\big)(x), \quad \text{for } \mathfrak{m}\text{-a.e. } x \in X,$$

for all $t > 0$, where $\mathsf{h}_t : L^2(X, \mathfrak{m}) \to L^2(X, \mathfrak{m})$ is the heat flow seen as gradient flow of Ch. If $(X, \mathsf{d}, \mathfrak{m})$ is doubling and supports a local Poincaré inequality, then also the Lipschitz regularity of the heat kernel is deduced (following an argument described in [15]).

Also, since in $RCD(K, \infty)$ spaces Ch is a quadratic form, if we define

$$\mathcal{E}(f, g) := \mathsf{Ch}(f + g) - \mathsf{Ch}(f) - \mathsf{Ch}(g), \quad \forall f, g \in W^{1,2}(X, \mathsf{d}, \mathfrak{m}),$$

we get a closed Dirichlet form on $L^2(X, \mathfrak{m})$ (closure follows from the L^2-lower semicontinuity of Ch). Hence it is natural to compare the calculus on $RCD(K, \infty)$ spaces with the abstract one available for Dirichlet forms (see [11]). The picture here is pretty clear and consistent. Recall that to any $f \in D(\mathcal{E})$ one can associate the energy measure $[f]$ defined by

$$[f](\varphi) := -\mathcal{E}(f, f\varphi) + \mathcal{E}\big(f^2/2, \varphi\big).$$

Then it is possible to show that the energy measure coincides with $|Df|_*^2\mathfrak{m}$. Also, the distance d coincides with the intrinsic distance $\mathsf{d}_{\mathcal{E}}$ induced by the form, defined by

$$\mathsf{d}_{\mathcal{E}}(x, y) := \sup\Big\{\big|g(x) - g(y)\big| : g \in D(\mathcal{E}) \cap C(X), [g] \le \mathfrak{m}\Big\}.$$

Taking advantage of these identification and of the locality of $\mathcal{E}$ (which is a consequence of the locality of the notion $|Df|_*$), one can also see that on $RCD(K, \infty)$ spaces a continuous Brownian motion with continuous sample paths associated to h_t exists and is unique.

Finally, for $RCD(K, \infty)$ spaces it is possible to prove tensorization and globalization properties which are in line with those available for $CD(K, \infty)$ spaces.

Acknowledgements The authors acknowledge the support of the ERC ADG GeMeThNES and the PRIN08-grant from MIUR for the project *Optimal transport theory, geometric and functional inequalities, and applications*.

The authors also thank A. Mondino for his careful reading of a preliminary version of this manuscript.

References

1. Ambrosio, L., Gigli, N.: User's guide to optimal transport theory. In: Piccoli, B., Poupaud, F. (Eds.) The CIME Lecture Notes in Mathematics (2011, to appear)
2. Ambrosio, L., Gigli, N., Mondino, A., Rajala, T.: Riemannian Ricci curvature lower bounds in metric measure spaces with σ-finite measure, 1–38. arXiv:1207.4924 (2012)
3. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures. Lectures in Mathematics ETH Zürich, 2nd edn. Birkhäuser, Basel (2008)
4. Ambrosio, L., Gigli, N., Savaré, G.: Calculus and heat flow in metric measure spaces and applications to spaces with Ricci bounds from below, 1–74. arXiv:1106.2090 (2011)
5. Ambrosio, L., Gigli, N., Savaré, G.: Density of Lipschitz functions and equivalence of weak gradients in metric measure spaces, 1–28. arXiv:1111.3730 (2011)

6. Ambrosio, L., Gigli, N., Savaré, G.: Metric measure spaces with Riemannian Ricci curvature bounded from below, 1–60. arXiv:1109.0222 (2011)
7. Ambrosio, L., Rajala, T.: Slopes of Kantorovich potentials and existence of optimal transport maps in metric measure spaces. Ann. Mat. Pura Appl. (2012, to appear)
8. Brézis, H.: Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert. North-Holland Mathematics Studies, vol. 5. North-Holland, Amsterdam (1973). Notas de Matemática, 50
9. Cheeger, J.: Differentiability of Lipschitz functions on metric measure spaces. Geom. Funct. Anal. **9**, 428–517 (1999)
10. Daneri, S., Savaré, G.: Eulerian calculus for the displacement convexity in the Wasserstein distance. SIAM J. Math. Anal. **40**, 1104–1122 (2008)
11. Fukushima, M.: Dirichlet Forms and Markov Processes. North-Holland Mathematical Library, vol. 23. North-Holland, Amsterdam (1980)
12. Gigli, N.: On the heat flow on metric measure spaces: existence, uniqueness and stability. Calc. Var. Partial Differ. Equ. **39**, 101–120 (2010)
13. Gigli, N.: On the differential structure of metric measure spaces and applications, 1–86. arXiv:1205.6622 (2012)
14. Gigli, N.: Optimal maps in non branching spaces with Ricci curvature bounded from below. Geom. Funct. Anal. **22**(4), 990–999 (2012)
15. Gigli, N., Kuwada, K., Ohta, S.: Heat flow on Alexandrov spaces. Comm. Pure Appl. Math. (2012). doi:10.1002/cpa.21431
16. Gigli, N., Ohta, S.-I.: First variation formula in Wasserstein spaces over compact Alexandrov spaces. Can. Math. Bull. **55**(4), 723–735 (2012)
17. Heinonen, J.: Nonsmooth calculus. Bull. Am. Math. Soc. (N.S.) **44**, 163–232 (2007)
18. Heinonen, J., Koskela, P.: Quasiconformal maps in metric spaces with controlled geometry. Acta Math. **181**, 1–61 (1998)
19. Heinonen, J., Koskela, P.: A note on Lipschitz functions, upper gradients, and the Poincaré inequality. N.Z. J. Math. **28**, 37–42 (1999)
20. Koskela, P., MacManus, P.: Quasiconformal mappings and Sobolev spaces. Stud. Math. **131**, 1–17 (1998)
21. Kuwada, K.: Duality on gradient estimates and Wasserstein controls. J. Funct. Anal. **258**, 3758–3774 (2010)
22. Lisini, S.: Characterization of absolutely continuous curves in Wasserstein spaces. Calc. Var. Partial Differ. Equ. **28**, 85–120 (2007)
23. Lott, J., Villani, C.: Ricci curvature for metric-measure spaces via optimal transport. Ann. Math. (2) **169**, 903–991 (2009)
24. Ohta, S.-I.: Finsler interpolation inequalities. Calc. Var. Partial Differ. Equ. **36**, 211–249 (2009)
25. Ohta, S.-I.: Gradient flows on Wasserstein spaces over compact Alexandrov spaces. Am. J. Math. **131**, 475–516 (2009)
26. Ohta, S.-I., Sturm, K.-T.: Heat flow on Finsler manifolds. Commun. Pure Appl. Math. **62**, 1386–1433 (2009)
27. Petrunin, A.: Alexandrov meets Lott–Villani–Sturm. arXiv:1003.5948v1 (2010)
28. Rajala, T.: Improved geodesics for the reduced curvature-dimension condition in branching metric spaces. Discrete Contin. Dyn. Syst. (2011, to appear)
29. Savaré, G.: Gradient flows and evolution variational inequalities in metric spaces (2010, in preparation)
30. Shanmugalingam, N.: Newtonian spaces: an extension of Sobolev spaces to metric measure spaces. Rev. Mat. Iberoam. **16**, 243–279 (2000)
31. Sturm, K.-T.: On the geometry of metric measure spaces. I. Acta Math. **196**, 65–131 (2006)
32. Villani, C.: Optimal Transport. Old and New. Grundlehren der Mathematischen Wissenschaften, vol. 338. Springer, Berlin (2009)
33. Zhang, H.-C., Zhu, X.-P.: Ricci curvature on Alexandrov spaces and rigidity theorems. Commun. Anal. Geom. **18**, 503–553 (2010)

Spaces of Finite Element Differential Forms

Douglas N. Arnold

Abstract We discuss the construction of finite element spaces of differential forms which satisfy the crucial assumptions of the finite element exterior calculus, namely that they can be assembled into subcomplexes of the de Rham complex which admit commuting projections. We present two families of spaces in the case of simplicial meshes, and two other families in the case of cubical meshes. We make use of the exterior calculus and the Koszul complex to define and understand the spaces. These tools allow us to treat a wide variety of situations, which are often treated separately, in a unified fashion.

1 Introduction

The gradient, curl, and divergence are the most fundamental operators of vector calculus, appearing throughout the differential equations of mathematical physics and other applications. The finite element solution of such equations requires finite element subspaces of the natural Hilbert space domains of these operators, namely H^1, $H(\mathrm{curl})$, and $H(\mathrm{div})$. The construction of subspaces with desirable properties has been an active research topic for half a century. Exterior calculus provides a framework in which these fundamental operators and spaces are unified and generalized, and their properties and inter-relations clarified. Each of the operators is viewed as a particular case of the exterior derivative operator $d = d^k$ taking differential k-forms on some domain $\Omega \subset \mathbb{R}^n$ to differential $(k+1)$-forms. We regard d^k as an unbounded operator between the Hilbert spaces $L^2\Lambda^k$ and $L^2\Lambda^{k+1}$ consisting of differential forms with L^2 coefficients. The domain of d^k is the Hilbert space

$$H\Lambda^k = \left\{ u \in L^2\Lambda^k \mid du \in L^2\Lambda^{k+1} \right\}, \tag{1.1}$$

In memory of Enrico Magenes, in gratitude for his deep and elegant mathematics, which taught us, and his profound humanity, which inspired us.

The work of the author was supported by NSF grant DMS-1115291.

D.N. Arnold (✉)
School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA
e-mail: arnold@umn.edu
url: http://www.ima.umn.edu/~arnold

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*,
Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_9,

and all the d^k and their domains combine to form the L^2 de Rham complex

$$0 \to H\Lambda^0 \xrightarrow{d^0} H\Lambda^1 \xrightarrow{d^1} \cdots \xrightarrow{d^{n-1}} H\Lambda^n \to 0.$$

Differential 0-forms and n-forms may be identified simply with functions on Ω and differential 1-forms and $(n-1)$-forms may be identified with vector fields. In three dimensions, we may use these proxies to write the de Rham complex as

$$0 \to H^1 \xrightarrow{\text{grad}} H(\text{curl}) \xrightarrow{\text{curl}} H(\text{div}) \xrightarrow{\text{div}} L^2 \to 0.$$

The finite element exterior calculus (FEEC) is a theory developed in the last decade [1, 5, 6] which enables the development and analysis of finite element spaces of differential forms. One major part of FEEC is carried out in the framework of Hilbert complexes, of which the L^2 de Rham complex is the most canonical example. One important outcome of FEEC is the realization that the finite dimensional subspaces $\Lambda_h^k \subset H\Lambda^k$ used in Galerkin discretizations of a variety of differential equations involving differential k-forms should satisfy two basic assumptions, beyond the obvious requirement that the spaces have good approximation properties. The first assumption is that the subspaces form a *subcomplex* of the de Rham complex, i.e., that $d\Lambda_h^k \subset \Lambda_h^{k+1}$. The second is that there exist projection operators π_h^k from $H\Lambda^k$ to Λ_h^k which commute with d in the sense that the following diagram commutes:

$$\begin{array}{ccccccccc}
H\Lambda^0 & \xrightarrow{d} & H\Lambda^1 & \xrightarrow{d} & \cdots & \xrightarrow{d} & H\Lambda^{n-1} & \xrightarrow{d} & H\Lambda^n \\
\downarrow{\scriptstyle \pi_h^0} & & \downarrow{\scriptstyle \pi_h^1} & & & & \downarrow{\scriptstyle \pi_h^{n-1}} & & \downarrow{\scriptstyle \pi_h^n} \\
\Lambda_h^0 & \xrightarrow{d} & \Lambda_h^1 & \xrightarrow{d} & \cdots & \xrightarrow{d} & \Lambda_h^{n-1} & \xrightarrow{d} & \Lambda_h^n
\end{array}$$

The second major part of FEEC, into which the present exposition falls, is concerned with the construction of specific finite element spaces Λ_h^k of differential forms. A special role is played by two families of finite element spaces $\mathcal{P}_r^-\Lambda^k(\mathcal{T}_h)$ and $\mathcal{P}_r\Lambda^k(\mathcal{T}_h)$, defined for any dimension n, any simplicial mesh $\mathcal{T}_h$, any polynomial degree $r \geq 1$, and any form degree $0 \leq k \leq n$. Both these spaces are subspaces of $H\Lambda^k(\Omega)$. The $\mathcal{P}_r^-\Lambda^k$ spaces with increasing k and constant r form a subcomplex of L^2 de Rham complex which admits commuting projections. The same is true of the $\mathcal{P}_r\Lambda^k$ family, except in that case the polynomial degree r decreases as the form degree k increases.

We also discuss cubical meshes. In this case, there is a well-known family of elements, denoted by $\mathcal{Q}_r^-\Lambda^k$ in our notation, obtained by a tensor product construction. As for the $\mathcal{P}_r^-\Lambda^k$ family, the $\mathcal{Q}_r^-\Lambda^k$ spaces with constant degree r combine to form a de Rham subcomplex with commuting projections. We also discuss a recently discovered second family on cubical meshes, the $\mathcal{S}_r\Lambda^k$ family of [3]. Like the $\mathcal{P}_r\Lambda^k$ family, the de Rham subcomplexes for this family are obtained with decreasing degree. Moreover for large r, the $\dim \mathcal{S}_r\Lambda^k(\mathcal{T}_h)$ is much smaller dimension

than $\dim \mathcal{Q}_r^- \Lambda^k$. The finite element subspaces of H^1, $H(\mathrm{curl})$, and $H(\mathrm{div})$ from this family in three dimensions are new.

The remainder of the paper is organized as follows. In the next section we cover some preliminary material (which the more expert reader may wish to skip). We recall the construction of finite element spaces from spaces of shape functions and unisolvent degrees of freedom. To illustrate we discuss the Lagrange elements and carry out the proof of unisolvence in a manner that will guide our treatment of differential form spaces of higher degree. We also give a brief summary of those aspects of exterior calculus most relevant to us. In Sect. 3 we discuss the two primary families of finite element spaces for differential forms on simplicial meshes mentioned above. A key role is played by the Koszul complex, which is introduced in this section. Then, in Theorem 3.5, we give a proof of unisolvence for the $\mathcal{P}_r^-$ family which we believe to be simpler than has appeared heretofore (a similar proof could be given for the $\mathcal{P}_r$ family as well). In the final section we review the two families mentioned for cubical meshes, including a description, without proofs, of the recently discovered $\mathcal{S}_r$ family.

2 Preliminaries

2.1 The Assembly of Finite Element Spaces

Recalling the definition of a finite element space [11], we assume that the domain $\Omega \subset \mathbb{R}^n$ is triangulated by finite elements, i.e., its closure is the union of a finite set $\mathcal{T}_h$ of closed convex polyhedral elements with nonempty interiors such that the intersection of any two elements is either empty or is a common face of each of some dimension. We denote by $\Delta_d(T)$ the set of faces of T of dimensions d, so, for example, $\Delta_0(T)$ is the set of vertices of T, and $\Delta_n(T)$ is the singleton set whose only element is T. We also define $\Delta(T) = \bigcup_{0\le d\le n} \Delta_d(T)$, the set of all faces of T. In this paper we consider the two cases of simplicial elements, in which each element T of the triangulation is an n-simplex, and cubical elements, in which element is an n-box (i.e., the Cartesian product of n intervals). To define a finite element space $\Lambda_h^k \subset H\Lambda^k(\Omega)$, we must supply, for each element $T \in \mathcal{T}_h$,

(1) A finite dimensional space $V(T)$, called the space of *shape functions*, consisting of differential k-forms on T with polynomial coefficients. The finite element space will consist of functions u which belong to the shape function spaces piecewise in the sense that $u|_T \in V(T)$ for all $T \in \mathcal{T}_h$ (allowing the possibility that u is multiply-valued on faces of dimension $< n$).
(2) A set of functionals $V(T) \to \mathbb{R}$, called the *degrees of freedom*, which are *unisolvent* (i.e., which form a basis for the dual space $V(T)^*$) and such that each degree of freedom is associated to a specific face of $f \in \Delta(T)$.

It is assumed that when two distinct elements T_1 and T_2 intersect in a common face f, the degrees of freedom of T_1 and T_2 which are associated to f are in a specific 1-to-1 correspondence. If u is a function which belongs to the shape function

spaces piecewise, then we say that the degrees of freedom are single-valued on u if whenever two elements $T_1 \neq T_2$ meet in a common face, then the corresponding degrees of freedom associated to the face take the same value on $u|_{T_1}$ and $u|_{T_2}$, respectively. With these ingredients, the finite element space Λ_h^k associated to the choice of triangulation $\mathcal{T}_h$, the shape function spaces $V(T)$, and the degrees of freedom, is defined as the set of all k-forms on Ω which belong to the shape function spaces piecewise and for which all the degrees of freedom are single-valued.

The choice of the degrees of freedom associated to faces of dimension $d < n$ determine the interelement continuity imposed on the finite element subspace. The use of degrees of freedom to specify the continuity, rather than imposing the continuity a priori in the definition of the finite element space, is of great practical significance in that it assures that the finite element space can be implemented efficiently. The dimension of the space is known (it is just the sum over the faces of the triangulation of the number of degrees of freedom associated to the face) and it depends only on the topology of the triangulation, not on the coordinates of the element vertices. Moreover, the degrees of freedom lead to a computable basis for Λ_h^k in which each basis element is associated to one degree of freedom. Further, the basis is local, in that the basis element for a degree of freedom associated to a face f is nonzero only on the elements that contain f.

The finite element space so defined does not depend on the specific choice of degrees of freedom in $V(T)^*$, but only on the span of the degrees of freedom associated to each face f of T, and we shall generally specify only the span, rather than a specific choice of basis for it.

2.2 The Lagrange Finite Element Family

To illustrate these definitions and motivate the constructions for differential forms, we consider the simplest example, the Lagrange family of finite element subspaces of $H^1 = H\Lambda^0$. The Lagrange space, which we denote $\mathcal{P}_r\Lambda^0(\mathcal{T}_h)$ in anticipation of its generalization below, is defined for any simplicial triangulation $\mathcal{T}_h$ in $\mathbb{R}^n$ and any polynomial degree $r \geq 1$. The shape function space is $V(T) = \mathcal{P}_r(T)$, the space of all polynomial functions on T of degree at most r. For a face f of T of dimension d, the span of the associated degrees of freedom are the functionals

$$u \in \mathcal{P}_r(T) \mapsto \int_f (\operatorname{tr}_f u)q, \quad q \in \mathcal{P}_{r-d-1}(f), \ f \in \Delta(T). \tag{2.1}$$

In interpreting this, we understand the space $\mathcal{P}_s(f)$ to be the space $\mathbb{R}$ of constants if f is 0-dimensional (a single vertex) and $s \geq 0$. Also the space $\mathcal{P}_s(f) = 0$ if $s < 0$ and f is arbitrary. The notation $\operatorname{tr}_f u$ denotes the trace of u on f, i.e., its restriction. Thus there is one degree of freedom associated to each vertex v, namely the evaluation functional $u \mapsto u(v)$. For $r \geq 2$ there are also degrees of freedom associated to

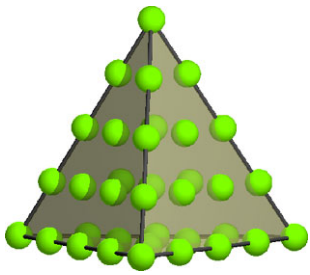


Fig. 1 Degrees of freedom for the Lagrange quartic space $\mathcal{P}_4\Lambda^0$ in 3 dimensions

the edges e of T, namely the moments of u on the edge of degree at most $r-2$:

$$u \mapsto \int_e (\operatorname{tr}_e u)q, \quad q \in \mathcal{P}_{r-2}(e).$$

For $r \geq 3$ there are degrees of freedom associated to the 2-faces, namely moments of degree at most $r-3$, etc. This is often indicated in a degree of freedom diagram, like that of Fig. 1, in which the number of symbols drawn in the interior of a face is equal to the number of degrees of freedom associated to the face.

A requirement of the definition of a finite element space is that the degrees of freedom be unisolvent. We present the proof for Lagrange elements in detail, since it will guide us when it comes to verifying unisolvence for more complicated spaces.

Theorem 2.1 (Unisolvence for the Lagrange elements) *For any $r \geq 1$ and any n-simplex T, the degrees of freedom (2.1) are unisolvent on $V(T) = \mathcal{P}_r(T)$.*

Proof It suffices to verify, first, that the number of degrees of freedom proposed for T does not exceed $\dim V(T)$, and, second, that if all the degrees of freedom vanish when applied to some $u \in V(T)$, then $u \equiv 0$. For the first claim, we have by (2.1) that the total number of degrees of freedom is at most

$$\sum_{d=0}^{n} \#\Delta_d(T) \dim \mathcal{P}_{r-d-1}(\mathbb{R}^d) = \sum_{d=0}^{n} \binom{n+1}{d+1}\binom{r-1}{d} = \binom{n+r}{n} = \dim \mathcal{P}_r(T),$$

where the second equality is a binomial identity which comes from expanding in the equation $(1+x)^{n+1}(1+x)^{r-1} = (1+x)^{n+r}$ and comparing the coefficients of x^n on both sides.

We prove the second claim by induction on the dimension n, the case $n = 0$ being trivial. Suppose that $u \in \mathcal{P}_r(T)$ for some simplex T of dimension n and that all the degrees of freedom in (2.1) vanish. We wish to show that u vanishes. Let $F \in \Delta_{n-1}(T)$ be a facet of T, and consider $\operatorname{tr}_F u$, which is a polynomial function of at most degree r on the $(n-1)$-dimensional simplex F, i.e., it belongs to $\mathcal{P}_r(F)$. Moreover, if we replace T by F and u by $\operatorname{tr}_F u$ in (2.1), the resulting functionals vanish by assumption (using the obvious fact that $\operatorname{tr}_f \operatorname{tr}_F u = \operatorname{tr}_f u$ for $f \subset F \subset T$). By induction we conclude that $\operatorname{tr}_F u$ vanishes on all the facets F of T. Therefore, u is divisible by the barycentric coordinate function λ_i which vanishes on F, and,

since this holds for all facets, $u = (\prod_{i=0}^{n} \lambda_i) p$ for some $p \in \mathcal{P}_{r-n-1}(T)$. Taking $f = T$ and $q = p$ in (2.1) we conclude that

$$\int_T \left(\prod_{i=0}^{n} \lambda_i \right) p^2 = 0,$$

which implies that p vanishes on T, and so u does as well. □

Let us note some features of the proof, which will be common to the unisolvence proofs for all of the finite element spaces we discuss here. After a dimension count to verify that the proposed degrees of freedom are correct in number, or at least no more than required, the proof proceeded by induction on the number of space dimensions. The inductive step relied on a trace property of the shape function space $V(T) = \mathcal{P}_r(T)$ for the family, namely that $\operatorname{tr}_F V(T) \subset V(F)$. Moreover, it used a similar trace property for the degrees of freedom: if $\xi_F \in V(F)^*$ is a degree of freedom for $V(F)$, then the pullback $\xi_F \circ \operatorname{tr}_F \in V(T)^*$ is a degree of freedom for $V(T)$. The induction reduced the unisolvence proof to verifying that if $u \in \mathring{V}(T)$, the space of functions in $V(T)$ whose trace vanishes on the entire boundary, and if the interior degrees of freedom (those associated to T itself) of u vanish, then u itself vanishes, which we showed by explicit construction.

Finally, we note that the continuity implied by the degrees of freedom is exactly what is required to insure that the Lagrange finite element space is contained in H^1:

$$\mathcal{P}_r \Lambda^0(\mathcal{T}_h) = \left\{ u \in H^1(\Omega) \mid u \text{ belongs to } \mathcal{P}_r(T) \text{ piecewise} \right\}. \tag{2.2}$$

Indeed, a piecewise smooth function belongs to $H^1(\Omega)$ if and only if its traces on faces are single-valued. Thus if a function in $H^1(\Omega)$ belongs piecewise to $\mathcal{P}_r(T)$, its traces are single-valued, so the degrees of freedom are single-valued, and the function belongs to $\mathcal{P}_r \Lambda^0(\mathcal{T}_h)$. On the other hand, if the function belongs to $\mathcal{P}_r \Lambda^0(\mathcal{T}_h)$, its traces on faces are single-valued, since, as we saw in the course of the unisolvence proof, they are determined by the degrees of freedom. Thus the function belongs to $H^1(\Omega)$.

2.3 Exterior Calculus

For the convenience of readers less familiar with differential forms and exterior calculus we now briefly review key definitions and properties. We begin with the space of *algebraic* k-forms on V: $\operatorname{Alt}^k V = \{L : V^k \to \mathbb{R} \mid k\text{-linear, skew-symmetric}\}$, where the multilinear form L is skew-symmetric, or alternating, if it changes sign under the interchange of any two of its arguments. The skew-symmetry condition is vacuous if $k < 2$, so $\operatorname{Alt}^1 V = V^*$ and, by convention, $\operatorname{Alt}^0 V = \mathbb{R}$. If ω is any k-linear map $V^k \to \mathbb{R}$, then $\operatorname{skw} \omega \in \operatorname{Alt}^k V$ where

$$(\operatorname{skw} \omega)(v_1, \ldots, v_k) = \frac{1}{k!} \sum_{\sigma} \operatorname{sign}(\sigma) \omega(v_{\sigma_1}, \ldots, v_{\sigma_k}),$$

with the sum taken over all the permutations of the integers 1 to k. The wedge product $\mathrm{Alt}^k V \times \mathrm{Alt}^l V \to \mathrm{Alt}^{k+l} V$ is defined

$$\omega \wedge \mu = \binom{k+l}{k} \mathrm{skw}(\omega \otimes \mu), \quad \omega \in \mathrm{Alt}^k V,\ \mu \in \mathrm{Alt}^l V.$$

Let $v_1, \ldots, v_n$ form a basis for V. Denoting by

$$\Sigma(k,n) = \big\{(\sigma_1, \ldots, \sigma_k) \in \mathbb{N}^k \mid 1 \le \sigma_1 < \cdots < \sigma_k \le n\big\},$$

an element of $\mathrm{Alt}^k V$ is completely determined by the values it assigns to the k-tuples $(v_{\sigma_1}, \ldots, v_{\sigma_k})$, $\sigma \in \Sigma_k$. Moreover, these values can be assigned arbitrarily. In fact, the k-form $\mu_{\sigma_1} \wedge \cdots \wedge \mu_{\sigma_k}$, where $\mu_1, \ldots, \mu_n$ is the dual basis to $v_1, \ldots, v_n$, takes the k-tuple $(v_{\sigma_1}, \ldots, v_{\sigma_k})$ to 1, and the other such k-tuples to 0. Thus $\dim \mathrm{Alt}^k V = \binom{n}{k}$, where $n = \dim V$.

We define differential forms on an arbitrary manifold, since we will be using them both when the manifold is a domain in $\mathbb{R}^n$ and when it is the boundary of such a domain. A differential k-form on a manifold Ω is a map ω which takes each point $x \in \Omega$ to an element $\omega_x \in \mathrm{Alt}^k T_x\Omega$, where $T_x\Omega$ is the tangent space to Ω at x. In other language, ω is a skew-symmetric covariant tensor field on Ω of order k. In particular, a differential 0-form is just a real-valued function on Ω and a differential 1-form is a covector field. In the case Ω is a domain in $\mathbb{R}^n$, then each tangent space can be identified with $\mathbb{R}^n$, and a differential k-form is simply a map $\Omega \to \mathrm{Alt}^k \mathbb{R}^n$. In this context, it is common to denote the dual basis to the canonical basis for $\mathbb{R}^n$ by $dx^1, \ldots, dx^n$, so dx^k applied to a vector $v = (v^1, \ldots, v^n) \in \mathbb{R}^n$ is its kth component v^k. With this notation, an arbitrary differential k-form can be written

$$u(x) = \sum_{\sigma \in \Sigma(k,n)} a_\sigma(x)\, dx^{\sigma_1} \wedge \cdots \wedge dx^{\sigma_k},$$

for some coefficients $a_\sigma : \Omega \to \mathbb{R}$.

Three basic operations on differential forms are the exterior derivative, the form integral, and the pullback. The exterior derivative $d\omega$ of a k-form ω is a $(k+1)$-form. In the case of a domain in $\mathbb{R}^n$, it is given by the intuitive formula

$$d\big(a_\sigma\, dx^{\sigma_1} \wedge \cdots \wedge dx^{\sigma_k}\big) = \sum_{j=1}^{n} \frac{\partial a_\sigma}{\partial x_j}\, dx^j \wedge dx^{\sigma_1} \wedge \cdots \wedge dx^{\sigma_k}.$$

It satisfies (in general) the identity $d^{k+1} \circ d^k = 0$ and the Leibniz rule $d(\omega \wedge \mu) = (d\omega) \wedge \mu + (-1)^k \omega \wedge (d\mu)$ if ω is a k-form.

The definition of the form integral requires that the manifold Ω be oriented. In this case we can define $\int_\Omega \omega \in \mathbb{R}$ for ω an n-form with $n = \dim \Omega$. The integral changes sign if the orientation of the manifold is reversed.

Finally, if $F : \Omega \to \Omega'$ is a differentiable map, then the pullback F^* takes a k-form on Ω' to one on Ω by

$$\big(F^*\omega\big)_x(v_1, \ldots, v_k) = \omega_{F(x)}(dF_x v_1, \ldots, dF_x v_k), \quad x \in \Omega,\ v_1, \ldots, v_k \in T_x\Omega.$$

The pullback respects the operations of wedge product, exterior derivative, and form integral:

$$F^*(\omega \wedge \mu) = \big(F^*\omega\big) \wedge \big(F^*\mu\big), \quad F^*(d\omega) = d\big(F^*\omega\big), \quad \int_\Omega F^*\omega = \int_{\Omega'} \omega,$$

for ω and μ differential forms on Ω'. The last relation requires that F be a diffeomorphism of Ω with Ω' which preserves orientation.

An important special case of pullback is when F is the inclusion of a submanifold Ω into a larger manifold Ω'. In this case the pullback is the trace operator taking a k-form on Ω' to a k-form on the submanifold Ω. All these operations combine elegantly into Stokes' theorem, which says that, under minimal hypothesis on the smoothness of the differential $(n-1)$-form ω and the n-manifold Ω,

$$\int_{\partial\Omega} \operatorname{tr}\omega = \int_\Omega d\omega.$$

If V is an inner product space, then there is a natural inner product on $\operatorname{Alt}^k V$. Thus for a Riemannian manifold, such as any manifold embedded in $\mathbb{R}^n$, the inner product $\langle \omega_x, \mu_x \rangle \in \mathbb{R}$ is defined for any k-forms ω, μ and any $x \in \Omega$. An oriented Riemannian manifold also has a unique volume form, vol, a differential n-form which at each point assigns the value 1 to a positively oriented orthonormal basis for the tangent space at that point. For a subdomain of $\mathbb{R}^n$ the volume form is the constant n-form with the value $dx^1 \wedge \cdots \wedge dx^n$ at each point. Combining these notions, we see that on any oriented Riemannian manifold we may define the L^2-inner product of k-forms:

$$\langle \omega, \mu \rangle_{L^2\Lambda^k(\Omega)} = \int_\Omega \langle \omega_x, \mu_x \rangle \,\mathsf{vol}.$$

The space $L^2\Lambda^k$ is of course the space of k-forms for which $\|\omega\|_{L^2\Lambda^k} := \sqrt{\langle \omega, \omega \rangle_{L^2\Lambda^k}} < \infty$, and then $H\Lambda^k$ is defined as in (1.1).

3 Families of Finite Element Differential Forms on Simplicial Meshes

Our goal now is to create finite element subspaces of the spaces $H\Lambda^k$ which fit together to yield a subcomplex with commuting projections. In this section the spaces will be constructed for a simplicial triangulation $\mathcal{T}_h$ of the domain $\Omega \subset \mathbb{R}^n$. Thus, for a simplex T, we must specify a space $V(T)$ of polynomial differential forms and a set of degrees of freedom for it.

3.1 The Polynomial Space $\mathcal{P}_r\Lambda^k$

An obvious choice for $V(T)$ is the space

$$\mathcal{P}_r\Lambda^k(T) = \left\{ \sum_{\sigma\in\Sigma(k,n)} p_\sigma\, dx^\sigma \;\middle|\; p_\sigma \in \mathcal{P}_r(T) \right\},$$

of a differential k-forms with polynomial coefficients of degree at most r. It is easy to compute its dimension:

$$\dim \mathcal{P}_r\Lambda^k(T) = \#\Sigma(k,n) \times \dim \mathcal{P}_r(T) = \binom{n}{k}\binom{n+r}{n} = \binom{n+r}{n-k}\binom{r+k}{r}. \tag{3.1}$$

Note that $d\mathcal{P}_r\Lambda^k \subset \mathcal{P}_{r-1}\Lambda^{k+1}$, i.e., the exterior derivative lowers the polynomial degree at the same time as it raises the form degree. Therefore, for each r we have a subcomplex of the de Rham complex:

$$\mathcal{P}_r\Lambda^0 \xrightarrow{d} \mathcal{P}_{r-1}\Lambda^1 \xrightarrow{d} \cdots \xrightarrow{d} \mathcal{P}_{r-n}\Lambda^n \to 0. \tag{3.2}$$

This complex is exact (we have left off the initial 0 since the first map, $d = \text{grad}$ acting on $\mathcal{P}_r\Lambda^0$ has a 1-dimensional kernel, consisting of the constant functions). That is, if $\omega \in \mathcal{P}_s\Lambda^k$ and $d\omega = 0$ then $\omega = d\mu$ for some $\mu \in \mathcal{P}_{s+1}\Lambda^{k-1}$. We prove this in Corollary 3.2 below, using an elementary but powerful tool called the *Koszul complex*. The same tool will also be used to define the degrees of freedom for $\mathcal{P}_r\Lambda^k(T)$, and to define an alternative space of shape functions.

3.2 The Koszul Complex

For a domain in $\Omega \subset \mathbb{R}^n$ (but not a general manifold), the identity map may be viewed as a vector field. It assigns to an arbitrary point $x \in \Omega \subset \mathbb{R}^n$ the point itself viewed as a vector in $\mathbb{R}^n$ and so an element of the tangent space $T_x\Omega$. Contracting a k-form ω with this identity vector field gives a $(k-1)$-form $\kappa\omega$:

$$(\kappa\omega)_x(v_1,\dots,v_{k-1}) = \omega_x(x, v_1,\dots,v_{k-1}), \quad x \in \Omega,\ v_1,\dots,v_{k-1} \in \mathbb{R}^n.$$

Since ω_x is skew-symmetric, $\kappa\kappa\omega = 0$, that is, κ is a differential. It satisfies a Leibniz rule:

$$\kappa(\omega\wedge\mu) = (\kappa\omega)\wedge\mu + (-1)^k\omega\wedge(\kappa\mu),$$

for a k-form ω and a second form μ. In particular $\kappa(f\omega) = f\kappa\omega$ if f is a function.

Also $\kappa dx^i = x^i$. These properties fully determine κ. Thus

$$\kappa\big(dx^i \wedge dx^j\big) = x^i\,dx^j - x^j\,dx^i,$$
$$\kappa\big(dx^i \wedge dx^j \wedge dx^k\big) = x^i\,dx^j \wedge dx^k - x^j\,dx^i \wedge dx^k + x^k\,dx^i \wedge dx^j,$$

and so forth. If we identify 1-forms with vector fields, then κ corresponds to the dot product of the vector field with x (or, more properly, with the identity vector field). On 2-forms in 3-D, κ is the cross product with x, and on 3-forms it is the product of a scalar field with x to get a vector field.

The Koszul differential κ maps the space $\mathcal{P}_r\Lambda^k$ of differential k-forms with coefficients in $\mathcal{P}_r(\Omega)$ to $\mathcal{P}_{r+1}\Lambda^{k-1}$, exactly the reverse of d. Thus both κd and $d\kappa$ map $\mathcal{P}_r\Lambda^k$ to itself. The following theorem points to an intimate relation between κ and d, called the *homotopy formula*. In it we write $\mathcal{H}_r\Lambda^k$ for the k-forms with *homogeneous* polynomial coefficients of degree r.

Theorem 3.1 (Homotopy formula)

$$(\kappa d + d\kappa)\omega = (k+r)\omega, \quad \omega \in \mathcal{H}_r\Lambda^k.$$

Remarks on the proof The case $k = 0$ is Euler's identity $x \cdot \operatorname{grad} p = r\,p$ for p a homogeneous polynomial of degree r. Using it, we can verify the theorem by direct computation. Alternatively, one may use Cartan's homotopy formula from differential geometry. For details on both proofs, see Theorem 3.1 of [5]. □

Corollary 3.2 *The polynomial de Rham complex* (3.2) *and the Koszul complex*

$$0 \to \mathcal{P}_{r-n}\Lambda^n \xrightarrow{\kappa} \mathcal{P}_{r-n+1}\Lambda^{n-1} \xrightarrow{\kappa} \cdots \xrightarrow{\kappa} \mathcal{P}_r\Lambda^0$$

are both exact.

Proof For the de Rham complex, it suffices to establish exactness of the homogeneous polynomial de Rham complex

$$\mathcal{H}_r\Lambda^0 \xrightarrow{d} \mathcal{H}_{r-1}\Lambda^1 \xrightarrow{d} \cdots \xrightarrow{d} \mathcal{H}_{r-n}\Lambda^n \to 0,$$

since then we can then just sum to get the result. We must show that if $\omega \in \mathcal{H}_s\Lambda^k$ and $d\omega = 0$ then ω is in the range of d. Indeed, by the homotopy formula

$$\omega = (s+k)^{-1}(d\kappa + \kappa d)\omega = (s+k)^{-1} d\kappa\omega.$$

A similar proof holds for the Koszul complex. □

Another important consequence is a direct sum decomposition:

Corollary 3.3 *For* $r \geq 1$, $0 \leq k \leq n$,

$$\mathcal{H}_r\Lambda^k = \kappa\mathcal{H}_{r-1}\Lambda^{k+1} \oplus d\mathcal{H}_{r+1}\Lambda^{k-1}. \tag{3.3}$$

Proof By the homotopy formula, any element of $\mathcal{H}_r\Lambda^k$ belongs to $\kappa\mathcal{H}_{r-1}\Lambda^{k+1} + d\mathcal{H}_{r+1}\Lambda^{k-1}$. Moreover the intersection of these two spaces is zero, since if ω belongs to the intersection, then $d\omega = 0$, $\kappa\omega = 0$, so $\omega = 0$ by the homotopy formula. □

3.3 The Polynomial Space $\mathcal{P}_r^-\Lambda^k$

We now define a second space of polynomial differential forms which can be used as shape functions. We have

$$\mathcal{P}_r\Lambda^k = \mathcal{P}_{r-1}\Lambda^k \oplus \mathcal{H}_r\Lambda^k = \mathcal{P}_{r-1}\Lambda^k \oplus \kappa\mathcal{H}_{r-1}\Lambda^{k+1} \oplus d\mathcal{H}_{r+1}\Lambda^{k-1}.$$

If we drop the last summand, we get a space intermediate between $\mathcal{P}_{r-1}\Lambda^k$ and $\mathcal{P}_r\Lambda^k$:

$$\mathcal{P}_r^-\Lambda^k := \mathcal{P}_{r-1}\Lambda^k + \kappa\mathcal{H}_{r-1}\Lambda^{k+1}. \tag{3.4}$$

Note that $\mathcal{P}_r^-\Lambda^0 = \mathcal{P}_r\Lambda^0$ and $\mathcal{P}_r^-\Lambda^n = \mathcal{P}_{r-1}\Lambda^n$, but for $0 < k < n$, $\mathcal{P}_r^-\Lambda^k$ is contained strictly between $\mathcal{P}_{r-1}\Lambda^k$ and $\mathcal{P}_r\Lambda^k$. We may compute the dimension of $\kappa\mathcal{H}_r\Lambda^k$, using the exactness of the Koszul complex and induction (see [5, Theorem 3.3]). This then yields a formula for the dimension of $\mathcal{P}_r^-\Lambda^k$:

$$\dim\mathcal{P}_r^-\Lambda^k = \binom{n+r}{n-k}\binom{r+k-1}{k}.$$

Comparing this with (3.1), we have

$$\dim\mathcal{P}_r^-\Lambda^k = \frac{r}{r+k}\dim\mathcal{P}_r\Lambda^k$$

(showing again that the spaces coincide for 0-forms).

Now

$$d\mathcal{P}_r^-\Lambda^k \subset d\mathcal{P}_r\Lambda^k \subset \mathcal{P}_{r-1}\Lambda^{k+1} \subset \mathcal{P}_r^-\Lambda^{k+1},$$

so we obtain another subcomplex of the de Rham complex:

$$\mathcal{P}_r^-\Lambda^0 \xrightarrow{d} \mathcal{P}_r^-\Lambda^1 \xrightarrow{d} \cdots \xrightarrow{d} \mathcal{P}_r^-\Lambda^n \to 0. \tag{3.5}$$

Note that, in contrast to (3.2), in this complex the degree r is held constant. However, like (3.2), the complex (3.5) is exact. Indeed,

$$\begin{aligned} d\mathcal{P}_r^-\Lambda^k &= d\big(\mathcal{P}_r^-\Lambda^k + d\mathcal{P}_{r+1}\Lambda^{k-1}\big) = d\mathcal{P}_r\Lambda^k \\ &= \mathcal{N}\big(d|\mathcal{P}_{r-1}\Lambda^{k+1}\big) = \mathcal{N}\big(d|\mathcal{P}_r^-\Lambda^{k+1}\big), \end{aligned}$$

where the penultimate equality follows from Corollary 3.2 and the last equality is a consequence of the definition (3.4) and the homotopy formula Theorem 3.1.

3.4 The $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ Family of Finite Element Differential Forms

Let $r \geq 1$, $0 \leq k \leq n$, and let $\mathcal{T}_h$ be a simplicial mesh of $\Omega \subset \mathbb{R}^n$. We define a finite element subspace $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ of $H\Lambda^k(\Omega)$. As shape functions on a simplex $T \in \mathcal{T}_h$ we take $V(T) = \mathcal{P}_r^- \Lambda^k(T)$. As degrees of freedom we take

$$u \in \mathcal{P}_r^- \Lambda^k(T) \mapsto \int_f (\operatorname{tr}_f u) \wedge q, \quad q \in \mathcal{P}_{r+k-d-1} \Lambda^{d-k}(f), \ f \in \Delta_d(T), \ d \geq k. \tag{3.6}$$

Note that, in the case $k = 0$, $V(T) = \mathcal{P}_r(T)$ and (3.6) coincides with (2.1), so the space $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ generalizes the Lagrange finite elements to differential forms of arbitrary form degree. We shall prove unisolvence for arbitrary polynomial degree, form degree, and space dimension at once. The proof will use the following lemma, which is proved via a simple construction using barycentric coordinates.

Lemma 3.4 *Let $r \geq 1$, $0 \leq k \leq n$, and let T be an n-simplex. If $u \in \mathring{\mathcal{P}}_{r-1} \Lambda^k(T)$ and*

$$\int_T u \wedge q = 0, \quad q \in \mathcal{P}_{r+k-n-1} \Lambda^{n-k}(T), \tag{3.7}$$

then $u \equiv 0$.

Proof Any element of $\mathcal{P}_{r-1} \Lambda^k(T)$ can be written in terms of barycentric coordinates as

$$u = \sum_{\sigma \in \Sigma(k,n)} u_\sigma \, d\lambda_{\sigma_1} \wedge \cdots \wedge d\lambda_{\sigma_k}, \quad u_\sigma \in \mathcal{P}_{r-1}(T).$$

Now let $1 \leq i \leq n$, and consider the trace of u on the face given by $\lambda_i = 0$. By the assumption that $u \in \mathring{\mathcal{P}}_{r-1} \Lambda^k(T)$, the trace vanishes. This implies that λ_i divides u_σ for any $\sigma \in \Sigma(k,n)$ whose range does not contain i. Thus

$$u_\sigma = p_\sigma \lambda_{\sigma_1^*} \cdots \lambda_{\sigma_{n-k}^*} \quad \text{for some } p_\sigma \in \mathcal{P}_{r+k-n-1}(T),$$

where $\sigma^* \in \Sigma(n-k,n)$ is the increasing sequence complementary to σ. Thus

$$u = \sum_{\sigma \in \Sigma(k,n)} p_\sigma \lambda_{\sigma_1^*} \cdots \lambda_{\sigma_{n-k}^*} d\lambda_{\sigma_1} \wedge \cdots \wedge d\lambda_{\sigma_k}, \quad p_\sigma \in \mathcal{P}_{r+k-n-1}(T).$$

Choosing

$$q = \sum_{\sigma \in \Sigma(k,n)} (-1)^{\operatorname{sign}(\sigma,\sigma^*)} p_\sigma \, d\lambda_{\sigma_1^*} \wedge \cdots \wedge d\lambda_{\sigma_{n-k}^*}$$

in (3.7), we get

$$0 = \int_T u \wedge q = \int_T \sum_{\sigma \in \Sigma(k,n)} p_\sigma^2 \lambda_{\sigma_1^*} \cdots \lambda_{\sigma_{n-k}^*} d\lambda_1 \wedge \cdots \wedge d\lambda_n.$$

However, the λ_i are positive on the interior of T and the n-form $d\lambda_1 \wedge \cdots \wedge d\lambda_n$ is a nonzero multiple of the volume form. Thus each p_σ must vanish, and so u vanishes. □

Theorem 3.5 (Unisolvence for $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$) *For any $r \geq 1$, $0 \leq k \leq n$, and n-simplex T, the degrees of freedom (3.6) are unisolvent for $V(T) = \mathcal{P}_r^- \Lambda^k(T)$.*

Proof First we do the dimension count. The number of degrees of freedom is at most

$$\sum_{d \geq k} \#\Delta_d(T) \dim \mathcal{P}_{r+k-d-1} \Lambda^k(\mathbb{R}^d) = \sum_{d \geq k} \binom{n+1}{d+1} \binom{r+k-1}{d} \binom{d}{k}$$
$$= \sum_{j \geq 0} \binom{n+1}{j+k+1} \binom{r+k-1}{j+k} \binom{j+k}{j}.$$

Simplifying with the binomial identities,

$$\binom{a}{b}\binom{b}{c} = \binom{a}{c}\binom{a-c}{a-b}, \quad \sum_{j \geq 0} \binom{a}{b+j}\binom{c}{j} = \binom{a+c}{a-b},$$

the right-hand side becomes

$$\binom{r+n}{r+k}\binom{r+k-1}{k} = \dim \mathcal{P}_r^- \Lambda^k(T).$$

It remains to show that if $u \in \mathcal{P}_r^- \Lambda^k(T)$ and the degrees of freedom in (3.6) vanish, then u vanishes. Since $\operatorname{tr}_f \mathcal{P}_r^- \Lambda^k(T) = \mathcal{P}_r^- \Lambda^k(f)$, we may use induction on dimension to conclude that $\operatorname{tr}_f u$ vanishes on each facet f, so $u \in \mathring{\mathcal{P}}_r^- \Lambda^k(T)$. Therefore $du \in \mathring{\mathcal{P}}_{r-1} \Lambda^{k+1}(T)$. Moreover,

$$\int_T du \wedge p = \pm \int_T u \wedge dp = 0, \quad p \in \mathcal{P}_{r+k-n} \Lambda^{n-k-1}(T),$$

where the first equality comes from Stoke's theorem and the Leibniz rule, and the second from the hypothesis that the degrees of freedom for u vanish. We may now apply the lemma (with k replaced by $k+1$) to du to conclude that du vanishes. But the homotopy formula implies that for $u \in \mathcal{P}_r^- \Lambda^k$ with $du = 0$, $u \in \mathcal{P}_{r-1} \Lambda^k$. Using the interior degrees of freedom from (3.6), we may apply the lemma to u, to conclude that u vanishes. □

It is easy to check that the degrees of freedom imply single-valuedness of the traces of elements of $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$, so that they indeed belong to $H\Lambda^k$. Moreover, it is easy to see that the complex (3.5) involving the shape functions, leads to a finite element subcomplex of the L^2 de Rham complex on Ω:

$$\mathcal{P}_r^- \Lambda^0(\mathcal{T}_h) \xrightarrow{d} \mathcal{P}_r^- \Lambda^1 \xrightarrow{d} (\mathcal{T}_h) \xrightarrow{d} \cdots \xrightarrow{d} \mathcal{P}_r^- \Lambda^n(\mathcal{T}_h).$$

Using the degrees of freedom to define projection operators π_h^k into $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ (the domain of π_h^k consists of all continuous k-forms in $H\Lambda^k(\Omega)$), we obtain projections that commute with d (this can be verified using Stokes' theorem), which is crucial to the analysis of the element via FEEC.

3.5 The $\mathcal{P}_r\Lambda^k(\mathcal{T}_h)$ Family of Finite Element Differential Forms

We may also use the full polynomial space $\mathcal{P}_r\Lambda^k(T)$ as shape functions for a finite element space. The corresponding degrees of freedom are

$$u \in \mathcal{P}_r\Lambda^k(T) \mapsto \int_f (\operatorname{tr}_f u) \wedge q, \quad q \in \mathcal{P}^-_{r+k-d}\Lambda^{d-k}(f), \ f \in \Delta_d(T), \ d \geq k. \tag{3.8}$$

Note that in this case the degrees of freedom involve $\mathcal{P}_r^-$ spaces, defined through the Koszul complex. The analysis of these spaces is very parallel to that of the last subsection, and we will not carry it out here. Again, we obtain unisolvence, and a finite element subcomplex of the de Rham complex

$$\mathcal{P}_r\Lambda^0(\mathcal{T}_h) \xrightarrow{d} \mathcal{P}_{r-1}\Lambda^1(\mathcal{T}_h) \xrightarrow{d} \cdots \xrightarrow{d} \mathcal{P}_{r-n}\Lambda^n(\mathcal{T}_h),$$

which admits a commuting projection defined via the degrees of freedom.

3.6 Historical Notes

In the case $k = 0$, the two shape function spaces $\mathcal{P}_r^-\Lambda^k$ and $\mathcal{P}_r\Lambda^k$ coincide, as do the spaces $\mathcal{P}_{r-d-1}\Lambda^{d-k}(f)$ and $\mathcal{P}^-_{r-d}\Lambda^{d-k}(f)$, $f \in \Delta_d(T)$, entering (3.6) and (3.8). Thus the two finite element families coincide for 0-forms, and provide two distinct generalizations of the Lagrange elements to differential forms of higher degree.

In n dimensions, n-forms may be viewed as scalar functions and the space $H\Lambda^n(\Omega)$ just corresponds to $L^2(\Omega)$. The finite element subspace $\mathcal{P}_r\Lambda^n(\mathcal{T}_h)$ is simply the space of all piecewise polynomial functions of degree r, with no interelement continuity required. The space $\mathcal{P}_r^-\Lambda^n(\mathcal{T}_h)$ coincides with $\mathcal{P}_{r-1}\Lambda^n(\mathcal{T}_h)$.

In two dimensions, the remaining spaces $\mathcal{P}_r^-\Lambda^1(\mathcal{T}_h)$ and $\mathcal{P}_r\Lambda^1(\mathcal{T}_h)$ can be identified, via vector proxies, with the Raviart–Thomas spaces [20] and the Brezzi–Douglas–Marini spaces [10]. In three dimensions, the $\mathcal{P}_r^-\Lambda^1(\mathcal{T}_h)$ and $\mathcal{P}_r^-\Lambda^2(\mathcal{T}_h)$ spaces are the finite element subspaces of $H(\operatorname{curl}, \Omega)$ and $H(\operatorname{div}, \Omega)$, respectively, called the Nédélec edge and face elements of the first kind [18]. The spaces $\mathcal{P}_r\Lambda^1(\mathcal{T}_h)$ and $\mathcal{P}_r\Lambda^2(\mathcal{T}_h)$ are the Nédélec edge and face elements of the second kind [19]. Diagrams for the two-dimensional and three-dimensional elements are shown in Figs. 2 and 3.

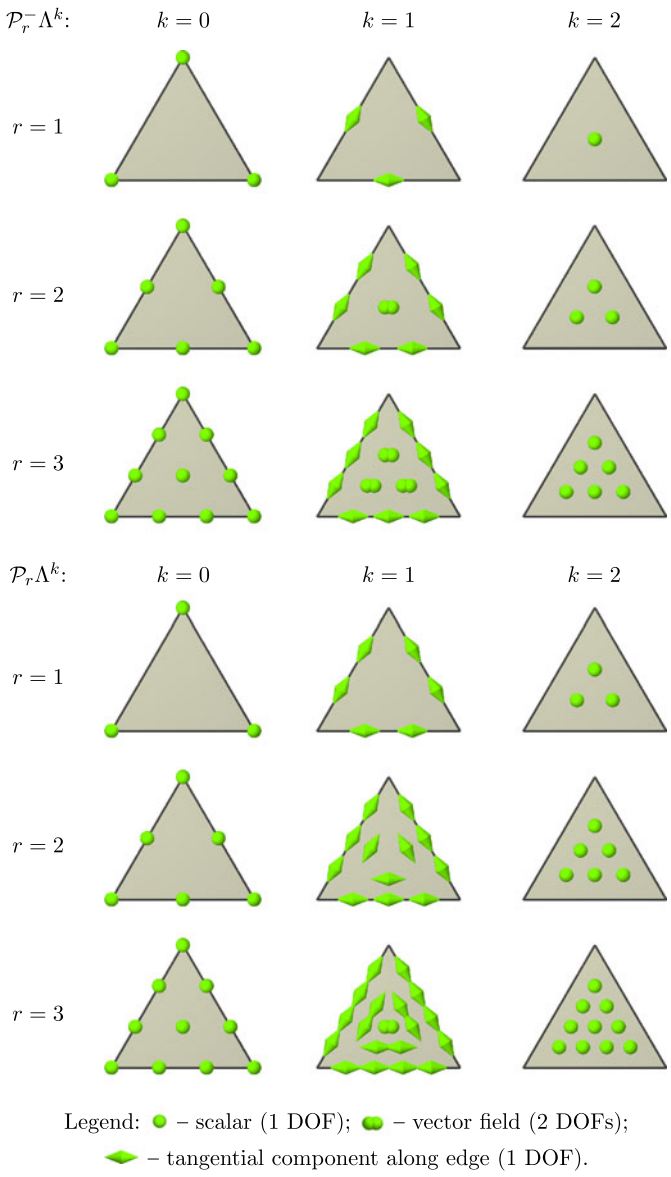


Fig. 2 The $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ and $\mathcal{P}_r \Lambda^k(\mathcal{T}_h)$ spaces in two dimensions

The lowest order spaces $\mathcal{P}_1^- \Lambda^k(\mathcal{T}_h)$ are very geometric, possessing precisely one degree of freedom per face of dimension k, and no others (see the top rows of Figs. 2 and 3). In fact these spaces first appeared in the geometry literature in the work of Whitney in 1957 [24] long before their first appearance as finite elements. In the 1970s, they were used by Dodziuk [13] and Dodziuk and Patodi [14] as a theoret-

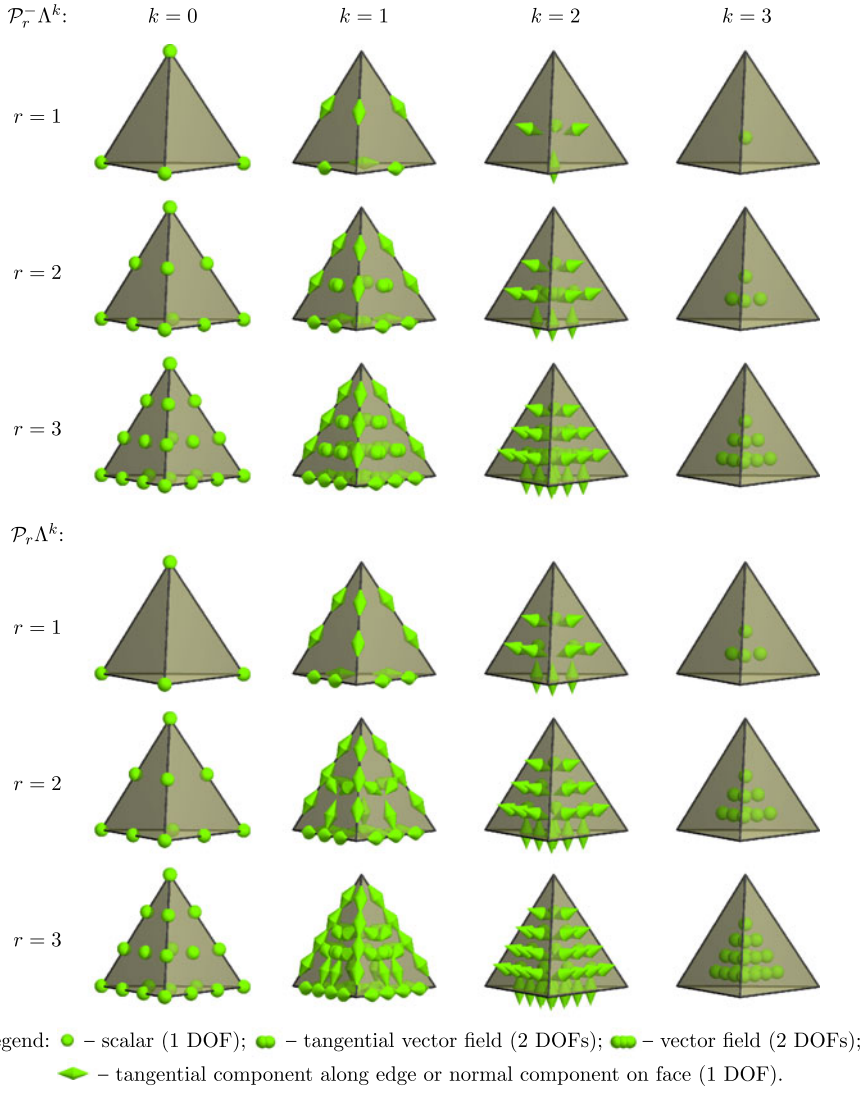


Fig. 3 The $\mathcal{P}_r^- \Lambda^k(\mathcal{T}_h)$ and $\mathcal{P}_r \Lambda^k(\mathcal{T}_h)$ spaces in three dimensions

ical tool to approximate the eigenvalues of the Hodge Laplacian on a Riemannian manifold. This then played an essential role in Müller's proof of the Ray–Singer conjecture [17]. The spaces $\mathcal{P}_r \Lambda^k(\mathcal{T}_h)$ also appeared in the geometry literature, introduced by Sullivan [22, 23]. In an early, largely overlooked paper bringing finite element analysis techniques to bear on geometry Baker [7] named these Sullivan–Whitney forms, and analyzed their convergence for the eigenvalue problem for the Hodge Laplacian. In 1988 Bossavit made the connection between Whitney forms and the mixed finite elements in use in electromagnetics [8], in part inspired by the

thesis of Kotiuga [16]. The first unified treatment of the $\mathcal{P}_r^- \Lambda^k$ spaces, which was based on exterior calculus and included a unisolvence proof, was in a seminal paper of Hiptmair [15] in 1999. In the 2006 paper of Arnold, Falk, and Winther [5], in which the term finite element exterior calculus first appeared, the Koszul complex was first applied to finite elements, simplifying many aspects and resulting in a simultaneous treatment of both the $\mathcal{P}_r^- \Lambda^k$ and $\mathcal{P}_r \Lambda^k$ spaces.

4 Families of Finite Element Differential Forms on Cubical Meshes

We now describe two families of spaces of finite element differential forms, which we denote $\mathcal{Q}_r^- \Lambda^k(\mathcal{T}_h)$ and $\mathcal{S}_r \Lambda^k(\mathcal{T}_h)$, defined for cubical meshes $\mathcal{T}_h$, i.e., meshes in which each element is the Cartesian product of intervals. In some sense, the $\mathcal{Q}_r^- \Lambda^k$ family can be seen as an analogue of the $\mathcal{P}_r^- \Lambda^k$ family for simplicial meshes, and the $\mathcal{S}_r \Lambda^k$ family an analogue of the $\mathcal{P}_r \Lambda^k$ family. The $\mathcal{Q}_r^- \Lambda^k$ family can be constructed from the one-dimensional case by a tensor product construction, and is long known. By contrast, the $\mathcal{S}_r \Lambda^k$ family first appeared in recent work of Arnold and Awanou [3]. Even in two and three dimensions, the spaces in this family were for the most part not known previously.

4.1 The $\mathcal{Q}_r^- \Lambda^k$ Family

We describe this family only very briefly. A more detailed description will be included in a forthcoming study of the approximation properties of these spaces under non-affine mappings [4]. Suppose we are given a subcomplex of the de Rham complex on an element $S \subset \mathbb{R}^m$ and a second such subcomplex on an element $T \subset \mathbb{R}^n$:

$$V^0(S) \xrightarrow{d} V^1(S) \xrightarrow{d} \cdots \xrightarrow{d} V^m(S), \qquad V^0(T) \xrightarrow{d} V^1(T) \xrightarrow{d} \cdots \xrightarrow{d} V^n(T).$$

We may then construct a subcomplex of the de Rham complex on $S \times T$ by a tensor product construction which is known in the theory of differential forms; see, e.g., [21, p. 61]. The canonical projection $\pi_S : S \times T \to S$ determines a pullback of i-forms on S to i-forms on $S \times T$, so $\pi_S^* V^i(S)$ is a space of i-forms on $S \times T$ and, similarly, $\pi_T^* V^j(T)$ is a space of j-forms on $S \times T$. Thus we may define a space of k-forms on $S \times T$ by

$$V^k(S \times T) = \bigoplus_{i+j=k} \pi_S^* V^i(S) \wedge \pi_T^* V^j(T).$$

We take the space $V^k(S \times T)$ as the shape functions for k-forms on $S \times T$. The construction of degrees of freedom for $V^k(S \times T)$ is simple. If $\eta \in V^i(S)^*$ is a

degree of freedom associated to a face f of S, and $\rho \in V^j(T)^*$ is associated to a face g of T, we define

$$\eta \wedge \rho \in \left[\pi_S^* V^i(S) \wedge \pi_T^* V^j(T)\right]^* \subset V^k(S \times T)^*,$$

by

$$(\eta \wedge \rho)\left(\pi_s^* u \wedge \pi_T^* v\right) = \eta(u)\rho(v),$$

and associate the degree of freedom $\eta \wedge \rho$ to $f \times g$, which is a face of $S \times T$.

The $\mathcal{Q}_r^-$ family is defined by applying this tensor product repeatedly, starting with a finite element de Rham complex on an interval in one dimension. In one dimension the $\mathcal{P}_r^-$ and $\mathcal{P}_r$ de Rham subcomplexes coincide. On an interval I, the shape functions for 0-forms are $V^0(I) = \mathcal{P}_r(I)$ with degrees of freedom at each end point, and moments of degree at most $r-1$ in the interior. The shape function for 1-forms are $V^1(I) = \mathcal{P}_{r-1}(I)$ with all degrees of freedom in the interior. Repeatedly using the tensor product construction just outlined, we obtain polynomial spaces and degrees of freedom on a box $I_1 \times \cdots \times I_n \subset \mathbb{R}^n$. We denote the shape function space so obtained by $\mathcal{Q}_r^- \Lambda^k(I_1 \times \cdots \times I_n)$. In $n = 2$ dimensions, for example,

$$\begin{aligned}
\mathcal{Q}_r^- \Lambda^0(I_1 \times I_2) &= \mathcal{Q}_r(I_1 \times I_2) = \mathcal{P}_r(I_1) \otimes \mathcal{P}_r(I_2), \\
\mathcal{Q}_r^- \Lambda^1(I_1 \times I_2) &= \left[\mathcal{P}_{r-1}(I_1) \otimes \mathcal{P}_r(I_2)\right] \times \left[\mathcal{P}_r(I_1) \otimes \mathcal{P}_{r-1}(I_2)\right], \\
\mathcal{Q}_r^- \Lambda^2(I_1 \times I_2) &= \mathcal{Q}_{r-1}(I_1 \times I_2).
\end{aligned}$$

Diagrams for these elements in two and three dimensions are shown in Fig. 4. The space $\mathcal{Q}_r^- \Lambda^0(\mathcal{T}_h)$ is the standard $\mathcal{Q}_r$ finite element subspace of $H^1(\Omega)$ and the space $\mathcal{Q}_r^- \Lambda^n(\mathcal{T}_h)$ is the discontinuous $\mathcal{Q}_{r-1}$ subspace of $L^2(\Omega)$. The space $\mathcal{Q}_r^- \Lambda^1(\mathcal{T}_h)$ goes back to Raviart and Thomas [20] in two dimensions, and the $\mathcal{Q}_r^- \Lambda^1(\mathcal{T}_h)$ and $\mathcal{Q}_r^- \Lambda^2(\mathcal{T}_h)$ were given by Nédélec in [18]. The spaces with r held fixed combine to create a finite element de Rham subcomplex,

$$\mathcal{Q}_r^- \Lambda^0(\mathcal{T}_h) \xrightarrow{d} \mathcal{Q}_r^- \Lambda^1(\mathcal{T}_h) \xrightarrow{d} \cdots \xrightarrow{d} \mathcal{Q}_r^- \Lambda^n(\mathcal{T}_h),$$

and the degrees of freedom determine commuting projections.

Recently, Cockburn and Qiu [12] have published a different family of finite element spaces in two and three dimensions, that seems to be related to these. They begin with the complex formed by the full spaces $\mathcal{Q}_r \Lambda^k$, which lie between $\mathcal{Q}_r^- \Lambda^k$ and $\mathcal{Q}_{r+1}^- \Lambda^k$. That complex (which was discussed in [19]) does *not* admit commuting projections. Cockburn and Qiu define a small space of bubble functions that can be added to each of the spaces so that the resulting spaces remain inside $\mathcal{Q}_{r+1}^- \Lambda^k$ but also form a de Rham subcomplex (with constant r) which admits commuting projections.

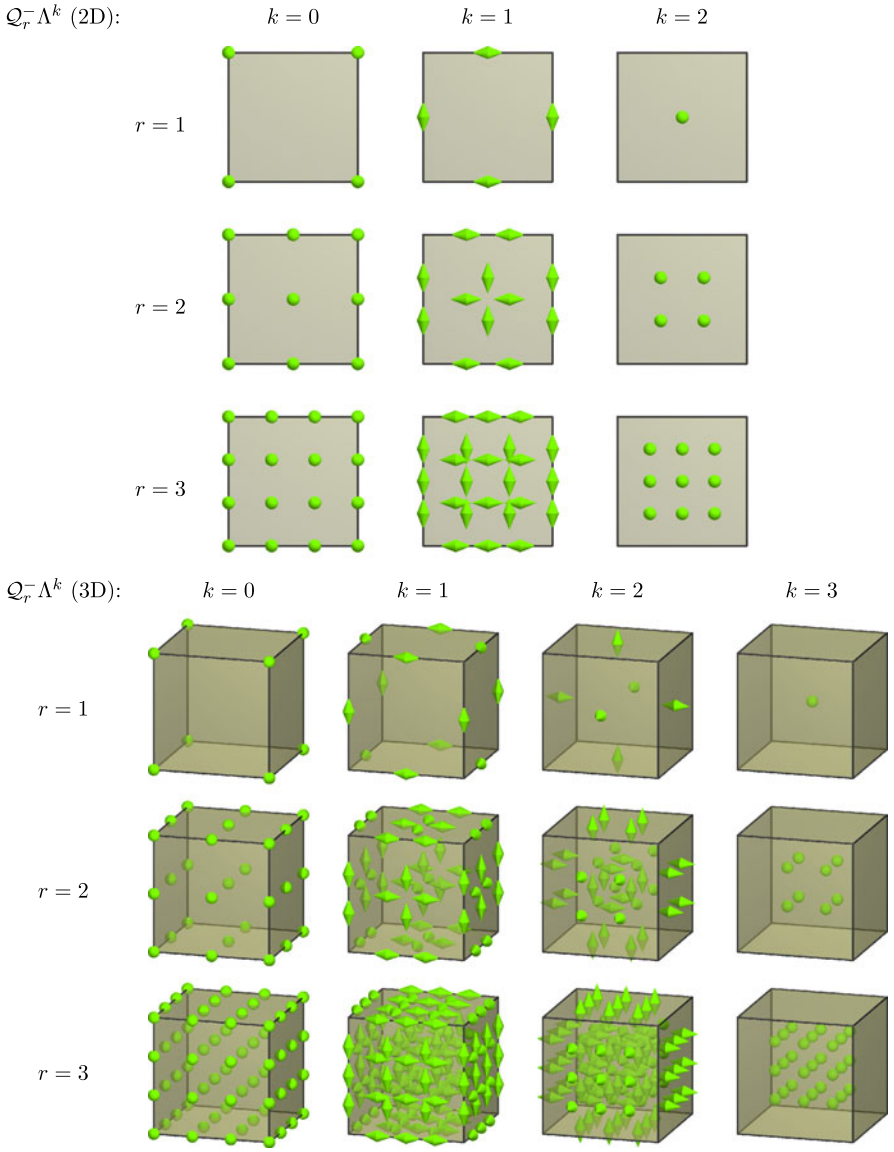


Fig. 4 The $\mathcal{Q}_r^- \Lambda^k(\mathcal{T}_h)$ spaces in two and three dimensions

4.2 A Second Family of Finite Element Differential Forms on Cubes

The $\mathcal{S}_r \Lambda^k$ family presented in this section was derived recently in [3]. It seems to be complementary to the $\mathcal{Q}_r^- \Lambda^k$ family much as the $\mathcal{P}_r \Lambda^k$ family complements

the $\mathcal{P}_r^-\Lambda^k$ family. To describe the new family we require some notation. A k-form monomial in n variables is the product of an ordinary monomial and a simple alternator:

$$m = \big(x^1\big)^{\alpha_1}\cdots\big(x^n\big)^{\alpha_n}\,dx^{\sigma_1}\wedge\cdots\wedge dx^{\sigma_k},$$

where α is a multi-index and $\sigma\in\Sigma(k,n)$. We define the degree of m to be the polynomial degree of its coefficient: $\deg m = \sum_i \alpha_i$. The *linear degree* of m is more complicated:

$$\operatorname{ldeg} m = \#\big\{i \mid \alpha_i = 1,\ \alpha_i\notin\{\sigma_1,\dots,\sigma_k\}\big\},$$

that is, the number of variables that enter the coefficient linearly, not counting the variables that enter the alternator. For example, if $m = x^1x^2(x^3)^5\,dx^1$, then $\deg m = 7$, $\operatorname{ldeg} m = 1$.

We now define the space of shape functions we shall use for k-forms on an n-dimensional box, T. Viewing monomial forms as differential forms on T, we define $\mathcal{H}_{r,l}\Lambda^k(T)\subset\mathcal{H}_r\Lambda^k(T)$ to be the span of all monomial k-forms m such that $\deg m = r$ and $\operatorname{ldeg} m\ge l$. Using this definition and the Koszul differential, we then define

$$\mathcal{J}_r\Lambda^k(T) = \sum_{l\ge1}\kappa\mathcal{H}_{r+l-1,l}\Lambda^{k+1}(T)\subset\mathcal{P}_{r+n-k-1}\Lambda^k(T).$$

Finally, we define the shape functions on T by

$$\mathcal{S}_r\Lambda^k(T) = \mathcal{P}_r\Lambda^k(T) + \mathcal{J}_r\Lambda^k(T) + d\mathcal{J}_{r+1}\Lambda^{k-1}(T),$$

defined for all $r\ge1$, $0\le k\le n$.

As the definition of the shape functions takes a while to absorb, we describe the spaces in more elementary terms in the case of three dimensions.

- The space $\mathcal{S}_r\Lambda^0$, the polynomial shape functions for the H^1 space, consists of all polynomials u with *superlinear degree* $\operatorname{sdeg} u\le r$. The superlinear degree of a monomial is its degree ignoring any variable that enters to the first power, and the superlinear degree of a polynomial is the maximum over its monomials. The criterion $\operatorname{sdeg} u\le r$ was introduced in [2] to generalize the serendipity elements from 2 to n-dimensions.
- The space $\mathcal{S}_r\Lambda^1$, the shape functions for the $H(\operatorname{curl})$ space, consists of vector fields of the form

  $$\big(v^1,v^2,v^3\big) + \big(x^2x^3\big(w^2-w^3\big), x^3x^1\big(w^3-w^1\big), x^1x^2\big(w^1-w^2\big)\big) + \operatorname{grad} u,$$

  with polynomials v^i, w^i, and u for which $\deg v^i\le r$, $\deg w^i\le r-1$, $\operatorname{sdeg} u\le r+1$, and w^i is independent of the variable x^i.
- The $H(\operatorname{div})$ space uses shape functions $\mathcal{S}_r\Lambda^2$, which are of the form

  $$\big(v^1,v^2,v^3\big) + \operatorname{curl}\big(x^2x^3\big(w^2-w^3\big), x^3x^1\big(w^3-w^1\big), x^1x^2\big(w^1-w^2\big)\big),$$

  with $\deg v^i\le r$, $\deg w^i\le r$, and w^i independent of the variable x^i.

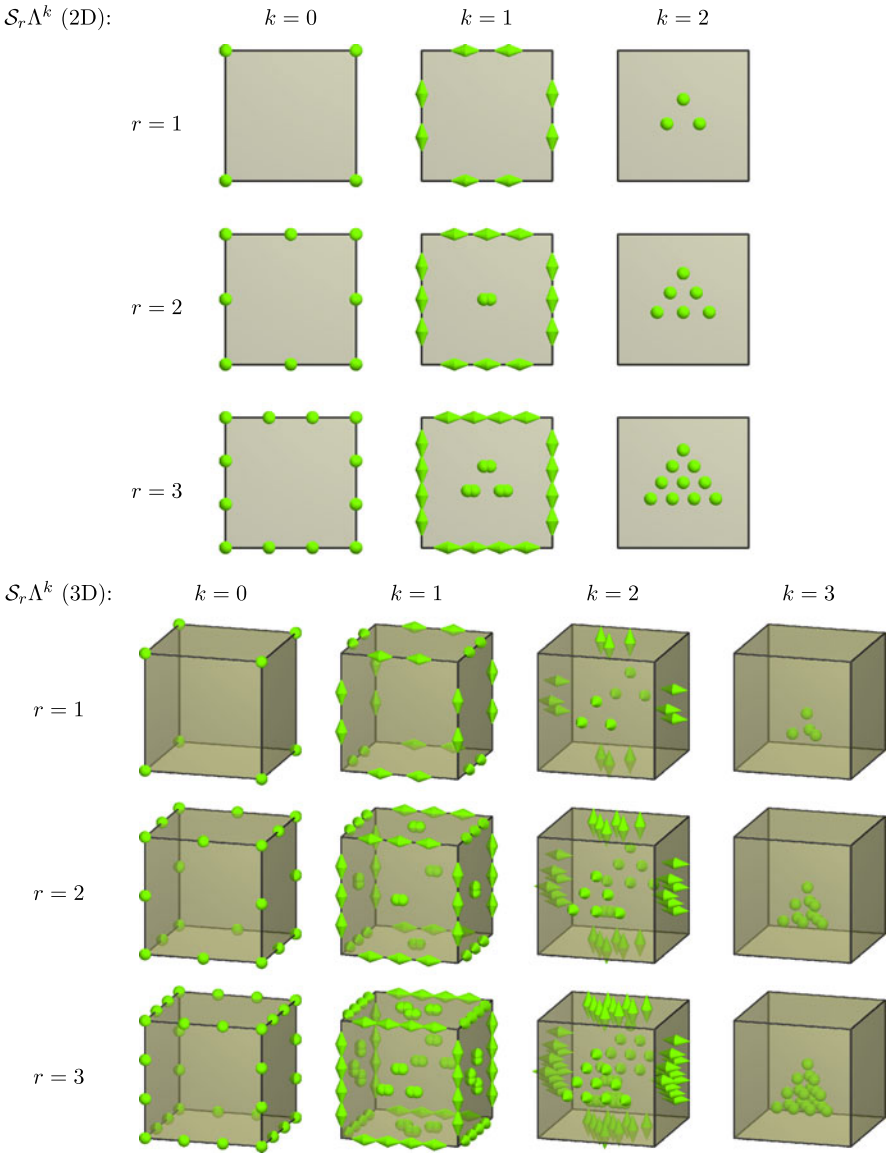


Fig. 5 The $\mathcal{S}_r\Lambda^k(\mathcal{T}_h)$ spaces in two and three dimensions

- Finally the L^2 space $\mathcal{S}_r\Lambda^3$ simply coincides with $\mathcal{P}_r$.

In [3] we establish the following properties of these spaces (in any dimension):

- degree property: $\mathcal{P}_r\Lambda^k(I^n) \subset \mathcal{S}_r\Lambda^k(I^n) \subset \mathcal{P}_{r+n-k}\Lambda^k(I^n)$;
- inclusion property: $\mathcal{S}_r\Lambda^k(I^n) \subset \mathcal{S}_{r+1}\Lambda^k(I^n)$;

Table 1 Dimension of $\mathcal{Q}_r^- \Lambda^k(I^n)$ and $\mathcal{S}_r \Lambda^k(I^n)$

k	r						r					
	1	2	3	4	5	6	1	2	3	4	5	6
$n=1$												
0	2	3	4	5	6	7	2	3	4	5	6	7
1	1	2	3	4	5	6	2	3	4	5	6	7
$n=2$												
0	4	9	16	25	36	49	4	8	12	17	23	30
1	4	12	24	40	60	84	8	14	22	32	44	58
2	1	4	9	16	25	36	3	6	10	15	21	28
$n=3$												
0	8	27	64	125	216	343	8	20	32	50	74	105
1	12	54	144	300	540	882	24	48	84	135	204	294
2	6	36	108	240	450	756	18	39	72	120	186	273
3	1	8	27	64	125	216	4	10	20	35	56	84
$n=4$												
0	16	81	256	625	1296	2401	16	48	80	136	216	328
1	32	216	768	2000	4320	8232	64	144	272	472	768	1188
2	24	216	864	2400	5400	10584	72	168	336	606	1014	1602
3	8	96	432	1280	3000	6048	32	84	180	340	588	952
4	1	16	81	256	625	1296	5	15	35	70	126	210

- trace property: for each face f of I^n, $\operatorname{tr}_f \mathcal{S}_r \Lambda^k(I^n) \subset \mathcal{S}_r \Lambda^k(f)$;
- subcomplex property: $d\mathcal{S}_r \Lambda^k(I^n) \subset \mathcal{S}_{r-1} \Lambda^{k+1}(I^n)$.

The degrees of freedom for $\mathcal{S}_r \Lambda^k(T)$ are quite simple:

$$u \in \mathcal{S}_r \Lambda^k(T) \mapsto \int_f (\operatorname{tr}_f u) \wedge q, \quad q \in \mathcal{P}_{r-2(d-k)} \Lambda^{d-k}(f), \ f \in \Delta_d(T), \ d \geq k. \tag{4.1}$$

These are illustrated in Fig. 5. Notice that weighting function q is sought in a $\mathcal{P}_s$ space, not a $\mathcal{Q}_s$ space. Moreover, as the face dimension d increases by 1, the degree s of the space used for q decreases by 2. A major result of [3] is a proof that the degrees of freedom are unisolvent. Further, we show there that the resulting finite element spaces combine into de Rham subcomplexes with commuting projections:

$$\mathcal{S}_r \Lambda^0(\mathcal{T}_h) \xrightarrow{d} S_{r-1} \Lambda^1(\mathcal{T}_h) \xrightarrow{d} \cdots \xrightarrow{d} S_{r-n} \Lambda^n(\mathcal{T}_h),$$

in which the degrees r decrease, as for the $\mathcal{P}_r \Lambda^K(\mathcal{T}_h)$ spaces on simplices.

For n-forms, the space $\mathcal{S}_r \Lambda^n(\mathcal{T}_h)$ is simply the discontinuous $\mathcal{P}_r$ space (but defined on boxes, rather than simplices). In 2-dimensions, the 0-form space $\mathcal{S}_r \Lambda^0(\mathcal{T}_h)$ is the well-known serendipity space, and the 1-form space is the rectangular BDM

space defined in [10]. Hence these spaces were all known in 2 dimensions. However, in 3 and more dimensions they were not. The 0-form space is the appropriate generalization of the serendipity space to higher dimensions, a space first defined in 2011 [2]. The space $\mathcal{S}_r\Lambda^2$ in 3-D is, we believe, the correct analogue of the BDM elements to cubical meshes. It has the same degrees of freedom as the space in [9] but the shape functions have better symmetry properties. For 1-forms in 3-D, $\mathcal{S}_r\Lambda^1$ is a finite element discretization of $H(\mathrm{curl})$. To the best of our knowledge, neither the degrees of freedom nor the shape functions for this space had been proposed previously. Finally, we note that the dimension of $\mathcal{S}_r^-\Lambda^k(T)$ tends to be much smaller than that of $\mathcal{Q}_r^-\Lambda^k(T)$, especially for r large, as can be observed in Table 1.

References

1. Arnold, D.N.: Differential complexes and numerical stability. In: Proceedings of the International Congress of Mathematicians, vol. I, Beijing, 2002, pp. 137–157. Higher Education Press, Beijing (2002). MR MR1989182 (2004h:65115)
2. Arnold, D.N., Awanou, G.: The serendipity family of finite elements. Found. Comput. Math. **11**(3), 337–344 (2011). doi:10.1007/s10208-011-9087-3
3. Arnold, D.N., Awanou, G.: Finite element differential forms on cubical meshes. Preprint (2012). URL: http://arxiv.org/pdf/1204.2595
4. Arnold, D.N., Boffi, D., Bonizzoni, F.: Approximation by tensor product finite element differential forms (2012, in preparation)
5. Arnold, D.N., Falk, R.S., Winther, R.: Finite element exterior calculus, homological techniques, and applications. Acta Numer. **15**, 1–155 (2006). MR MR2269741 (2007j:58002)
6. Arnold, D.N., Falk, R.S., Winther, R.: Finite element exterior calculus: from Hodge theory to numerical stability. Bull. Am. Math. Soc. **42**(2), 281–354 (2010)
7. Baker, G.A.: Combinatorial Laplacians and Sullivan-Whitney forms. In: Differential Geometry, College Park, MD, 1981/1982. Progr. Math., vol. 32, pp. 1–33. Birkhäuser, Boston (1983). MR MR702525 (84m:58005)
8. Bossavit, A.: Whitney forms: a class of finite elements for three-dimensional computations in electromagnetism. IEEE Trans. Magn. **135**(Part A), 493–500 (1988)
9. Brezzi, F., Douglas, J. Jr., Durán, R., Fortin, M.: Mixed finite elements for second order elliptic problems in three variables. Numer. Math. **51**, 237–250 (1987). MR MR890035 (88f:65190)
10. Brezzi, F., Douglas, J. Jr., Marini, L.D.: Two families of mixed finite elements for second order elliptic problems. Numer. Math. **47**, 217–235 (1985). MR MR799685 (87g:65133)
11. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978). MR MR0520174 (58 #25001)
12. Cockburn, B., Qiu, W.: Commuting diagrams for the TNT elements on cubes. Math. Comput. (2012, to appear)
13. Dodziuk, J.: Finite-difference approach to the Hodge theory of harmonic forms. Am. J. Math. **98**(1), 79–104 (1976). MR MR0407872 (53 #11642)
14. Dodziuk, J., Patodi, V.K.: Riemannian structures and triangulations of manifolds. J. Indian Math. Soc. (N.S.) **40**(1–4), 1–52 (1976). MR MR0488179 (58 #7742)
15. Hiptmair, R.: Canonical construction of finite elements. Math. Comput. **68**, 1325–1346 (1999). MR MR1665954 (2000b:65214)
16. Kotiuga, P.R.: Hodge decompositions and computational electromagnetics. PhD in Electrical Engineering, McGill University (1984)
17. Müller, W.: Analytic torsion and R-torsion of Riemannian manifolds. Adv. Math. **28**(3), 233–305 (1978). MR MR498252 (80j:58065b)

18. Nédélec, J.-C.: Mixed finite elements in $\mathbf{R}^3$. Numer. Math. **35**, 315–341 (1980). MR MR592160 (81k:65125)
19. Nédélec, J.-C.: A new family of mixed finite elements in $\mathbf{R}^3$. Numer. Math. **50**, 57–81 (1986). MR MR864305 (88e:65145)
20. Raviart, P.-A., Thomas, J.-M.: A mixed finite element method for 2nd order elliptic problems. In: Mathematical Aspects of Finite Element Methods, Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975, pp. 292–315. Lecture Notes in Mathematics, vol. 606. Springer, Berlin (1977). MR MR0483555 (58 #3547)
21. Robin, J.W., Salamon, D.A.: Introduction to differential topology (2011). Lecture notes for a course at ETH Zürich. URL: http://www.math.ethz.ch/~salamon/PREPRINTS/difftop.pdf
22. Sullivan, D.: Differential forms and the topology of manifolds. In: Manifolds—Tokyo 1973, Proc. Internat. Conf., Tokyo, 1973, pp. 37–49. Univ. Tokyo Press, Tokyo (1975). MR MR0370611 (51 #6838)
23. Sullivan, D.: Infinitesimal computations in topology. Publ. Math. IHÉS **1977**(47), 269–331 (1978). MR MR0646078 (58 #31119)
24. Whitney, H.: Geometric Integration Theory. Princeton University Press, Princeton (1957). MR MR0087148 (19,309c)

A Priori Bounds for Solutions of a Nonlocal Evolution PDE

Luis Caffarelli and Enrico Valdinoci

Abstract We obtain L^∞ bounds, independent of ε for the equation

$$u_t^{(\varepsilon)} = \Delta\big(u^{(\varepsilon)} + f\big(u^{(\varepsilon)}\big)\big) - \varepsilon L\big(\Delta u^{(\varepsilon)}\big).$$

Here, f is supposed to be constant outside a bounded interval and L is a suitable nonlocal operator, such as the fractional Laplacian.

1 Introduction

Given $\varepsilon > 0$, we consider the evolution PDE

$$\begin{cases} u_t^{(\varepsilon)} = \Delta\big(u^{(\varepsilon)} + f\big(u^{(\varepsilon)}\big)\big) - \varepsilon L\big(\Delta u^{(\varepsilon)}\big), \\ u^{(\varepsilon)}(x,0) = u_0^{(\varepsilon)}(x). \end{cases} \tag{1}$$

Here above $u = u(x,t)$, $x \in \mathbb{R}^n$, $t \in (0,+\infty)$. As usual, $u_t = \partial_t u$ is the derivative with respect to the time variable t, and Δ is the Laplace operator in the space variable x (that is, the sum of the pure second derivatives). We suppose that $f \in C^2(\mathbb{R})$ and that L is an integral operator of the type that we now describe. Equation (1) may be seen as a nonlocal variant of the classical Cahn-Hilliard phase coexistence model. In the classical (i.e., local case) the model deals with a chemical mixture that, while cooling down, separates into two different phases.

L. Caffarelli (✉)
Department of Mathematics, University of Texas at Austin, 1 University Station C1200, Austin, TX 78712-0257, USA
e-mail: caffarel@math.utexas.edu

E. Valdinoci
Dipartimento di Matematica "F. Enriques", Università di Milano, via Saldini 50, 20133 Milan, Italy
e-mail: enrico@math.utexas.edu

E. Valdinoci
Istituto di Matematica Applicata e Tecnologie Informatiche "E. Magenes", Consiglio Nazionale delle Ricerche, via Ferrata 1, 27100 Pavia, Italy

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*, Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_10,

The modification in (1) takes into account a nonlocal operator that drives the phase segregation (this may be due, for instance, to long-range particle interactions or to boundary effects of the container).

The recent literature has taken into consideration several types of nonlocal versions of the Cahn-Hilliard equation, also for concrete applications, such as the denoising, segmentation and reconstruction of signals in image processing, in order to reduce the unpleasant effect of the white noises, see e.g. [5]. Other applications also occur in fluid dynamics, see [2, 3] and references therein. In this paper we consider quite a severe type of nonlocal feature of the equation, namely the one driven by a fractional Laplacian type operator. When $\varepsilon = 0$, the equation in (1) may be ill-posed, so the nonlocal operator plays also a stabilising effect (though oscillations may increase as $\varepsilon \to 0^+$). Fixed

$$a \in \left(\frac{1}{2}, 1\right), \tag{2}$$

we consider a measurable, $(n+2a)$-positive homogeneous kernel $K : \mathbb{R}^n \setminus \{0\} \to (0, +\infty)$ that is bounded and bounded from zero on S^{n-1}: hence

$$K(\lambda x) = \frac{K(x)}{\lambda^{n+2a}} \tag{3}$$

for any $\lambda > 0$ and any $x \in \mathbb{R}^n \setminus \{0\}$, and

$$\frac{c}{|x|^{n+2a}} \leqslant K(x) \leqslant \frac{C}{|x|^{n+2a}} \tag{4}$$

for suitable $C \geqslant c > 0$. The study of this type of kernels is very popular in singular integrals (see, e.g., [6]) and the fractional Laplacian is a particular case of interest (in such a case, $K(y) = |x|^{-n-2a}$, up to normalisation factors). In relation with the PDE in (1), such kernels play an important role for long-range interaction systems in statistical mechanics (see, e.g., [4]).

We consider the spaces

$$\mathscr{D} := \left\{ v \in C^2\big(\mathbb{R}^n\big) \text{ s.t. } \frac{v(x)}{1+|x|^{n+2a}} \in L^1\big(\mathbb{R}^n\big) \right\}$$

and, for any $\ell \in \mathbb{N}$,

$$\mathscr{D}_\ell := \big\{ v : \mathbb{R}^n \to \mathbb{R} \text{ s.t. } D^\beta v \in \mathscr{D} \text{ for any } \beta \in \mathbb{N}^n \text{ with } \beta_1 + \cdots + \beta_\ell \leqslant \ell \big\}.$$

Notice that $\mathscr{D}_0 = \mathscr{D}$. For any $v \in \mathscr{D}$, we define

$$(Lv)(x) := \int_{\mathbb{R}^n} \big(v(x+y) + v(x-y) - 2v(x)\big) K(y)\, dy. \tag{5}$$

Notice that the definition in (5) is well-posed, thanks to (4). We also define

$$\mathscr{W} := \big\{ v \in C^0\big(\mathbb{R}^n \times [0,+\infty)\big) \text{ s.t.}$$
$$x \mapsto v(x,t) \in \mathscr{D}_2 \text{ for any fixed } t > 0 \text{ and}$$
$$t \mapsto v(x,t) \in C^1\big((0,+\infty)\big) \text{ for any fixed } x \in \mathbb{R}^n \big\}.$$

The main result we prove here is the following uniform bound, independent of ε, for solutions[1] of (1):

Theorem 1.1 *Suppose that*

$$f'(r) = 0 \quad \text{if } |r| \geqslant 1. \tag{6}$$

Let $u^{(\varepsilon)} \in \mathscr{W}$ be a sequence of solutions of (1) with

$$\sup_{\substack{x\in\mathbb{R}^n \\ \varepsilon>0}} \big|u_0^{(\varepsilon)}(x)\big| + \varepsilon^{1/(2a)}\big|\nabla u_0^{(\varepsilon)}(x)\big| < +\infty. \tag{7}$$

Then, there exists $C > 0$ such that

$$\big|u^{(\varepsilon)}(x,t)\big| \leqslant C$$

for any $x \in \mathbb{R}^n$ and any $t \in [0,+\infty)$.

Of course, the C in the statement of Theorem 1.1 is not universal, but it depends on the quantity in (7) (as well as on n, a and the structural constants of K).

We will obtain Theorem 1.1 from scaling: if we set

$$v^{(\varepsilon)}(x,t) := u^{(\varepsilon)}\big(\varepsilon^{1/(2a)}x, \varepsilon^{1/a}t\big)$$

and $v_0^{(\varepsilon)}(x) := v^{(\varepsilon)}(x,0)$, we deduce from (3) that

$$\begin{aligned}
&\Delta\big[v^{(\varepsilon)}(x,t) + f\big(v^{(\varepsilon)}(x,t)\big)\big] - L\big(\Delta v^{(\varepsilon)}\big)(x,t)\\
&\quad= \varepsilon^{1/a}\big[\Delta u^{(\varepsilon)}\big(\varepsilon^{1/(2a)}x, \varepsilon^{1/a}t\big) + \big(\Delta f\big(u^{(\varepsilon)}\big)\big)\big(\varepsilon^{1/(2a)}x, \varepsilon^{1/a}t\big)\\
&\qquad - \varepsilon L\big(\Delta u^{(\varepsilon)}\big)\big(\varepsilon^{1/(2a)}x, \varepsilon^{1/a}t\big)\big]\\
&\quad= \varepsilon^{1/a} u_t^{(\varepsilon)}\big(\varepsilon^{1/(2a)}x, \varepsilon^{1/a}t\big)\\
&\quad= v_t^{(\varepsilon)}(x,t).
\end{aligned}$$

That is, $v^{(\varepsilon)}$ solves the PDE in (1) with $\varepsilon := 1$. Also, by (7),

$$\sup_{\substack{x\in\mathbb{R}^n \\ \varepsilon>0}} \big|v_0^{(\varepsilon)}(x)\big| + \big|\nabla v_0^{(\varepsilon)}(x)\big| < +\infty.$$

[1] For concreteness, in this paper, the solutions are taken in the classical sense (e.g. they are smooth, according to the definition of $\mathscr{W}$).

Consequently, Theorem 1.1 would easily follow by scaling the following result:

Theorem 1.2 *Suppose that* (6) *holds and let* $u \in \mathscr{W}$ *be a solution of*

$$\begin{cases} u_t = \Delta\big(u + f(u)\big) - L(\Delta u), \\ u(x,0) = u_0(x), \end{cases} \tag{8}$$

with $u_0 \in W^{1,\infty}(\mathbb{R}^n)$.

Then, there exists $C > 0$ *such that*

$$\big|u(x,t)\big| \leqslant C \tag{9}$$

for any $x \in \mathbb{R}^n$ *and any* $t \in [0, +\infty)$.

The proof of Theorem 1.2 relies on some techniques developed in [1], with the suitable modifications needed here to take into account the nonlocal operator L.

Roughly speaking, the main idea borrowed from [1] is to split u into two terms, one, say u_1, which is driven by the heat equation, and a remainder u_2. The kernel representing u_2 behaves like the heat kernel for large t, but the nonlocal operator plays a role for small t. The contribution of the kernel are studied via Fourier analysis in order to obtain the desired bounds.

The organisation of the paper is as follows. Sections 2 and 3 are devoted to the Fourier analysis of the operator, which will be applied in Sect. 4 to obtain several useful bounds on the kernel.

The gradient of the solution will be estimated in Sect. 5 (in fact, the gradient bound of Theorem 5.3 may be of interest in itself).

The proof of Theorem 1.2 is contained in Sect. 6 and it is obtained by first controlling the size of u_2, thanks to the kernel estimates, and then the one of u_1, exploiting the gradient bound of Sect. 5 and the standard maximum principle for the heat equation (assumption (6) makes it possible to reduce to it).

2 Fourier Multipliers of K

As usual, the Schwartz space of rapidly decreasing functions will be denoted by $\mathcal{S}$.

Given $\phi \in \mathcal{S}$, its Fourier transform will be denoted by either $\widehat{\phi}$ or $\mathscr{F}\phi$, and the anti-Fourier transform by either $\check{\phi}$ or $\mathscr{F}^{-1}\phi$.

Given $v, w \in L^2(\mathbb{R}^n, \mathbb{C})$, we define

$$\langle v, w\rangle := \int_{\mathbb{R}^n} v(x)w(x)dx.$$

Given a tempered distribution $T \in \mathcal{S}^\star$, we also denote by $\langle T, \phi\rangle$ its action on a test function $\phi \in \mathcal{S}$: of course, no ambiguity should arise from the use of the notation $\langle\cdot,\cdot\rangle$ in the two settings.

As usual, if $w \in L^1_{\rm loc}(\mathbb{R}^n)$, its Fourier transform is meant in the distribution sense, i.e.

$$\langle \widehat{w}, \phi \rangle := \langle w, \widehat{\phi} \rangle \quad \text{for any } \phi \in \mathcal{S}.$$

Also, if $f = f(x, y) : X \times Y \to \mathbb{R}$ we write

$$\int_X \int_Y f(x, y) dy dx$$

to mean

$$\int_X \left(\int_Y f(x, y) dy \right) dx.$$

With this notation, we can perform a Fourier analysis on the operator L. We start by finding an explicit representation of the symbol of L in Fourier space and by observing its natural scaling properties.

Lemma 2.1 *For any $v \in \mathscr{D}$,*

$$Lv \in L^1_{\rm loc}\big(\mathbb{R}^n\big) \tag{10}$$

and

$$\mathscr{F}(Lv) = S\widehat{v} \quad \textit{with } S(\xi) := \frac{1}{2(2\pi)^n} \int_{\mathbb{R}^n} \big(\cos(y \cdot \xi) - 1\big) K(y) dy \leqslant 0. \tag{11}$$

Also,

$$-S(\xi) \geqslant \frac{1}{C_\star} |\xi|^{2a} \quad \textit{for any } \xi \in \mathbb{R}^n \setminus B_1, \tag{12}$$

and, for any $\beta \in \mathbb{N}^n$,

$$\big| D^\beta S(\xi) \big| \leqslant C_\star |\xi|^{2a - (\beta_1 + \cdots + \beta_n)} \quad \textit{for any } \xi \neq 0, \tag{13}$$

for a suitable $C_\star \geqslant 1$.

Proof We start by proving (10). For this, fix $R > 0$.

We observe that if $z \in \mathbb{R}^n \setminus B_{2R}$ and $x \in B_R$, then

$$|z \pm x| \geqslant |z| - |x| = \frac{|z|}{2} + \frac{|z|}{2} - |x| \geqslant \frac{|z|}{2} + \frac{2R}{2} - R = \frac{|z|}{2}.$$

Therefore, recalling (4), we see that, if $z \in \mathbb{R}^n \setminus B_{2R}$ and $x \in B_R$,

$$K(z - x) + K(z + x) + 2K(z) \leqslant \frac{C_0}{|z|^{n+2a}},$$

for a suitable $C_0 > 0$.

Consequently,

$$\begin{aligned}
&\int_{B_R}\int_{\mathbb{R}^n\setminus B_{3R}} \big|v(x+y)+v(x-y)-2v(x)\big|K(y)dydx\\
&\quad\leqslant \int_{B_R}\int_{\mathbb{R}^n\setminus B_{3R}} \big|v(x+y)\big|K(y)dydx+\int_{B_R}\int_{\mathbb{R}^n\setminus B_{3R}} \big|v(x-y)\big|K(y)dydx\\
&\qquad+2\int_{B_R}\int_{\mathbb{R}^n\setminus B_{3R}} \big|v(y)\big|K(y)dydx\\
&\quad= \int_{B_R}\int_{\mathbb{R}^n\setminus B_{3R}(x)} \big|v(z)\big|K(z-x)dzdx+\int_{B_R}\int_{\mathbb{R}^n\setminus B_{3R}(-x)} \big|v(z)\big|K(z+x)dzdx\\
&\qquad+2\int_{B_R}\int_{\mathbb{R}^n\setminus B_{3R}} \big|v(z)\big|K(z)dzdx\\
&\quad\leqslant \int_{B_R}\int_{\mathbb{R}^n\setminus B_{2R}} \big|v(z)\big|\big(K(z-x)+K(z+x)+2K(x)\big)dzdx\\
&\quad\leqslant \int_{B_R}\int_{\mathbb{R}^n\setminus B_{2R}} \frac{C_0|v(z)|}{|z|^{n+2a}}dzdx\\
&\quad\leqslant C_R,
\end{aligned}$$

for a suitable $C_R>0$, since $v\in\mathscr{D}$.

Accordingly, using (4) once more,

$$\begin{aligned}
\|Lv\|_{L^1(B_R)} &\leqslant \int_{B_R}\int_{B_{3R}} \big|v(x+y)+v(x-y)-2v(x)\big|K(y)dydx\\
&\qquad+\int_{B_R}\int_{\mathbb{R}^n\setminus B_{3R}} \big|v(x+y)+v(x-y)-2v(x)\big|K(y)dydx\\
&\leqslant \int_{B_R}\int_{B_{3R}} \|v\|_{C^2(B_{4R})}|y|^2K(y)dydx+C_R\\
&\leqslant C'_R\left(\int_{B_{3R}}|y|^{2-(n+2a)}+1\right)\\
&\leqslant C'_R\left(\frac{\omega_{n-1}(3R)^{2(1-a)}}{2(1-a)}+1\right),
\end{aligned}$$

for a suitable $C'_R>0$, and this proves (10).

Now, we prove (11). If $v\in\mathcal{S}$, a direct computation (see, e.g., Sect. 2 of [7]) proves (11).

Let us now check that (11) holds for any $v\in\mathscr{D}$. For this, first notice that $\mathscr{F}(Lv)$ is well-defined, in the distribution sense, thanks to (10).

Given a function w, we define $\widetilde{w}(x):=w(-x)$ and we observe that $\widetilde{Lv}=L\widetilde{v}$.

Take any $\phi \in \mathcal{S}$ and let $\psi := \check{\phi}$. We remark that $\mathscr{F}(L\psi) = S\widehat{\psi}$, since we know that (11) holds in $\mathcal{S}$, and so

$$\begin{aligned}\langle S\widehat{v}, \phi\rangle &= \langle \widehat{v}, S\phi\rangle = \langle \widehat{v}, S\widehat{\psi}\rangle \\ &= \big\langle \widehat{v}, \mathscr{F}(L\psi)\big\rangle = \big\langle v, \mathscr{F}^2(L\psi)\big\rangle = \langle v, \widetilde{L\psi}\rangle \\ &= \langle \widetilde{v}, L\psi\rangle = \langle L\widetilde{v}, \psi\rangle = \langle \widetilde{Lv}, \psi\rangle \\ &= \langle Lv, \widetilde{\psi}\rangle = \big\langle Lv, \mathscr{F}^2\psi\big\rangle = \big\langle Lv, \mathscr{F}^2\check{\phi}\big\rangle \\ &= \langle Lv, \mathscr{F}\phi\rangle = \big\langle \mathscr{F}(Lv), \phi\big\rangle.\end{aligned}$$

This completes the proof of (11) when $v \in \mathscr{D}$.

Now we consider $\xi \neq 0$, we set $\vec{\xi} := \xi/|\xi|$, we take a rotation R such that $\vec{\xi} = Re_1$, and we change variable $w := R^T(|\xi|y)$: since $dw = |\xi|^n dy$, we obtain in (11) that

$$\begin{aligned}-2(2\pi)^n S(\xi) &= \int_{\mathbb{R}^n} \big(1 - \cos(y\cdot\xi)\big)K(y)dy \\ &= \int_{\mathbb{R}^n} \big(1 - \cos\big(y|\xi|\cdot(Re_1)\big)\big)K(y)dy \\ &= |\xi|^{-n}\int_{\mathbb{R}^n} \big(1 - \cos(w_1)\big)K\big(|\xi|^{-1}Rw\big)dy \\ &= c_\star|\xi|^{2a}, \quad \text{with } c_\star := \int_{\mathbb{R}^n} \big(1 - \cos(w_1)\big)K(Rw)dy. \end{aligned} \tag{14}$$

Noticed that we have used (3) here, and that $c_\star \in (0, +\infty)$ thanks to (4). Then, (12) and (13) follow from (14). □

3 Fourier Analysis

This is the most technical part of the paper, in which some techniques of [1] need to be adapted to our case by taking into account the different natural scaling of the operator. The estimates obtained will be uniform in time, and this requires a different analysis for the short and long times asymptotics.

Lemma 3.2 *Let S be as in (11). Let $P : \mathbb{R}^n \to \mathbb{R}$ be a polynomial. Given $\gamma \in \{0, a\}$, we define the interval I_γ as*

$$I_\gamma := \begin{cases} [1, +\infty) & \text{if } \gamma = 0, \\ (0, 1] & \text{if } \gamma = a. \end{cases} \tag{15}$$

Then, there exists $C_\star > 0$ such that

$$\sup_{t\in I_\gamma} \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} P(\eta) e^{-t^{\gamma/(\gamma+1)}|\eta|^2(1-S(t^{-1/(2(\gamma+1))}\eta))} e^{i\eta\cdot z} d\eta \right| dz \leqslant C_\star. \tag{16}$$

Proof Let

$$\lambda = \lambda_\gamma := \frac{1}{2(\gamma+1)}.$$

We observe that

$$\sup_{t\in I_\gamma} t^{2(\gamma-a)\lambda} \leqslant \max\left\{ \sup_{t\in[1,+\infty)} t^{-a}, \sup_{t\in(0,1]} t^0 \right\} = 1. \tag{17}$$

Also, a direct computation shows that

$$e^{t^{2\gamma\lambda}|\eta|^2(1-S(t^{-\lambda}\eta))} \cdot \sum_{\substack{\beta\in\mathbb{N}^n \\ \beta_1+\cdots+\beta_n\leqslant 2n}} \left| \left(\frac{\partial}{\partial\eta}\right)^\beta e^{-t^{2\gamma\lambda}|\eta|^2(1-S(t^{-\lambda}\eta))} \right|$$

is bounded by an appropriate polynomial in the variables $|\eta|$, $t^{2\gamma\lambda}$ and $t^{(2\gamma-m)\lambda} \times |\partial^\beta S(t^{-\lambda}\eta)|$, for $m = \beta_1 + \cdots + \beta_n \leqslant 2n$.

That is, recalling (13), it is bounded by a polynomial in the variables $|\eta|$, $t^{2\gamma\lambda}$ and $t^{2(\gamma-a)\lambda}$.

Accordingly, due to (15) and (17), for any $t \in I_\gamma$, it is bounded by a polynomial in the variable $|\eta|$, and so

$$\sup_{t\in I_\gamma} \sum_{\substack{\beta\in\mathbb{N}^n \\ \beta_1+\cdots+\beta_n\leqslant 2n}} \left| \left(\frac{\partial}{\partial\eta}\right)^\beta e^{-t^{2\gamma\lambda}|\eta|^2(1-S(t^{-\lambda}\eta))} \right| \leqslant \bar{C}\big(1+|\eta|^{\bar{C}}\big) e^{-t^{2\gamma\lambda}|\eta|^2(1-S(t^{-\lambda}\eta))},$$

for a suitable $\bar{C} > 0$.

Hence, if P is as in the statement of Lemma 3.2,

$$\begin{aligned} &\sup_{t\in I_\gamma} \sum_{\substack{\beta\in\mathbb{N}^n \\ \beta_1+\cdots+\beta_n\leqslant 2n}} \left| \left(\frac{\partial}{\partial\eta}\right)^\beta \big(P(\eta) e^{-t^{2\gamma\lambda}|\eta|^2(1-S(t^{-\lambda}\eta))}\big) \right| \\ &\qquad \leqslant \widetilde{C}\big(1+|\eta|^{\widetilde{C}}\big) e^{-t^{2\gamma\lambda}|\eta|^2(1-S(t^{-\lambda}\eta))}, \end{aligned} \tag{18}$$

for some $\widetilde{C} > 0$.

We now localise the integral in (16). For this, we fix a function $\phi \in C^\infty(\mathbb{R}, [0,1])$ such that $\phi = 1$ in $[-1,1]$ and $\phi = 0$ outside $[-2,2]$. We set $\psi := 1 - \phi$ and we write

$$1 = \prod_{j=1}^n \big(\psi(z_j) + \phi(z_j)\big).$$

So the integrand in (16) becomes

$$\int_{\mathbb{R}^n}\prod_{j=1}^{n}\big(\psi(z_j)+\phi(z_j)\big)\left|\int_{\mathbb{R}^n}P(\eta)e^{-t^{\gamma/(\gamma+1)}|\eta|^2(1-S(t^{-1/(2(\gamma+1))}\eta))}e^{i\eta\cdot z}d\eta\right|dz.$$

After expanding the sums in the product, we see (by possibly reordering the coordinates) that the "typical" term has the form

$$\int_{\mathbb{R}^n}\prod_{j=1}^{k}\psi(z_j)\prod_{\ell=k+1}^{n}\phi(z_\ell)\left|\int_{\mathbb{R}^n}P(\eta)e^{-t^{\gamma/(\gamma+1)}|\eta|^2(1-S(t^{-1/(2(\gamma+1))}\eta))}e^{i\eta\cdot z}d\eta\right|dz,$$

for $k=0,\dots,n$, with the obvious notation that $\prod_{j=m}^{q}$ equals 1 whenever $q<m$.

We integrate by parts $2k$-times, concluding that the above quantity equals to

$$\begin{aligned}
&\int_{\mathbb{R}^n}\left|\int_{\mathbb{R}^n}P(\eta)e^{-t^{\gamma/(\gamma+1)}|\eta|^2(1-S(t^{-1/(2(\gamma+1))}\eta))}\right.\\
&\qquad\left.\times\prod_{j=1}^{k}\big(\psi(z_j)e^{i\eta_j z_j}\big)\prod_{\ell=k+1}^{n}\phi(z_\ell)e^{i\eta_\ell z_\ell}d\eta\right|dz\\
&=\int_{\mathbb{R}^n}\left|\int_{\mathbb{R}^n}P(\eta)e^{-t^{\gamma/(\gamma+1)}|\eta|^2(1-S(t^{-1/(2(\gamma+1))}\eta))}\right.\\
&\qquad\left.\times\prod_{j=1}^{k}\left(\frac{\psi(z_j)}{(iz_j)^2}\left(\frac{\partial}{\partial\eta_j}\right)^2e^{i\eta_j z_j}\right)\prod_{\ell=k+1}^{n}\phi(z_\ell)e^{i\eta_\ell z_\ell}d\eta\right|dz\\
&=\int_{\mathbb{R}^n}\left|\int_{\mathbb{R}^n}\left(\frac{\partial}{\partial\eta_1}\right)^2\cdots\left(\frac{\partial}{\partial\eta_k}\right)^2\big[P(\eta)e^{-t^{\gamma/(\gamma+1)}|\eta|^2(1-S(t^{-1/(2(\gamma+1))}\eta))}\big]\right.\\
&\qquad\left.\times\prod_{j=1}^{k}\left(\frac{\psi(z_j)}{(iz_j)^2}e^{i\eta_j z_j}\right)\prod_{\ell=k+1}^{n}\phi(z_\ell)e^{i\eta_\ell z_\ell}d\eta\right|dz\\
&=\int_{\mathbb{R}^n}\prod_{j=1}^{k}\frac{\psi(z_j)}{|z_j|^2}\prod_{\ell=k+1}^{n}\phi(z_\ell)\\
&\qquad\times\left|\int_{\mathbb{R}^n}\left(\frac{\partial}{\partial\eta_1}\right)^2\cdots\left(\frac{\partial}{\partial\eta_k}\right)^2\big[P(\eta)e^{-t^{\gamma/(\gamma+1)}|\eta|^2(1-S(t^{-1/(2(\gamma+1))}\eta))}\big]e^{i\eta\cdot z}d\eta\right|dz.
\end{aligned}$$

This and (18) imply that, for any $t\in I_\gamma$, the integrand in (16) is bounded by

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\prod_{j=1}^{k}\frac{\psi(z_j)}{|z_j|^2}\prod_{\ell=k+1}^{n}\phi(z_\ell)\big(1+|\eta|^{\widetilde{C}}\big)e^{-t^{2\gamma\lambda}|\eta|^2(1-S(t^{-\lambda}\eta))}d\eta dz, \tag{19}$$

where the multiplicative constant has been neglected.

Now, we claim that

$$t^{2\gamma\lambda}|\eta|^2\big(1-S\big(t^{-\lambda}\eta\big)\big)\geqslant\frac{1}{C_*}|\eta|^{2(1+\gamma)}, \tag{20}$$

for a suitable $C_*>0$.

Indeed, if $t^{-\lambda}|\eta|\leqslant 1$ we use (11) and we obtain

$$t^{2\gamma\lambda}|\eta|^2\big(1-S\big(t^{-\lambda}\eta\big)\big)\geqslant t^{2\gamma\lambda}|\eta|^2\geqslant|\eta|^{2(1+\gamma)},$$

proving (20) in this case. If, conversely, $t^{-\lambda}|\eta|>1$, we exploit (12) and we obtain

$$t^{2\gamma\lambda}|\eta|^2\big(1-S\big(t^{-\lambda}\eta\big)\big)\geqslant -t^{2\gamma\lambda}|\eta|^2S\big(t^{-\lambda}\eta\big)\geqslant\frac{t^{2\lambda(\gamma-a)}|\eta|^{2(1+a)}}{C_\star}\geqslant|\eta|^{2(1+\gamma)},$$

which is (20) in this case.

As a consequence of (19) and (20), for any $t\in I_\gamma$, we control the integrand in (16) with

$$\underline{C}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\prod_{j=1}^{k}\frac{\psi(z_j)}{|z_j|^2}\prod_{\ell=k+1}^{n}\phi(z_\ell)\big(1+|\eta|^{\widetilde{C}}\big)e^{-|\eta|^{2(1+\gamma)}/C_*}d\eta dz,$$

for a suitable $\underline{C}>0$. This plainly yields the desired result, by exploiting the supports of ϕ and ψ. □

As an immediate consequence of (16), we have that:

Corollary 3.1 *If $P,Q:\mathbb{R}^n\to\mathbb{R}$ are polynomials,*

$$\sup_{t\in(0,1]}\int_{\mathbb{R}^n}\left|\int_{\mathbb{R}^n}P(\eta)e^{-t^{a/(a+1)}|\eta|^2(1-S(t^{-1/(2(a+1))}\eta))}e^{i\eta\cdot z}d\eta\right|dz$$
$$+\sup_{t\in[1,+\infty)}\int_{\mathbb{R}^n}\left|\int_{\mathbb{R}^n}Q(\eta)e^{-|\eta|^2(1-S(t^{-1/2}\eta))}e^{i\eta\cdot z}d\eta\right|dz<+\infty. \tag{21}$$

Also, a useful variation of Lemma 3.2 is given by the following result:

Lemma 3.3 *Let S be as in (11). Let $P:\mathbb{R}^n\to\mathbb{R}$ be a polynomial. Then, there exists $C_0>0$ such that, for any $t\geqslant 1$,*

$$\int_{\mathbb{R}^n}\left|\int_{\mathbb{R}^n}P(\eta)e^{-|\eta|^2}\big(e^{|\eta|^2S(t^{-1/2}\eta)}-1\big)e^{i\eta\cdot z}d\eta\right|dz\leqslant\frac{C_0}{t^a}. \tag{22}$$

Proof Though the proof is a simple variation of the one given in Lemma 3.2, we give the technical details for the facility of the reader.

For any $(\eta,t)\in\mathbb{R}^n\times[1,+\infty)$, let

$$g(\eta,t):=e^{|\eta|^2S(t^{-1/2}\eta)}-1.$$

We observe that, from (11), $S(t^{-1/2}\eta) \leqslant 0$ and so, from (13),

$$\begin{aligned}\big|g(\eta,t)\big| &= 1 - e^{|\eta|^2 S(t^{-1/2}\eta)} = \int_{|\eta|^2 S(t^{-1/2}\eta)}^{0} e^s ds \leqslant \int_{|\eta|^2 S(t^{-1/2}\eta)}^{0} 1 ds \\ &= -|\eta|^2 S\big(t^{-1/2}\eta\big) \leqslant \frac{|\eta|^{2(1+a)}}{t^a}. \end{aligned} \tag{23}$$

We take ϕ and ψ as in the proof of Lemma 3.2 and we follow the same arguments as there to see that, after expanding the sums in the product, the "typical" term obtained from the integral in (22) has the form

$$\begin{aligned}&\int_{\mathbb{R}^n} \prod_{j=1}^{k} \psi(z_j) \prod_{\ell=k+1}^{n} \phi(z_\ell) \left| \int_{\mathbb{R}^n} P(\eta) e^{-|\eta|^2} \big(e^{|\eta|^2 S(t^{-1/2}\eta)} - 1\big) e^{i\eta\cdot z} d\eta \right| dz \\ &\quad = \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} P(\eta) e^{-|\eta|^2} g(\eta,t) \left(\prod_{j=1}^{k} \psi(z_j) e^{i\eta_j z_j} \right) \left(\prod_{\ell=k+1}^{n} \phi(z_\ell) e^{i\eta_\ell z_\ell} \right) d\eta \right| dz \\ &\quad = \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} P(\eta) e^{-|\eta|^2} g(\eta,t) \left(\prod_{j=1}^{k} \frac{\psi(z_j)}{(iz_j)^2} \left(\frac{\partial}{\partial \eta_j} \right)^2 e^{i\eta_j z_j} \right) \right. \\ &\qquad \left. \times \left(\prod_{\ell=k+1}^{n} \phi(z_\ell) e^{i\eta_\ell z_\ell} \right) d\eta \right| dz \\ &\quad = \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} \left(\frac{\partial}{\partial \eta_1} \right)^2 \cdots \left(\frac{\partial}{\partial \eta_k} \right)^2 \big[P(\eta) e^{-|\eta|^2} g(\eta,t) \big] \right. \\ &\qquad \left. \times \left(\prod_{j=1}^{k} \frac{\psi(z_j)}{(iz_j)^2} e^{i\eta_j z_j} \right) \left(\prod_{\ell=k+1}^{n} \phi(z_\ell) e^{i\eta_\ell z_\ell} \right) d\eta \right| dz, \end{aligned} \tag{24}$$

where $2k$ integrations by parts have been performed.

Now, an explicit computation gives that

$$e^{|\eta|^2} \sum_{\substack{\beta\in\mathbb{N}^n \\ \beta_1+\cdots+\beta_n \leqslant 2n}} \left| \left(\frac{\partial}{\partial \eta} \right)^{\beta} \big[P(\eta) e^{-|\eta|^2} g(\eta,t) \big] \right| \tag{25}$$

is bounded by

$$P_0(\eta)\big|g(\eta,t)\big| + e^{|\eta|^2 S(t^{-1/2}\eta)} \sum_{q=1}^{2n} P_q(\eta) \left| \sum_{\substack{\beta\in\mathbb{N}^n \\ \beta_1+\cdots+\beta_n = q}} \left(\frac{\partial}{\partial \eta} \right)^{\beta} \big[|\eta|^2 S\big(t^{-1/2}\eta\big) \big] \right|,$$

for suitable polynomials $P_0, P_1, \dots, P_{2n}$, and so, recalling (13), it is bounded by

$$P_0(\eta)\big|g(\eta,t)\big| + \frac{\widetilde{P}(\eta)}{t^a},$$

for a suitable polynomial $\widetilde{P}$.

That is, in the light of (23), the quantity in (25) is bounded by

$$\frac{M(1+|\eta|^M)}{t^a},$$

for a suitably large $M \in \mathbb{N}$.

This and (24) imply that the integral in (22) is bounded by

$$\frac{M}{t^a}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}\prod_{j=1}^{k}\frac{\psi(z_j)}{z_j^2}\prod_{\ell=k+1}^{n}\phi(z_\ell)\big(1+|\eta|^M\big)e^{-|\eta|^2}d\eta dz.$$

Using the supports of ϕ and ψ, we obtain the estimate claimed in (22). □

4 Kernel Estimates

Here, we perform useful estimates on the kernel of the homogeneous equation associated to (8).

For this, we define

$$N(x,t) := \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} e^{-t|\xi|^2(1-S(\xi))}e^{i\xi\cdot x}d\xi,$$

where S is the one introduced in (11).

We will also exploit the following standard notation, for $1 \leqslant p \leqslant \infty$:

$$\|\cdot\|_{L^p_x} := \|\cdot\|_{L^p(\mathbb{R}^n)} \quad \text{and}$$
$$\|\cdot\|_{L^p_t} := \|\cdot\|_{L^p((0,+\infty))}.$$

Also, ∂_j will be a short notation for the spatial derivative $\frac{\partial}{\partial x_j}$, for $j = 1, \dots, n$.

We observe that N is the fundamental solution of the linearised equation:

$$N_t = \Delta N - L(\Delta N) \quad \text{for } t > 0, \qquad \lim_{t\to 0^+} N(x,t) = \delta_0.$$

Scope of these pages is to point out the following integral bound on N:

$$\sup_{t>0}\big\|N(\cdot,t)\big\|_{L^1_x} + \int_0^1 \|\Delta N\|_{L^1_x}dt + \sup_{j\in\{1,\dots,n\}}\int_0^{+\infty}\|\partial_j\Delta N\|_{L^1_x}dt < +\infty.$$

This will be achieved in the following Lemmata 4.4 and 4.5. From the estimate above, we will obtain a gradient bound via convolution in the forthcoming Theorem 5.3. This and a control on the long-time behaviour will finally provide the proof of Theorem 1.2.

Lemma 4.4

$$\sup_{t>0} \left\| N(\cdot,t) \right\|_{L^1_x} < +\infty.$$

Proof Let

$$\omega(t) := \begin{cases} t^{1/(2(a+1))} & \text{if } t \in (0,1), \\ t^{1/2} & \text{if } t \in [1,+\infty). \end{cases} \tag{26}$$

It is useful to observe that if $P, Q : \mathbb{R}^n \to \mathbb{R}$ are polynomials and

$$\mathscr{P}(\eta,t) := \begin{cases} P(\eta) & \text{if } t \in (0,1), \\ Q(\eta) & \text{if } t \in [1,+\infty), \end{cases}$$

then (21) says that

$$\begin{aligned} &\sup_{t\in(0,+\infty)} \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} \mathscr{P}(\eta,t) e^{-t(|\eta|/|\omega(t)|)^2(1-S(\eta/\omega(t)))} e^{i\eta\cdot z} d\eta \right| dz \\ &\quad \leqslant \sup_{t\in(0,1)} \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} P(\eta) e^{-t^{a/(a+1)}|\eta|^2(1-S(t^{-1/(2(a+1))}\eta))} e^{i\eta\cdot z} d\eta \right| dz \\ &\qquad + \sup_{t\in[1,+\infty)} \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} Q(\eta) e^{-|\eta|^2(1-S(t^{-1/2}\eta))} e^{i\eta\cdot z} d\eta \right| dz \\ &\quad < +\infty. \end{aligned} \tag{27}$$

Now, we use the substitution

$$\eta := \omega(t)\xi, \qquad z := x/\big(\omega(t)\big) \tag{28}$$

to conclude that

$$\begin{aligned} \left\| N(\cdot,t) \right\|_{L^1_x} &= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} e^{-t|\xi|^2(1-S(\xi))} e^{i\xi\cdot x} d\xi \right| dx \\ &= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} e^{-t(|\eta|/|\omega(t)|)^2(1-S(\eta/\omega(t)))} e^{i\eta\cdot z} d\eta \right| dz. \end{aligned}$$

This and (27) (applied here with $\mathscr{P} := 1$) readily give the desired result. □

A useful result, analogous to Lemma 4.4 is the following:

Lemma 4.5

$$\int_0^1 \|\Delta N\|_{L^1_x} dt < +\infty \tag{29}$$

and

$$\sup_{j\in\{1,\dots,n\}} \int_0^{+\infty} \|\partial_j \Delta N\|_{L^1_x} dt < +\infty. \tag{30}$$

Proof The proof is an appropriate modification of the one of Lemma 4.4. Let us first deal with (30). Fix $j \in \{1,\dots,n\}$. We have that

$$\left|\sum_{k=1}^n \partial_{jkk} N(x,t)\right| = \left|\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} \xi_j |\xi|^2 e^{-t|\xi|^2(1-S(\xi))} e^{i\xi\cdot x} d\xi\right|.$$

Thus, we use once more the substitution introduced in (26) and (28) to obtain that

$$\begin{aligned}
&\left\|\sum_{k=1}^n \partial_{jkk} N(\cdot,t)\right\|_{L^1_x} \\
&\quad\leqslant \frac{1}{(\omega(t))^3(2\pi)^n}\int_{\mathbb{R}^n}\left|\int_{\mathbb{R}^n} \eta_j|\eta|^2 e^{-t(|\eta|/|\omega(t)|)^2(1-S(\eta/\omega(t)))} e^{i\eta\cdot z} d\eta\right| dz \\
&\quad\leqslant \frac{C(j)}{(\omega(t))^3},
\end{aligned}$$

for a suitable $C(j) > 0$, thanks to (27), applied here with

$$\mathscr{P}(\eta,t) := \eta_j|\eta|^2.$$

As a consequence,

$$\begin{aligned}
\int_0^{+\infty}\left\|\sum_{k=1}^n \partial_{jkk} N\right\|_{L^1_x} dt &\leqslant \int_0^{+\infty} \frac{C(j)dt}{(\omega(t))^3} \\
&= \int_0^1 \frac{C(j)dt}{t^{3/(2(a+1))}} + \int_1^{+\infty} \frac{C(j)dt}{t^{3/2}} \leqslant \widetilde{C}(j),
\end{aligned}$$

for a suitable $\widetilde{C}(j) > 0$. Notice that here we used the fact that $3/(2(a+1)) < 1$, thanks to (2). The above estimate implies (30).

The proof of (29) is analogous and, in fact, much simpler: just consider in this case only the integral in $t \in (0,1)$, drop η_j, and replace $\omega^3(t)$ with $\omega^2(t) = t^{1/(a+1)}$. □

5 Gradient Bounds

Now, we apply the previous estimates to deduce a uniform gradient bound:

Theorem 5.3 *Let $u \in \mathscr{W}$ be a solution of*

$$\begin{cases} u_t = \Delta(u+g) - L(\Delta u), \\ u(x,0) = u_0(x), \end{cases} \tag{31}$$

with $g = g(x,t) \in L^\infty(\mathbb{R}^n \times [0,+\infty))$, $g(\cdot,t) \in C^2(\mathbb{R}^n)$ for any $t > 0$, and $\nabla u_0 \in (L^\infty_x)^n$.

Then,

$$\sup_{t\in(0,+\infty)} \|\nabla u(\cdot,t)\|_{(L^\infty_x)^n} \leqslant C_\star\big(\|\nabla u_0\|_{(L^\infty_x)^n} + \|g\|_{L^\infty(\mathbb{R}^n\times(0,+\infty))}\big)$$

for a suitable $C_\star > 0$.

Proof Let $h(x,t) := \Delta(g(x,t))$ and $h_0(x) := h(x,0)$. We remark that, for fixed $t > 0$, all the terms of the PDE in (31) are in $L^1_{\mathrm{loc}}(\mathbb{R}^n)$, because of (10). Consequently, by Fourier transforming (31),

$$\widehat{u}_t = -\kappa\widehat{u} + \widehat{h},$$

with

$$\kappa(\xi) := |\xi|^2\big(1 - S(\xi)\big),$$

thanks to (11).

So, by solving the ODE,

$$\widehat{u}(\xi,t) = \int_0^t \widehat{h}(\xi,s)e^{-\kappa(\xi)(t-s)}ds + \widehat{u}_0(\xi)e^{-\kappa(\xi)t}$$

and then, by antitransforming,

$$u = N \ast\ast h + N * u_0,$$

where $*$ denotes the convolution in $x \in \mathbb{R}^n$ and $\ast\ast$ the convolution in $x \in \mathbb{R}^n$ and t (up to a finite time).

As a consequence,

$$\partial_j u = \partial_j N \ast\ast h + N * \partial_j u_0. \tag{32}$$

Now, we recall Lemma 4.4, according to which $\|N(\cdot,t)\|_{L^1_x} \leqslant C$, for any $t > 0$, for a suitable $C > 0$.

Therefore, for any $x \in \mathbb{R}^n$ and any $t > 0$,

$$\big|(N * \partial_j u_0)(x,t)\big| \leqslant \big\|N(\cdot,t)\big\|_{L^1_x} \|\partial_j u_0\|_{L^\infty_x} \leqslant C\|\nabla u_0\|_{(L^\infty_x)^n}. \tag{33}$$

Moreover, possibly renaming C, by (30),

$$\int_0^{+\infty} \|\partial_j \Delta N\|_{L^1_x} dt \leqslant C,$$

and so, for any $x \in \mathbb{R}^n$ and any $t > 0$,

$$\begin{aligned}
\big|(\partial_j N ** h)(x,t)\big| &= \big|(\partial_j \Delta N) ** g(x,t)\big| \\
&\leqslant \int_0^{+\infty} \left[\int_{\mathbb{R}^n} \big|(\partial_j \Delta N)(y,s)\big| \big|g(x-y,t-s)\big| dy \right] ds \\
&\leqslant \int_0^{+\infty} \|\partial_j \Delta N\|_{L^1_x} \|g\|_{L^\infty(\mathbb{R}^n \times (0,+\infty))} ds \\
&\leqslant C\|g\|_{L^\infty(\mathbb{R}^n \times (0,+\infty))}.
\end{aligned} \tag{34}$$

Then, the desired claim follows from (32), (33) and (34). □

6 Proof of Theorem 1.2

Let $g(x,t) := f(u(x,t))$ and $h = h(x,t) := \Delta g$. We observe that, for fixed $t > 0$, all the terms of the PDE in (8) are in $L^1_{\rm loc}(\mathbb{R}^n)$, due to (10). Therefore, we may Fourier transform (8) and solve the associated ODE (in analogy with what we did in the beginning of the proof of Theorem 5.3), to obtain that

$$u = N ** h + N * u_0,$$

where $*$ denotes the convolution in $x \in \mathbb{R}^n$ and $**$ the convolution in $x \in \mathbb{R}^n$ and t (up to a finite time).

Hence, if we define the heat kernel

$$H(x,t) := \frac{1}{(2\sqrt{\pi t})^n} e^{-|x|^2/4t}$$

and

$$G := N - H,$$

we obtain that

$$u = u_1 + u_2,$$

with

$$u_1 = H **h + H * u_0 \tag{35}$$

and

$$u_2 = G **h + G * u_0.$$

We remark that

$$\begin{aligned}
&\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} e^{-|\eta|^2}\big(e^{|\eta|^2 S(t^{-1/2}\eta)} - 1\big)e^{i\eta\cdot x}d\eta \\
&\quad = \left(\frac{\sqrt{t}}{2\pi}\right)^n \int_{\mathbb{R}^n} e^{-t|\xi|^2}\big(e^{t|\xi|^2 S(\xi)} - 1\big)e^{i\xi\cdot(t^{1/2}x)}d\xi \\
&\quad = t^{n/2}\big(N\big(t^{1/2}x,t\big) - H\big(t^{1/2}x,t\big)\big) \\
&\quad = t^{n/2}G\big(t^{1/2}x,t\big),
\end{aligned}$$

thanks to the substitution $\xi := t^{-1/2}\eta$ and to the standard Fourier transform property of H.

Therefore, for any $t \geqslant 1$, by (22),

$$\begin{aligned}
\big\|\Delta G(\cdot,t)\big\|_{L^1_x} &= \int_{\mathbb{R}^n} t^{n/2}\big|\Delta G\big(t^{1/2}z,t\big)\big|dz \\
&= \frac{1}{(2\pi)^n t}\int_{\mathbb{R}^n}\left|\int_{\mathbb{R}^n}|\eta|^2 e^{-|\eta|^2}\big(e^{|\eta|^2 S(t^{-1/2}\eta)} - 1\big)e^{i\eta\cdot z}d\eta\right|dz \\
&\leqslant \frac{C_0}{t^{1+a}},
\end{aligned} \tag{36}$$

for a suitable $C_0 > 0$.

Now, we give a uniform bound for u_2. For this, first we use Lemma 4.4 (together with a standard property of the heat kernel) to obtain that

$$\begin{aligned}
\sup_{\substack{x\in\mathbb{R}^n\\ t>0}}\big|G * u_0(x,t)\big| &\leqslant \sup_{\substack{x\in\mathbb{R}^n\\ t>0}}\big|N * u_0(x,t)\big| + \sup_{\substack{2\\ x\in\mathbb{R}^n\, t>0}}\big|H * u_0(x,t)\big| \\
&\leqslant \sup_{t>0}\big\|N(\cdot,t)\big\|_{L^1_x}\|u_0\|_{L^\infty_x} + \sup_{t>0}\big\|H(\cdot,t)\big\|_{L^1_x}\|u_0\|_{L^\infty_x} \\
&\leqslant C_1\|u_0\|_{L^\infty_x},
\end{aligned} \tag{37}$$

for a suitable $C_1 > 0$.

On the other hand,

$$\int_0^1 \big\|\partial_k H(\cdot,s)\big\|_{L^1_x} ds = \frac{1}{2^{n+1}\pi^{n/2}} \int_0^1 \left| \int_{\mathbb{R}^n} \frac{x_k e^{-|x|^2/(4s)}}{s^{(n/2)+1}} dx \right| ds$$
$$\leqslant \frac{1}{2^{n+1}\pi^{n/2}} \int_0^1 \int_{\mathbb{R}^n} \frac{|z| e^{-|z|^2/4}}{\sqrt{s}} dz ds \leqslant C_2,$$

for some $C_2 > 0$, where the change of variable $z = x/\sqrt{s}$ was used.

Accordingly, making use also of Theorem 5.3,

$$\begin{aligned}
\sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \big|H **h(x,t)\big| &= \sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \big|H ** \Delta g(x,t)\big| \\
&\leqslant \sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \int_0^t \left| \int_{\mathbb{R}^n} H(y,s)\Delta g(x-y,t-s)dx \right| ds \\
&\leqslant \sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \int_0^t \left| \int_{\mathbb{R}^n} \nabla H(y,s)\cdot\nabla g(x-y,t-s)dx \right| ds \\
&\leqslant \sum_{k=1}^n \int_0^1 \big\|\partial_k H(\cdot,s)\big\|_{L^1_x} \|\partial_k g\|_{L^\infty_x} ds \\
&\leqslant \sum_{k=1}^n C_2 \big\|f'\big\|_{L^\infty([-1,1])} \sup_{\substack{x\in\mathbb{R}^n\\ t>0}} |\nabla u| \\
&\leqslant C_3,
\end{aligned}$$

for a suitable $C_3 > 0$, and so, recalling (29),

$$\begin{aligned}
\sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \big|G **h(x,t)\big| &\leqslant C_3 + \sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \big|N **h(x,t)\big| \\
&\leqslant C_3 + \sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \int_0^t \left| \int_{\mathbb{R}^n} N(y,s)\Delta g(x-y,t-s)dx \right| ds \\
&\leqslant C_3 + \sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \int_0^t \left| \int_{\mathbb{R}^n} \Delta N(y,s) g(x-y,t-s)dx \right| ds \\
&\leqslant C_3 + \int_0^1 \big\|\Delta N(\cdot,s)\big\|_{L^1_x} \|g\|_{L^\infty(\mathbb{R}^n\times(0,+\infty))} ds \\
&\leqslant C_3 + C_4\|f\|_{L^\infty([-1,1])} \\
&\leqslant C_5, \qquad (38)
\end{aligned}$$

for suitable $C_4, C_5 > 0$.

From (37) and (38), we conclude that

$$\sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \big|u_2(x,t)\big| \leqslant \sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \big|G**h(x,t)\big| + \sup_{\substack{x\in\mathbb{R}^n\\ t\in(0,1]}} \big|G*u_0(x,t)\big| \leqslant C_6, \tag{39}$$

for a suitable $C_6 > 0$, and our goal is now to bound u_2 when $t \geqslant 1$.

For this, we use again (38) to obtain the following estimate:

$$\begin{aligned}
\sup_{\substack{x\in\mathbb{R}^n\\ t\geqslant 1}} \big|G**h(x,t)\big| &= \sup_{\substack{x\in\mathbb{R}^n\\ t\geqslant 1}} \left| \int_0^t \left(\int_{\mathbb{R}^n} G(y,s)h(x-y,t-s)dy \right) ds \right| \\
&\leqslant \sup_{\substack{x\in\mathbb{R}^n\\ t\geqslant 1}} \left| \int_0^1 \left(\int_{\mathbb{R}^n} G(y,s)h(x-y,t-s)dy \right) ds \right| \\
&\quad + \left| \int_1^t \left(\int_{\mathbb{R}^n} G(y,s)h(x-y,t-s)dy \right) ds \right| \\
&\leqslant C_5 + \sup_{\substack{x\in\mathbb{R}^n\\ t\geqslant 1}} \left| \int_1^t \left(\int_{\mathbb{R}^n} \Delta G(y,s)g(x-y,t-s)dy \right) ds \right|.
\end{aligned}$$

As a consequence, recalling (36),

$$\begin{aligned}
\sup_{\substack{x\in\mathbb{R}^n\\ t\geqslant 1}} \big|G**h(x,t)\big| &\leqslant C_5 + \|f\|_{L^\infty([-1,1])} \sup_{t\geqslant 1} \int_1^t \big\|\Delta G(\cdot,s)\big\|_{L^1_x} ds \\
&\leqslant C_5 + \|f\|_{L^\infty([-1,1])} \int_1^{+\infty} \frac{C_0}{s^{1+a}} ds \\
&\leqslant C_7,
\end{aligned} \tag{40}$$

for a suitably large C_7.

This and (37) imply that

$$\sup_{\substack{x\in\mathbb{R}^n\\ t\geqslant 1}} \big|u_2(x,t)\big| \leqslant \sup_{\substack{x\in\mathbb{R}^n\\ t\geqslant 1}} \big|G**h(x,t)\big| + \sup_{\substack{x\in\mathbb{R}^n\\ t\geqslant 1}} \big|G*u_0(x,t)\big| \leqslant C_8,$$

for some $C_8 > 0$, and so, in the light of (39),

$$\sup_{\substack{x\in\mathbb{R}^n\\ t>0}} \big|u_2(x,t)\big| \leqslant C_6 + C_8. \tag{41}$$

Having completed the uniform bound on u_2, our goal is now to bound u_1. Though this will be obtained by following almost verbatim the argument on pages 141–143

of [1], we provide full details for the convenience of the reader. Exploiting (41), we can define

$$m := 1 + \sup_{x\in\mathbb{R}^n} \big|u_0(x)\big| + \sup_{\substack{x\in\mathbb{R}^n\\ t>0}} \big|u_2(x,t)\big|.$$

We claim that

$$\sup_{\substack{x\in\mathbb{R}^n\\ t>0}} \big|u_1(x,t)\big| \leqslant 2m+1. \tag{42}$$

The proof of (42) is by contradiction, hence we assume, say, that

$$\sup_{\substack{x\in\mathbb{R}^n\\ t>0}} u_1(x,t) > 2m+1.$$

Therefore, there exists $T>0$ for which

$$S := \big\{(x,t)\in\mathbb{R}^n\times[0,T) \text{ s.t. } u_1(x,t) > 2m+1\big\}$$

is not empty.

Notice that

$$\inf_S |u| \geqslant \inf_S |u_1| - \sup_S |u_2| \geqslant (2m+1) - m = m+1 > 1.$$

So, if $(x,t)\in S$, we have that $h(x,t) = \Delta f(u(x,t)) = 0$, thanks to (6), and therefore, by (35),

$$u_1 = H * u_0 \quad \text{in } S. \tag{43}$$

That is, since H is the heat kernel,

$$\partial_t u_1 = \Delta u_1 \quad \text{in } S. \tag{44}$$

Moreover, since $u\in\mathscr{W}$, we have that the map $t\mapsto u(0,t)$ belongs to $C^0([0,+\infty)) \subseteq L^\infty([0,T])$, hence

$$\sup_{t\in[0,T]} \big|u(0,t)\big| \leqslant \widetilde{C}(T),$$

for a suitable $\widetilde{C}(T)$, and therefore, exploiting (41) and Theorem 5.3,

$$\begin{aligned}\sup_{\substack{x\in\mathbb{R}^n\\ t\in[0,T]}} \big|u_1(x,t)\big| &\leqslant \sup_{\substack{x\in\mathbb{R}^n\\ t\in[0,T]}} \big|u_2(x,t)\big| + \sup_{\substack{x\in\mathbb{R}^n\\ t\in[0,T]}} \big|u(x,t)\big|\\ &\leqslant C_6 + C_8 + \sup_{t\in[0,T]} \Big(\big|u(0,t)\big| + C_9|x|\Big)\\ &\leqslant C(T)\big(1+|x|\big),\end{aligned} \tag{45}$$

for a suitable $C(T)>0$.

For any $\delta \in (0,1)$, we define

$$u_\delta(x,t) := u_1(x,t) - \delta\left(\frac{|x|^2}{2n} + t\right)$$

and

$$S_\delta := \big\{(x,t) \in \mathbb{R}^n \times [0,T) \text{ s.t. } u_\delta(x,t) > 2m+1\big\}.$$

It is easily seen that

$$S_\delta \subseteq S, \tag{46}$$

and so, from (44),

$$\partial_t u_\delta = \Delta u_\delta \quad \text{in } S_\delta. \tag{47}$$

Notice also that

$$u_\delta = 2m+1 \quad \text{on } \partial\big(\overline{S_\delta} \cap \{t = t_0\}\big), \text{ for any } t_0 \leqslant T. \tag{48}$$

Furthermore,

$$S_\delta \subseteq B_{(2+4nC(T))/\delta} \times [0,T]. \tag{49}$$

To prove (49), just take $(x,t) \in S_\delta$ with $|x| \geqslant 4nC(T)/\delta$, and use (45) to deduce that

$$\begin{aligned} 0 \leqslant 2m+1 \leqslant u_\delta(x,t) &\leqslant \big|u_1(x,t)\big| - \delta\left(\frac{|x|^2}{2n} + t\right) \\ &\leqslant C(T)\big(1+|x|\big) - \frac{\delta|x|^2}{2n} \leqslant C(T) - \frac{C(T)|x|}{2}, \end{aligned}$$

hence $|x| \leqslant 2$, which proves (49).

In fact, (49) can be improved as follows:

$$S_\delta \subseteq B_{(2+4nC(T))/\delta} \times [\mu, T], \tag{50}$$

for a suitable $\mu > 0$.

To prove (50), we argue by contradiction, supposing that there exists a sequence $(x_j, t_j) \in S_\delta$, with $t_j \to 0^+$ as $j \to +\infty$. By (49), we may suppose that $x_j \to x_\infty$ as $j \to +\infty$. Since, recalling (43) and (46), we have that $u_1(x_j,t_j) = H * u_0(x_j,t_j)$, we obtain that

$$\begin{aligned} 2m+1 &\leqslant \lim_{j\to+\infty} u_\delta(x_j,t_j) \leqslant \lim_{j\to+\infty} u_1(x_j,t_j) \\ &= H * u_0(x_\infty, 0) = u_0(x_\infty) \leqslant m. \end{aligned}$$

This contradiction proves (50).

Now, we show that

$$S_\delta \text{ is empty}, \quad \text{for any } \delta \in (0,1). \tag{51}$$

Suppose the contrary, hence S_δ is nonempty for some fixed $\delta \in (0,1)$. For any $j \in \mathbb{N}$, $j \geqslant 10$, let

$$u_j(x,t) := u_\delta(x,t) - \frac{t}{j}.$$

By (50), we can take $(x_j, t_j) \in \overline{S_\delta}$ such that

$$u_j(x_j,t_j) = \max_{\overline{S_\delta}} u_j. \tag{52}$$

Up to subsequence, we may and do suppose that

$$(x_j,t_j) \text{ lies in the interior of } \overline{S_\delta} \cap \{t = t_j\}, \tag{53}$$

otherwise (48) would give that

$$2m+1 = \lim_{j\to+\infty} u_j(x_j,t_j) \geqslant \lim_{j\to+\infty} u_j(x,t) = u_\delta(x,t) \quad \text{for any } (x,t) \in \overline{S_\delta},$$

which would say that S_δ is void.

Analogously, up to subsequence, we may and do suppose that for any j there exists $\tau_j > 0$ such that

$$(x_j, t_j - s) \in \overline{S_\delta} \quad \text{for any } s \in [0,\tau_j], \tag{54}$$

otherwise, for fixed j, we would have that $(x_j, t_j - \eta_\ell) \notin \overline{S_\delta}$ for a suitable infinitesimal sequence η_ℓ, and so, recalling (50), $u_\delta(x_j, t_j - \eta_\ell) \leqslant 2m+1$, hence

$$\begin{aligned} 2m+1 &\geqslant \lim_{j\to+\infty} \lim_{\ell\to+\infty} u_\delta(x_j, t_j - \eta_\ell) \\ &= \lim_{j\to+\infty} u_\delta(x_j,t_j) = \lim_{j\to+\infty} u_j(x_j,t_j) + \frac{t_j}{j} \\ &\geqslant \lim_{j\to+\infty} u_j(x,t) = u_\delta(x,t) \quad \text{for any } (x,t) \in \overline{S_\delta}, \end{aligned}$$

which, once more, would say that S_δ is void.

As a consequence of (52) and (53), we have that

$$\Delta u_j(x_j,t_j) \leqslant 0,$$

while the use of (54) gives that

$$\partial_t u_j(x_j,t_j) \geqslant 0.$$

Therefore,

$$0 \leqslant \partial_t u_j(x_j, t_j) - \Delta u_j(x_j, t_j) = \partial_t u_\delta(x_j, t_j) - \Delta u_\delta(x_j, t_j) - \frac{1}{j}.$$

This is in contradiction with (47), and so (51) is proved.

Since

$$S = \bigcup_{\delta \in (0,1)} S_\delta,$$

we obtain that S has to be empty as well. This contradiction proves (42).

Then, (41) and (42) imply (9) and the proof of Theorem 1.2 is thus completed.

Acknowledgements The first author is supported by NSF grants. The second author is supported by ICES *Oden Fellowship*, FIRB *A&B Analysis and Beyond* and ERC ε *Elliptic PDE's and Symmetry of Interfaces and Layers for Odd Nonlinearities*. Part of this work was performed while the second author was visiting UT, the very warm hospitality of which has been greatly appreciated.

References

1. Caffarelli, L.A., Muler, N.E.: An L^∞ bound for solutions of the Cahn-Hilliard equation. Arch. Ration. Mech. Anal. **133**(2), 129–144 (1995)
2. Frigeri, S., Grasselli, M.: Global and trajectory attractors for a nonlocal Cahn-Hilliard-Navier-Stokes system (2012). Preprint. http://arxiv.org/abs/1107.5933
3. Gal, C.G., Grasselli, M.: Longtime behavior of nonlocal Cahn-Hilliard equations (2012). Preprint. http://arxiv.org/abs/1207.4018
4. Giacomin, G., Lebowitz, J.L.: Phase segregation dynamics in particle systems with long range interactions. I. Macroscopic limits. J. Stat. Phys. **87**(1–2), 37–61 (1997)
5. Liu, G.: Existence of solution to initial-boundary value problems of the Cahn-Hilliard equation with nonlocal terms. J. Math. Res. **1**(2), 179 (2009). http://ccsenet.org/journal/index.php/jmr/article/download/3787/3398
6. Stein, E.M.: Singular Integrals and Differentiability Properties of Functions. Princeton Mathematical Series, vol. 30, Princeton University Press, Princeton (1970), xiv+290 pp.
7. Valdinoci, E.: From the long jump random walk to the fractional Laplacian. Bol. Soc. Esp. Mat. Apl. **49**, 33–44 (2009)

On the Numerical Analysis of Adaptive Spectral/*hp* Methods for Elliptic Problems

Claudio Canuto and Marco Verani

The transcendental is not infinite and unattainable tasks,
but the neighbor who is within reach in any given situation.
(D. Bonhoeffer)

Abstract We provide an overview of the state of the art of adaptive strategies for high-order *hp* discretizations of partial differential equations; at the same time, we draw attention on some recent results of ours concerning the convergence and complexity analysis of adaptive algorithm of spectral and spectral-element type. Complexity is studied under the assumption that the solution belongs to a sparsity class of exponential type, which means that its best N-term approximation error in the chosen piecewise polynomial basis decays at an exponential rate with respect to N.

1 Introduction

The present authors are some generations apart, yet both of them have been deeply influenced by the gigantic human and professional figure of Enrico Magenes; this paper is a modest yet heartfelt tribute to his memory.

On the scientific ground, professor Magenes not only gave outstanding contributions to the mathematical theory of partial differential equations, but he was a pioneer in fostering the development of a Computational Mathematics at the same time soundly tied to functional analysis and strongly immersed in real-life applications. The "Laboratorio di Analisi Numerica", that he founded in Pavia in the Seventies and directed for several decades, had a paramount impact on the Italian

C. Canuto (✉)
Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy
e-mail: claudio.canuto@polito.it

M. Verani
MOX-Dipartimento di Matematica, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milan, Italy
e-mail: marco.verani@polimi.it

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*, Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_11,

Applied Mathematics community, and soon become a recognized reference for the whole international community.

Enrico Magenes was able to create, and maintain over the years, a scientific environment extremely open and favorable to the development and exchange of new ideas. Many young researchers have been attracted by that atmosphere; some of them, such as the senior author (CC), had the chance to become members of that institution, and bring everlasting gratitude for the opportunity they had of being exposed day by day to the charismatic personality of the founder; some others, such as the junior author (MV), have remained fascinated and deeply influenced by his human and scientific legacy.

Enrico Magenes was always very careful in granting freedom of research to the members of his group, the only conditions to respect being the quality and the interest of the undertaken investigations. Good projects pushed forward by younger collaborators had the chance of being supported as more mature lines of research. The onset of interest for spectral methods, which are the object of the present contribution, at the "Laboratorio" in the early Eighties is precisely an example of this favorable environment. Influenced by a stimulating visit at the "Laboratoire d'Analyse Numérique" (created in Paris by Magenes' coworker and friend Jacques-Louis Lions), Alfio Quarteroni initiated a fruitful and long-lasting collaboration with the present senior author on the numerical analysis of spectral and high-order methods for boundary-value problems; while in the early stage the scientific guidance was provided by Franco Brezzi, the full and continuous strategic and logistic support granted by professor Magenes was certainly a key ingredient for the success of that research.

Since then, high-order methods such as spectral(-element) methods or the hp-version of finite element methods (the two categories being often hardly distinguishable from each other) have reached their maturity, both in the full understanding of their theoretical properties and in the penetration into the scientific computing practices, in various applicative environments (see, e.g., [12, 13, 50]). Yet, some relevant aspects of these methods are still far from being in a satisfactory shape, and deserve further investigations. An example is given by the so-called hp-adaptivity; indeed, even for steady problems, a full and rigorous understanding of the selection strategies between h-refinement and p-enrichment, and their influence on the complexity and optimality of the related algorithms, is still lacking.

The purpose of the present paper is to provide a soft overview of the state of the art of adaptive strategies for high-order discretizations of partial differential equations; at the same time, we aim at drawing attention on some recent results of ours concerning the convergence and complexity analysis of adaptive algorithm of spectral and spectral-element type.

We begin by recalling various approximation results which show that a proper choice of the mesh and the polynomial degree distribution over the mesh guarantees an exponential decay of the error even if the solution of the equation exhibits singularities inside the domain or at the boundary, provided their position is known. The free parameter is the cardinality of the set of active degrees of freedom. An elementary derivation is detailed, in the case of algebraic singularity or piecewise analyticity. Obviously, such results on optimal "a-priori adaptivity" constitute a benchmark

for the "a-posteriori" adaptive strategies, which need to detect the singularities and properly allocate the degrees of freedom around them. Thus, we are led to review the main error estimators proposed in the literature for hp methods, and the various adaptation strategies which exploit their information. While the number of different strategies is fairly large, with different mathematical sources and different practical performances, very few algorithms can be rigorously proven to be convergent, with a precise estimate of their rate of convergence. The situation becomes even worse if complexity or cost issues are to be taken into account.

In this respect, we devote the second part of the paper to illustrate some recent results we have obtained in collaboration with Ricardo H. Nochetto; we consider adaptive spectral methods of Legendre type, and actually we extend them to cover the case of spectral-element discretizations (or p-type finite elements). A representative algorithm (out of several possible variants) is described, and its convergence properties are discussed. Furthermore, we investigate its complexity, by comparing the output of the algorithm with the best possible approximation of the exact solution in the chosen piecewise polynomial basis, for the same accuracy—this point of view being related to the so-called "best N-term approximation" of a function. The novelty of the analysis, compared to the available results in the literature, is that optimality is discussed with respect to an assumed exponential (or sub-exponential) decay of the best N-approximation error; this assumption appears indeed to be coherent with the use of spectral-type methods in the discretization of the boundary-value problem.

Notation Throughout the paper, by $A \lesssim B$ we mean that the quantity A can be bounded by a multiple of B, the multiplicator being independent of those parameters A and B may depend on. Likewise $A \simeq B$ means $A \lesssim B$ and $B \lesssim A$, whereas $A \sim B$ means $A = B + o(B)$, with $o(B)$ negligible with respect to B.

2 From Approximation Theory to a-Priori Adaptive hp Methods

We begin by recalling some classical results concerning the approximation of a univariate function having Sobolev or analytical regularity, by means of algebraic polynomials. These estimates will be useful in the subsequent analysis of piecewise polynomial approximation.

Let I denote the reference interval $(-1, 1)$. If v belongs to the Sobolev space $H^m(I)$, $m \geq 0$, then

$$\inf_{w \in \mathbb{P}_p(I)} \|v - w\|_{L^2(I)} \leq C(v, m) p^{-m}, \tag{1}$$

where the positive constant $C(v, m)$ can be bounded by the norm of v in $H^m(I)$ (actually, the bound holds under weaker assumptions, see, e.g., [12, 50]). On the other hand, if v can be extended to an analytic function on the closed ellipse $E(-1, 1; \sigma)$

in the complex plane having foci at $z = \pm 1$ and semiaxes' sum $\sigma > 1$, then setting $\eta = \log \sigma$ one has

$$\inf_{w \in \mathbb{P}_p(I)} \|v - w\|_{L^2(I)} \leq C(v, \eta) p^{-1/2} \mathrm{e}^{-\eta p} \tag{2}$$

(see, e.g., [30, 50]). A different estimate involves the maximum modulus of v over the ellipse $E(-1, 1; \sigma)$ and reads as follows:

$$\inf_{w \in \mathbb{P}_p(I)} \|v - w\|_{L^2(I)} \leq C \frac{1}{\sinh \eta} \mathrm{e}^{-\eta p} \max_{z \in E(-1,1;\sigma)} \left| v(z) \right| \tag{3}$$

(see, e.g., [12]).

The previous results can be easily combined to provide bounds of the approximation error for piecewise smooth functions on a finite partition of the domain by piecewise polynomials; in such cases, the estimates involve not only the polynomial degrees but also the sizes of the subdomains. Less trivial is the problem of expressing the error in terms of the total number, say N, of employed degrees of freedom, and, even more, of optimizing the allocation of degrees of freedom for a given target accuracy. This is precisely the crucial problem of hp adaptivity.

The earliest attempts to study the adaptive approximation of a univariate function, having a finite number of singularities and otherwise smooth, by means of piecewise polynomials of variable degree dates back to the late Seventies, with the pioneering works [20] and [24] (see also [21] and the references therein).

In [20], the best N-term approximation of a univariate function in the maximum norm by piecewise polynomials of variable degree is studied, and in particular it is proven that for certain classes of analytic functions the best N-term approximation is achieved by a single polynomial over the entire domain. On the other hand, [24] deals with functions with singularities of the type x^α near the origin, and proves that a proper combination of graded mesh and linear increase of polynomial degrees (see below) yields exponential decay of the best N-term approximation error, with exponent proportional to $\sqrt{N}$.

This result influenced Gui and Babuška [30] in their study of the convergence rate of the hp approximation to a model elliptic problem in 1D. As usual, a Céa Lemma argument reduces the problem to estimate the best N-term approximation error in the energy norm. Let us give some detail.

Let $\Omega = (0, 1)$. Suppose that the solution u of the underlying elliptic PDE is real analytic in $(0, 1]$ and behaves like $u(x) = x^\alpha$ for some $\alpha > 1/2$. Consider a partition of $\overline{\Omega}$ into contiguous intervals K_j $(0 \leq j \leq J)$ and a corresponding distribution of polynomial degrees $p_j \geq 1$. Define the subspace

$$V_\delta = \left\{ v \in H^1(\Omega) : v_{|K_j} \in \mathbb{P}_{p_j}(K_j), 0 \leq j \leq J \right\}. \tag{4}$$

Let us assume that the mesh is geometrically graded towards 0, i.e., there exists $0 < \rho < 1$ such that

$$K_j = \left[\rho^{j+1}, \rho^j \right] \quad \text{for } 0 \leq j < J, \qquad K_J = \left[0, \rho^J \right], \tag{5}$$

whereas the polynomial degrees grow linearly away from 0, i.e.,

$$p_j = \max\bigl(1, \bigl[\lambda(J-j)\bigr]\bigr) \tag{6}$$

for some $\lambda > 0$. Under these assumptions, there exists λ such that one has for $N = \dim V_\delta$

$$\inf_{v_\delta \in V_\delta} \|u - v_\delta\|_{H^1(\Omega)} \leq C\mathrm{e}^{-b\sqrt{N}}, \tag{7}$$

where the constants $C > 0$ and $b > 0$ are independent of N. In particular, the choice $\rho^* = (\sqrt{2}-1)^2$ for the grading factor is optimal for any α.

The result can be extended to more general functions; in particular, to those belonging to the class $B^\ell_\beta(\Omega)$ for $\ell = 1, 2$ and some $\beta \in (0, 1)$; these are the functions $u \in H^{\ell-1}(\Omega)$ such that, setting $\Phi_\gamma(x) = |x|^\gamma$, the functions $\Phi_{\beta+k-\ell} D^k u$ belong to $L^2(\Omega)$ for any $k \geq \ell$, and there exist constants $C > 0$ and $d \geq 1$ for which

$$\bigl\|\Phi_{\beta+k-\ell} D^k u\bigr\|_{L^2(\Omega)} \leq C d^{k-l}(k-l)!. \tag{8}$$

Based on the previously described a-priori analysis, Gui and Babuška [31] proposed what is probably the first *hp* adaptive algorithm (see also Sect. 3.2 below). Given an elemental error estimator satisfying suitable assumptions, the elements of the partition on which the error estimator is larger than a fixed fraction of the largest estimator are marked for refinement/enrichment; by inspecting the ratio between two error estimators on the marked interval with two consecutive polynomial degrees, it is decided whether to divide the interval into two parts carrying the same polynomial degree as before, or to keep the interval unchanged but increase the polynomial degree by one. The algorithm is proven to be convergent, with a predicted rate. However, the assumptions on the admissible error estimators appear to be overly restrictive, essentially they are taylored on the x^α-type singularity, for which indeed the algorithm produces a nearly optimal discretization.

In 2D, the counterpart of the previous a-priori analysis is as follows. Consider a bounded polygon Ω, having the origin 0 as a vertex. Let u be the solution of an elliptic problem in Ω, which is real analytic in $\overline{\Omega} \setminus \{0\}$ and behaves like $|x|^\alpha$ as $|x| \to 0$. Consider a conforming and regular partition of $\overline{\Omega}$ by J layers of elements around the origin; all the elements in the j-th layer have diameter of the order of ρ^j for some fixed $0 < \rho < 1$. Assume that all elements in the j-th layer carry polynomial degrees of the order of p_j, with $p_j = \max(1, [\lambda(J-j)])$ for some $\lambda > 0$. Let the subspace $V_\delta \subset H^1(\Omega)$ be defined in the standard way, and let again $N = \dim V_\delta$. Then, the following error bound is proven in [32, 33]:

$$\inf_{v_\delta \in V_\delta} \|u - v_\delta\|_{H^1(\Omega)} \leq C\mathrm{e}^{-b\sqrt[3]{N}}, \tag{9}$$

with $C > 0$ and $b > 0$ independent of N. The result extends to solutions in the class $B^2_\beta(\Omega)$, locally defined in a neighborhood of each vertex in a manner similar as above. This is relevant, since the solution of elliptic problems in Ω with data having

suitable piecewise-analytic regularity can be shown to belong to such a class, see [5].

The situation in 3D is more complex, since in polyhedra singularities occur not only at vertices, but also along edges. Thus, an adapted mesh has a different structure in different regions of the domain, to accommodate for the local structure of the solution: it is quasi-uniform away from the boundary, it is isotropically graded towards a vertex, it is anisotropically graded towards the central part of an edge (being quasi-uniform in the tangential direction), and finally it has a transitional nature near the portion of an edge that gets close to a vertex. Then, a proper distribution of the polynomial degrees over such a mesh guarantees the following behavior of the best-approximation error vs the dimension N of the corresponding subspace V_δ, for the solution u of an elliptic problem in a polyhedron, with piecewise analytic data:

$$\inf_{v_\delta \in V_\delta} \|u - v_\delta\|_{H^1(\Omega)} \leq C \mathrm{e}^{-b\sqrt[5]{N}}, \tag{10}$$

again with $C > 0$ and $b > 0$ independent of N. The result (that should be compared to (7) in 1D and (9) in 2D) was first asserted by Guo and Babuška [6]; for the proof, we refer to [48, 49], where both Continuous- and Discontinuous-Galerkin hp discretizations are considered. The analysis relies on very accurate estimates of suitable weighted norms of the solution, of the type (8); we refer to [34] and to the more recent and comprehensive result [19].

2.1 An Elementary Analysis of hp Approximations over Dyadic Partitions

Herafter, we use elementary arguments based on the repeated application of the error estimate (3) in order to establish the exponential convergence of suitable hp approximations over dyadic partitions to singular functions or piecewise-analytic functions in one space dimension.

To this end, given real numbers $r < s$ and $\sigma > h := s - r$, let $E(r, s; \sigma)$ denote the closed ellipse in the complex plane having foci at $z = r, s$ and semiaxes' sum σ; let us set $c = (r + s)/2$ and $f = h/2$. Let v be a function defined on the interval (r, s) of the real line, that can be extended to an analytic function on the closed ellipse $E(r, s; \sigma)$. Then, if we apply the change of variable

$$\hat{x} = \frac{x - c}{f}, \qquad \hat{y} = \frac{y}{f},$$

it is easily seen that the function $\hat{v}$ such that $\hat{v}(\hat{x}) = v(x) = v(c + f\hat{x})$ is defined on the reference interval $I = (-1, 1)$ and can be extended to an analytic function on the closed ellipse $E(-1, 1; \hat{\sigma})$ with $\hat{\sigma} = \sigma/f$. Thus, we can apply the bound (3)

with $\eta = \log\hat{\sigma} = \log\sigma - \log f$, to obtain

$$\begin{aligned}\inf_{w\in\mathbb{P}_p((r,s))} \|v-w\|_{L^2((r,s))} &= h^{1/2} \inf_{\hat{w}\in\mathbb{P}_p(I)} \|\hat{v}-\hat{w}\|_{L^2(I)} \\ &\le Ch^{1/2}\frac{1}{\sinh\eta}\mathrm{e}^{-\eta p} \max_{z\in E(-1,1;\hat{\sigma})} \big|\hat{v}(z)\big| \\ &= Ch^{1/2}\frac{1}{\sinh\eta}\mathrm{e}^{-\eta p} \max_{z\in E(r,s;\sigma)} \big|v(z)\big|. \end{aligned} \tag{11}$$

The Case of an Algebraic Singularity

Let $\Omega = (0,1)$. Suppose again that the solution u is real analytic in $(0,1]$ and behaves like $u(x) = x^\alpha$ for some $\alpha > 1/2$. Let us consider the subspace of $L^2(\Omega)$ defined similarly to (4), i.e.,

$$V_\delta = \big\{v \in L^2(\Omega) : v_{|K_j} \in \mathbb{P}_{p_j}(K_j),\ 0 \le j \le J\big\}, \tag{12}$$

where the K_j are defined in (5) with $\rho = \frac{1}{2}$ and the p_j are to be determined in the sequel. If v_δ denotes any function in V_δ, we split the approximation error as

$$\|u - v_\delta\|^2_{L^2(\Omega)} = \sum_{j=0}^{J} \|u_j - v_j\|^2_{L^2(K_j)}$$

with $u_j = u_{|K_j}$ and $v_j = v_{\delta|K_j}$. For $j = J$, we take as v_J the linear interpolant of u_J. This yields, with $h_J = 2^{-J}$,

$$\|u_J - v_J\|^2_{L^2(K_J)} \simeq \int_0^{h_J} \big(s^\alpha - h_J^{\alpha-1}s\big)^2 ds \simeq h_J^{2\alpha+1} \simeq 2^{-(2\alpha+1)J}.$$

Consider now any interval K_j with $0 \le j < J$ and set $h_j = 2^{-(j+1)}$. We can think u_j as a real analytic function in K_j which can be extended to an analytic function in any closed ellipse $E(\rho^{j+1}\rho^j;\sigma)$ with $\sigma < \sigma_j = \frac{3}{2}h_j$. Hence, setting $\hat{\sigma}_j = \sigma_j/f_j = (\frac{3}{2}h_j)/(\frac{1}{2}h_j) = 3$ and $\eta_j = \log\hat{\sigma}_j = \log 3$, we can apply (11) in K_j and find $v_j \in \mathbb{P}_{p_j}(K_j)$ such that

$$\begin{aligned}\|u_j - v_j\|^2_{L^2(K_j)} &\le Ch_j\mathrm{e}^{-2\eta p_j} \max_{z\in E(\rho^{j+1}\rho^j;\sigma_j)} \big|u_j(z)\big|^2 \\ &\le C(u)h_j\mathrm{e}^{-2\eta p_j} = C(u)2^{-(j+1+\eta^* p_j)},\end{aligned}$$

with $\eta^* = 2(\log_2\mathrm{e})\,\eta$.

Let λ, μ be fixed constants ≥ 0 such that $\lambda + \mu = 1$. Then,

$$\|u - v_\delta\|^2_{L^2(\Omega)} \le C\sum_{j=0}^{J-1} 2^{-\lambda j}2^{-(\mu j+1+\eta^* p_j)} + 2^{-(2\alpha+1)J}.$$

Let us enforce that

$$\|u - v_\delta\|^2_{L^2(\Omega)} \leq C 2^{-2M}$$

for any given M. The bound on the error given above suggests to choose $J \sim 2M/(2\alpha + 1)$ as well as $\mu j + \eta^* p_j \sim 2M$, i.e., $p_j \sim (2M - \mu j)/\eta^*$. With such choices, it is readily seen that the total number N of activated degrees of freedom satisfies

$$N = \sum_{j=0}^{J} p_j \simeq M^2,$$

i.e., $M \simeq \sqrt{N}$. We conclude that the best approximation error satisfies

$$\inf_{v_\delta \in V_\delta} \|u - v_\delta\|_{L^2(\Omega)} \leq C \mathrm{e}^{-b\sqrt{N}} \tag{13}$$

for some $b > 0$, i.e., a bound of the same type as (7). Note that the definition of p_j given above is of the same type as (6).

The best approximation error in the $H^1(\Omega)$-norm can be estimated in a similar manner.

The Case of a Piecewise-Analytic Function

Assume now that u is a piecewise analytic function in $\overline{\Omega}$. It is not restrictive to assume the existence of just one singular point, say $x_s \in \Omega$. Thus, both $u_l := u_{|[0,x_s]}$ and $u_r := u_{|[x_s,1]}$ can be extended to analytic functions in a neighborhood of their intervals of definition in the complex plane.

With the aim of mimiking an adaptive algorithm which detects the position of the singularity by some error indicator, we consider the approximation procedure that generates a dyadic partition of $\overline{\Omega}$ by recursively halving the subinterval which contains the singular point x_s. Obviously, if x_s itself is a dyadic point, the procedure stops after a finite number of subdivisions, and we are just required to approximate by polynomials a finite number of analytic functions over a partitions of $\overline{\Omega}$; then, it is enough to apply (11) to each of them. If x_s is not a dyadic point, then at iteration $J \geq 0$ of the recursive algorithm we have a partition of the domain into $J + 1$ subintervals K_j, such that $h_j := |K_j| = 2^{-j}$ for $0 \leq j \leq J$, and such that K_J is the only interval containing x_s. Let us set $u_j := u_{|K_j}$.

If u_J is of class C^ℓ In the interval K_J, for some $\ell \geq -1$ ($\ell = -1$ meaning that u_J has a jump at x_s), then we can find a polynomial v_J of degree $\ell + 1$ such that

$$\|u_J - v_J\|^2_{L^2(K_J)} \simeq h_J^{2\ell+3} \simeq 2^{-(2\ell+3)J}.$$

On the other hand, in any interval K_j with $0 \leq j < J$, u_j is a real analytic function which can be extended to an analytic function in some closed ellipse $E(r_j, s_j; \sigma_j)$ where $K_j = [r_j, s_j]$ and $\sigma_j \simeq 1$ depending on the size of the ellipse of analyticity of

either u_l or u_r. In view of applying (11) in K_j we observe that $f_j = h_j/2 = 2^{-(j+1)}$, so that $\eta_j = \log(\sigma_j/f_j) \simeq a + bj$ and $\sinh \eta_j \sim \frac{1}{2}\mathrm{e}^{\eta_j} \simeq 2^j$. Hence,

$$\begin{aligned}\|u_j - v_j\|^2_{L^2(K_j)} &\le C h_j \mathrm{e}^{-2\eta_j p_j} \max_{z \in E(r_j, s_j; \sigma_j)} \left|u_j(z)\right|^2 \\ &\le C(u) 2^{-3j} \mathrm{e}^{-2(a+bj)p_j} = C(u) 2^{-(3j + (a^* + b^* j)p_j)}\end{aligned}$$

with $a^* = a \log_2 \mathrm{e}$ and $b^* = b \log_2 \mathrm{e}$. Summing-up, we obtain

$$\|u - v_\delta\|^2_{L^2(\Omega)} \le C \sum_{j=0}^{J-1} 2^{-j} 2^{-(2j + (a^* + b^* j)p_j)} + 2^{-(2\ell+3)J}.$$

Let us enforce that

$$\|u - v_\delta\|^2_{L^2(\Omega)} \le C 2^{-2M}$$

for any given M. The bound on the error given above suggests to choose $J \sim 2M/(2\ell + 3)$ as well as $2j + (a^* + b^* j)p_j \sim 2M$, i.e., $p_j \sim 2(M - j)/(a^* + b^* j)$. With such choices, it is readily seen that the total number N of activated degrees of freedom satisfies

$$N = \sum_{j=0}^{J} p_j \simeq \sum_{j=0}^{J-1} \frac{M - j}{a^* + b^* j} \simeq M \log J \simeq M \log M,$$

i.e., $M \simeq \phi(N)$, where $x = \phi(y)$ is the inverse function of $y = x \log x$ for $x \ge 1$. We conclude that the best approximation error satisfies

$$\inf_{v_\delta \in V_\delta} \|u - v_\delta\|_{L^2(\Omega)} \le C \mathrm{e}^{-b\phi(N)} \tag{14}$$

for some $b > 0$. The result indicates that the behavior of the best approximation error in the presence of piecewise analyticity is only marginally worse than the one in the case of full analyticity, see (7).

Again, the best approximation error in the $H^1(\Omega)$-norm can be estimated in a similar manner.

3 *hp* Adaptivity

Over the last few decades, adaptive algorithms have become a standard technique for solving partial differential equations via the finite element method. The general form of an adaptive algorithm can be stated as follows:

$$\cdots \to \texttt{SOLVE} \to \texttt{ESTIMATE} \to \texttt{MARK} \to \texttt{ENRICH} \to \texttt{SOLVE} \to \cdots.$$

Generally speaking, the algorithm starts computing the discrete solution (SOLVE) employing a low-dimensional approximation space. Thereafter, in order to improve the accuracy of the approximation, an error indicator is employed (ESTIMATE) to obtain information about the error distribution. Based on this error distribution, a set of elements are flagged (MARK) to be enriched and a suitable enrichment of the approximation space is chosen (ENRICH). A new approximation of higher accuracy is computed and a new adaptive iteration is performed in case the approximation is not sufficiently accurate. In the adaptive h-FEM, the enrichment of the finite element space is simply done by subdividing into smaller elements all those elements with a large error indicator. However, in the hp-FEM one has the option to split an element or to increase its approximation order. Thus, as already pointed out, a main difficulty in hp-adaptivity is to decide whether to increase the approximation order p or to split an element whose error is large. The importance of making the correct decisions is highlighted by the a priori results mentioned in Sect. 2, from which it is evident that for a large class of problems an exponential rate of convergence can be achieved if the mesh and the polynomial degree distribution are chosen suitably.

Although considerable progress has been made in the context of adaptive h-FEM on both the a posteriori error analysis and the theoretical and computational assessment of the convergence properties of the adaptive refinement strategies (see, e.g., [43] for a comprehensive introduction), in contrast the theory of adaptive hp-FEM is far less advanced. Below we provide a brief review of existing a-posteriori hp error estimates (Sect. 3.1) and hp adaptive methods (Sect. 3.2).

3.1 A-Posteriori hp Error Estimates

In the hp framework, similarly to the case of h-FEM, error indicators can be subdivided into the following categories:

- *Estimators based on the (approximate) solution of suitably defined local problems*. This includes [1–4, 44]. The estimators of [1–4] are based on solving local problems with Neumann type boundary conditions; a forerunner of this approach is [44]. Additionally, [44] discusses in detail other techniques known from h-FEM that can be extended to the hp-context such as solving local Dirichlet problems on patches, employing duality theory from convex optimization to derive upper and lower bounds of the local errors and employing various interpolation/postprocessing techniques to obtain more accurate approximations.

  At this point, we also mention the equilibrated residual estimators introduced in [10]. Although the method of [10] uses equilibrated fluxes, it differs from estimators via local Neumann problems as the estimators are obtained by the hypercircle method.
- *Residual based a-posteriori error estimators*. In the pioneering work [8], a posteriori error indicators of residual type have been considered. However, the two-dimensional analysis of [8] is restricted to meshes consisting of axiparallel rectangles. In [40] the results of [8] are extended to meshes containing quadrilaterals

and triangles and a family η_α, $\alpha \in [0, 1]$, of error indicators given by weighted residuals on the elements and on the edges is introduced. It is shown that η_0 is reliable. As in [8], the reason for considering a family of indicators is that simultaneous reliability and efficiency cannot be proved for any fixed $\alpha \in [0, 1]$ due to the poor p-dependence of polynomial inverse estimates. For a related residual based a posteriori error estimate in one dimension, see also [47].

- *Estimators based on more accurate approximate solutions of the global problem.* This approach is based on the following steps: (1) a reference (finer) solution is computed by performing a global hp-refinement, i.e., breaking each element isotropically and enriching the polynomial order of approximation by one; (2) an error indicator is built by computing (and localizing) a suitable projection-based interpolation error of the reference solution. Roughly speaking, the indicator is computed by projecting the reference solution onto a finite element space employing the original mesh, but with a local polynomial degree incremented by one, as well as on a sequence of finite element spaces corresponding to a local h-refinement of the element that results in the same increase in the number of degrees of freedom as the p-enrichment. This approach has been introduced in [23] and further developed in [38].

Finally, we refer to [35] for goal-oriented hp-type error estimators.

3.2 Adaptive hp Methods

Classical h adaptive finite element methods simply subdivide elements where the local error indicator is large, while keeping the polynomial degree fixed (at some low value). In general, this may not be the most efficient strategy in terms of error reduction per unit cost. For example, if the analytical solution to the underlying partial differential equation is smooth, or at least locally smooth, an enrichment of the polynomial degree (p-refinement) may be much more effective in reducing the local error per unit cost than a simple element subdivision (h-refinement). Generally speaking, a local p-refinement is expected to be more efficient on elements where the solution is smooth, while local h-refinement is preferable for regions where the solution is not smooth.

In the following we will briefly review existing hp adaptive strategies. In particular, we will highlight the mechanism driving the choice between h or p refinement.

1. *Optimization strategy based on reference solution.* In this strategy a reference solution is computed on a finer finite element space, which is obtained by uniformly refining all the elements and globally incrementing the polynomial degree by one. Then, on each element in the coarser finite element mesh, the projection-based interpolation error of the reference solution is computed (see in the previous subsection the description of the estimators based on more accurate approximate solutions of the global problem). The optimal refinement of each element is then chosen to be the one which leads to the smallest projection-based interpolation error; elements in the mesh are then refined based on those that will lead to

the greatest decrease in the projection error per degree of freedom. This strategy was first introduced by [22, 44, 46]; see also [23] for more recent work.

2. *Relative size of the error estimators*. In this strategy (originally introduced in [31]) it is assumed the existence of a local error indicator $\eta_K(u_{h,p}, h_K, p_K)$ which depends on the element K, the approximate solution $u_{h,p}$, the local mesh-size h_K and the local polynomial degree p_K. Then, the choice between h and p refinement is based on the ratio $r_K = \eta_K(u_{h,p_K}, h_K, p_K)/\eta_K(u_{h,p_K-1}, h_K, p_K - 1)$. In particular, if $c_k \leq \gamma$, $0 < \gamma < 1$ then p-enrichment should be performed as the error decreases when the polynomial degree is raised. On the other hand, if $c_K > \gamma$ then the element K is subdivided.
3. *Comparison of estimated and predicted error*. This strategy has been proposed in [40] (see also [27]) where the decision whether to subdivide an element or to increase its polynomial degree depends on the refinement history of the element. In particular, it is introduced a predicted local error indicator η_K^{pred} which can be viewed as a simple extrapolation of the error indicators computed during the previous refinement steps under the assumption that the solution is (locally) smooth. If the computed error indicator (which reflects the actual error) is larger than the predicted one, then an h-refinement is performed since the assumption of (local) smoothness, under which the computation of the predicted indicator is performed, is false. Conversely, if the indicated error is smaller than the predicted one, then p-refinement is performed.
4. *Analyticity check by estimating the decay rate of expansion coefficients*. In [21], the authors propose to determine whether the solution is locally smooth or non-smooth by calculating the decay rate of the Legendre expansion coefficients of the solution; this is performed by a least-squares best fit. More recently, a strategy has been developed in [37] for estimating the size of the Bernstein ellipse of the solution, thereby determining whether the solution is analytic. In the case when it is not analytic, a second strategy, based on the work developed in [36], seeks to directly compute the local Sobolev index of the solution.
5. *Local regularity estimation*. This strategy, first proposed in [4], relies on estimating in each element K the local Sobolev regularity index m_k, by using a local error indicator which is computed by solving a series of local problems with different polynomial degrees. The local Sobolev regularity is then employed to perform h or p refinement. In particular, if $p_K + 1 \leq m_K$, where p_K denotes the current local polynomial order, then p-refinement is performed in K, otherwise h-refinement is selected. This latter criterium relies on two ingredients: (i) the following hp a-priori error estimate on quasi-uniform mesh of size h and elements of uniform polynomial degree p:

$$\|u - u_{hp}\|_{H^1(\Omega)} \leq C h^{\mu} p^{-(m-1)} \|u\|_{H^m(\Omega)}$$

where $\mu = \min(p, m-1)$ and $u \in H^m(\Omega)$, and (ii) the idea that if the regularity of the solution is such that the rate of convergence of the h-type finite element method with elements of fixed degree p turns out to be sub-optimal, then an h refinement is needed. Recently, related ideas have been exploited in [26, 47]

where the residual, instead of the error, has been employed to choose between p enrichment or h refinement. More recently, an approach based on the Sobolev embedding of H^1 into an appropriate L^p space has been proposed for the one-dimension case in [54].

6. *"Texas Three step"*. This strategy was first introduced [45] and is based on a three-step scheme where only three solutions of the problem are needed. First, the initial mesh details as well as intermediate and final error tolerances are specified and the problem is solved. Then, the h-refinements take place in order to guarantee that the (intermediate) error (measured in some appropriate norm) is less than the intermediate tolerance. In the final third step, the mesh is kept fixed and the p-refinements are carried out to achieve the final error tolerance. For related work, we refer, e.g., to [53] and the references cited therein.

At last, we mention that a thorough comparison among various hp adaptive strategies has been recently accomplished in [41]. Algorithms have been tested on different kinds of representative solutions (analytic solution, corner singularity, peak, boundary layer, wavefront, and so on) and their performance has been evaluated according to different measures of efficiency, such as the number of activated degrees of freedom or the computational time. None of the considered strategies has emerged as the best one in all situations, although some strategies perform better than the others for specific kinds of solutions.

3.3 Convergence of Adaptive Spectral/hp Methods

The theory of h adaptive finite element (AFEM) schemes for elliptic problems is quite satisfactory: it started with the convergence results of [25] and [42]; the first optimality result was derived in [9] for $d = 2$ and extended by [51] to any d. The most comprehensive results for AFEM are contained in [16] for any d and L^2 data, and [18] for $d = 2$ and H^{-1} data; we refer to the survey [43]. In contrast, very little is known on convergence and optimality properties of adaptive hp methods. The first pioneering result goes back to [31] where an adaptive hp algorithm (see Sect. 3.2 for the description) is proven to be convergent, with a predicted rate. However, due to the assumptions on the admissible error estimators which appear to be overly restrictive, the results in [31] cannot be considered completely satisfactory. Only after two decades, in [26] a contraction result of the form

$$\|u - u_{n+1}\|_{H^1} \leq \mu \|u - u_n\|_{H^1}, \quad \mu < 1,$$

has been proven, where u_n and u_{n+1} are the coarse and the enriched discrete solutions built by an adaptive hp algorithm approximating a one dimensional elliptic problem. More recently, the convergence result of [26] has been extended to higher dimensions in [11]. In this respect, it is also worth mentioning the result in [47] where an estimation (from above and below) of the error between the actual discrete solution and its (h or p) enriched version is presented. However, to the best of

the authors' knowledge, there are no optimality results for hp adaptive algorithms: this is still a completely open issue.

4 Spectral Adaptive Algorithms with Optimality Properties

Inspired by the analysis performed in the wavelet framework by [17, 29, 52] and in the finite element framework by [9, 16, 25, 42, 43, 51], the present authors in collaboration with Ricardo H. Nochetto have recently initiated the study of the convergence and complexity properties of adaptive Fourier Galerkin methods in arbitrary d-dimension [14], and of adaptive Legendre-Galerkin methods in one-dimension [15]. Hereafter, we present a short account of the latter results, which incorporates their extension to the case of discretizations by spectral elements.

We consider the elliptic problem in $\Omega = (a, b)$

$$\begin{cases} Au = -D \cdot (\nu Du) + \sigma u = f & \text{in } \Omega, \\ u(a) = u(b) = 0, \end{cases} \tag{15}$$

where ν and σ are sufficiently smooth real coefficients satisfying $0 < \nu_* \le \nu(x) \le \nu^* < \infty$ and $0 < \sigma_* \le \sigma(x) \le \sigma^* < \infty$ in Ω; let us set

$$\alpha_* = \min(\nu_*, \sigma_*) \quad \text{and} \quad \alpha^* = \max\big(\nu^*, \sigma^*\big).$$

We formulate this problem variationally as

$$u \in V := H_0^1(\Omega): \quad a(u, v) = \langle f, v \rangle \quad \forall v \in H_0^1(\Omega), \tag{16}$$

where $a(u, v) = \int_\Omega \nu Du Dv + \int_\Omega \sigma uv$. We denote by $|||v||| = \sqrt{a(v, v)}$ the energy norm of any $v \in H_0^1(\Omega)$, which satisfies

$$\sqrt{\alpha_*}\|v\| \le |||v||| \le \sqrt{\alpha^*}\|v\|. \tag{17}$$

Our error estimators will be of residual type. Therefore, for any w belonging to some finite dimensional subspace V_Λ of $H_0^1(\Omega)$, we define the residual $r(w) = f - Aw \in H^{-1}(\Omega)$. Then, by the continuity and coercivity of the bilinear form a, one has

$$\frac{1}{\alpha^*}\big\|r(w)\big\| \le \|u - w\| \le \frac{1}{\alpha_*}\big\|r(w)\big\|, \tag{18}$$

or, equivalently,

$$\frac{1}{\sqrt{\alpha^*}}\big\|r(w)\big\| \le |||u - w||| \le \frac{1}{\sqrt{\alpha_*}}\big\|r(w)\big\|. \tag{19}$$

4.1 Bases and Norm Representations

Form now on, we assume that the coefficients and data of the problem are piecewise analytic on a finite partition $\mathcal{T} = \{K\}$ of $\overline{\Omega}$. Let us introduce the subspace of $H_0^1(\Omega)$ of the piecewise linear functions on $\mathcal{T}$, i.e.,

$$V_L(\mathcal{T}) = \left\{ v \in H_0^1(\Omega) \;\middle|\; v_{|K} \in \mathbb{P}_1(K)\ \forall K \in \mathcal{T} \right\};$$

then,

$$V := H_0^1(\Omega) = V_L(\mathcal{T}) \oplus \bigoplus_{K \in \mathcal{T}} H_0^1(K),$$

where, for convenience, we assume that functions in $H_0^1(K)$ are extended by 0 outside the interval K; indeed, for any $v \in V$, we have

$$v = v_L + \sum_{K \in \mathcal{T}} v_K,$$

where $v_L \in V_L(\mathcal{T})$ is the piecewise linear interpolant of v on $\mathcal{T}$ and $v_K = (v - v_L)_{|K} \in H_0^1(K)$. Since

$$(v_L, v_K)_{H_0^1(\Omega)} = \int_K v_L'(v - v_L)' = v_{L|K}' \int_K (v - v_L)' = 0 \quad \text{and}$$
$$(v_K, v_{K'})_{H_0^1(\Omega)} = 0 \quad \text{if } K \neq K',$$

we have

$$\|v\|^2_{H_0^1(\Omega)} = \|v_L\|^2_{H_0^1(\Omega)} + \sum_{K \in \mathcal{T}} \|v_K\|^2_{H_0^1(K)}.$$

Given any $F \in V' = H^{-1}(\Omega)$, let $F_L \in V_L(\mathcal{T})'$ denote the restriction of F to $V_L(\mathcal{T})$; similarly, for each $K \in \mathcal{T}$, let $F_K \in H^{-1}(K)$ denote the restriction of F to $H_0^1(K)$. Then,

$$\langle F, v \rangle = \langle F_L, v_L \rangle + \sum_{K \in \mathcal{T}} \langle F_K, v_K \rangle \quad \forall v \in V,$$

which easily implies

$$\|F\|^2_{H^{-1}(\Omega)} = \|F_L\|^2_{V_L(\mathcal{T})'} + \sum_{K \in \mathcal{T}} \|F_K\|^2_{H^{-1}(K)}. \tag{20}$$

Let us now introduce the Lagrangian basis functions ψ_q in $V_L(\mathcal{T})$ associated with the internal nodes of the partition, say x_q for $1 \le q \le Q$, so that

$$V_L(\mathcal{T}) = \operatorname{span}\{\psi_q \mid 1 \le q \le Q\}.$$

On the other hand, on the reference element $\hat{I} = (-1, 1)$, we consider the Babuška-Shen basis, made of polynomials of strictly increasing degree,

$$\hat{\phi}_k(\hat{x}) = \sqrt{\frac{2k-1}{2}} \int_{\hat{x}}^{1} L_{k-1}(s)\, ds = \frac{1}{\sqrt{4k-2}} \big(L_{k-2}(\hat{x}) - L_k(\hat{x}) \big) \quad k \geq 2, \quad (21)$$

where $L_k(\hat{x})$, $k \geq 0$, stands for the k-th Legendre polynomial, which satisfies $\deg L_k = k$, $L_k(1) = 1$ and

$$\int_{-1}^{1} L_k(\hat{x}) L_m(\hat{x}) d\hat{x} = \frac{2}{2k+1} \delta_{km}, \quad m \geq 0.$$

It is easily seen that the basis functions satisfy

$$(\hat{\phi}_k, \hat{\phi}_m)_{H_0^1(\hat{I})} = \int_{-1}^{1} \hat{\phi}_k'(\hat{x}) \hat{\phi}_m'(\hat{x}) d\hat{x} = \delta_{km}, \quad k, m \geq 2,$$

i.e., they form an orthonormal system with respect to the $H_0^1(\hat{I})$-inner product. Going back to our partition, for any element $K \in \mathcal{T}$ of size h_K, we can map the reference element to the element K via an affine transformation $x = h_K \hat{x} + c$, yielding the functions

$$\phi_{K,k}(x) = \sqrt{\frac{h_K}{2}} \hat{\phi}_k(\hat{x}), \quad k \geq 2, \quad (22)$$

which form an orthonormal system with respect to the $H_0^1(K)$-inner product.

At this point, we are ready to give a representation of the norms in $H_0^1(\Omega)$ and $H^{-1}(\Omega)$. Precisely, any $v \in H_0^1(\Omega)$ is expanded as

$$v = \sum_{q=1}^{Q} \hat{v}_q \psi_q + \sum_{K \in \mathcal{T}} \sum_{k=2}^{\infty} \hat{v}_{K,k} \phi_{K,k} =: \sum_{\lambda \in \mathbb{L}} \hat{v}_\lambda \varphi_\lambda,$$

so that

$$\|v\|^2_{H_0^1(\Omega)} = \sum_{q=1}^{Q} |\hat{v}_q|^2 + \sum_{K \in \mathcal{T}} \sum_{k=2}^{\infty} |\hat{v}_{K,k}|^2 =: \sum_{\lambda \in \mathbb{L}} |\hat{v}_\lambda|^2, \quad (23)$$

where the new notation on the right-most side has been introduced for subsequent convenience. Similarly, for any $F \in H^{-1}(\Omega)$, we set

$$\hat{F}_q = \langle F, \psi_q \rangle \quad \text{and} \quad \hat{F}_{K,k} = \langle F, \phi_{K,k} \rangle,$$

and we obtain

$$\|F\|^2_{H^{-1}(\Omega)} = \sum_{q=1}^{Q} |\hat{F}_q|^2 + \sum_{K \in \mathcal{T}} \sum_{k=2}^{\infty} |\hat{F}_{K,k}|^2 =: \sum_{\lambda \in \mathbb{L}} |\hat{F}_\lambda|^2. \quad (24)$$

The formal analogy between (23) and (24) suggests us to use the notation $\|.\|$ to indicate both the $H_0^1(\Omega)$-norm of a function v, or the $H^{-1}(\Omega)$-norm of a linear form F; the specific meaning will be clear from the context.

Moreover, given any finite index set $\Lambda \subset \mathbb{L}$, we define the subspace of $H_0^1(\Omega)$

$$V_\Lambda := \operatorname{span}\{\varphi_\lambda \mid \lambda \in \Lambda\};$$

we set $|\Lambda| = \operatorname{card} \Lambda$, so that $\dim V_\Lambda = |\Lambda|$. If g admits an expansion $g = \sum_{\lambda\in\mathbb{L}} \hat{g}_\lambda \varphi_\lambda$ (converging in an appropriate norm), then we define its projection $P_\Lambda g$ onto V_Λ by setting

$$P_\Lambda g = \sum_{\lambda\in\Lambda} \hat{g}_\lambda \varphi_\lambda.$$

Also note that if $r = r(v_\Lambda)$ is a residual, then its norm is given by

$$\|r\|^2 = \sum_{\lambda\in\mathbb{L}} |\hat{r}_\lambda|^2, \tag{25}$$

with $\hat{r}_\lambda = \langle f, \varphi_\lambda\rangle - a(v_\Lambda, \varphi_\lambda)$.

Algebraic Representation and Properties of the Stiffness Matrix

Let us introduce the semi-infinite, symmetric and positive-definite matrix

$$\mathbf{A} = (a_{\lambda,\mu})_{\lambda,\mu\in\mathbb{L}} \quad \text{with } a_{\lambda,\mu} = a(\varphi_\mu, \varphi_\lambda). \tag{26}$$

Then, Problem (16) can be equivalently written as

$$\mathbf{Au} = \mathbf{f}, \tag{27}$$

where the vectors $\mathbf{u} = (\hat{u}_\mu)$ and $\mathbf{f} = (\hat{f}_\lambda)$ collect, respectively, the coefficients of the solution $u = \sum_\mu \hat{u}_\mu \varphi_\mu$ of Problem (16), and of the right-hand side f.

For any element $K \in \mathcal{T}$, let us denote by $\mathbf{A}_K$ the square block of $\mathbf{A}$ associated with the basis functions $\{\phi_{K,n}\}$ in K, i.e.,

$$\mathbf{A}_K = (a_{\lambda,\mu})_{\lambda,\mu\in\mathbb{L}(K)} \quad \text{where } \mathbb{L}(K) = \big\{\lambda \in \mathbb{L} : \lambda = (K, k) \text{ for some } k \geq 2\big\}.$$

For convenience, let us write $k = k(\lambda)$.

Let us assume that the restrictions to any $K \in \mathcal{T}$ of the coefficients ν and σ of the differential operator in (16) are analytic functions, so that their (scaled) Legendre coefficients in K decay at an exponential rate. Then, one can prove the existence of strictly positive constants η_A and c_A such that

$$|a_{\lambda,\mu}| \leq c_A \mathrm{e}^{-\eta_A|k(\lambda)-k(\mu)|} \quad \forall \lambda, \mu \in \mathbb{L}(K). \tag{28}$$

We say that any $\mathbf{A}_K$ belongs to the exponential class $\mathcal{D}_e(\eta_A, c_A)$. Note that it is not restrictive to assume η_A and c_A independent of K, since the partition $\mathcal{T}$ has been fixed once and for all.

The following properties hold (see [14, 15]).

Proposition 4.1 *Assume that the constant c_A satisfying (28) is such that for each $K \in \mathcal{T}$*

$$c_A < \frac{1}{2}\left(\mathrm{e}^{\eta_A} - 1\right) \min_{\lambda \in \mathbb{L}(K)} a_{\lambda,\lambda}. \tag{29}$$

Then each $\mathbf{A}_K$ is invertible and $\mathbf{A}_K^{-1} \in \mathcal{D}_e(\bar{\eta}_A, \bar{c}_A)$ for some $\bar{\eta}_A \in (0, \eta_A]$ and $\bar{c}_A > 0$.

Proposition 4.2 *For any $K \in \mathcal{T}$ and any integer $J \geq 0$, the truncated matrices $(\mathbf{A}_K)_J$ such that*

$$\left((\mathbf{A}_K)_J\right)_{\lambda,\mu} = \begin{cases} a_{\lambda,\mu} & \text{if } |k(\lambda) - k(\mu)| \leq J, \\ 0 & \text{elsewhere}, \end{cases} \tag{30}$$

satisfy the inequalities

$$\left\|\mathbf{A}_K - (\mathbf{A}_K)_J\right\| \leq C_{\mathbf{A}} \mathrm{e}^{-\eta_A J}$$

for some constant $C_{\mathbf{A}} > 0$. Furthermore, under the assumptions of Proposition 4.1, there exists a constant $\bar{C}_{\mathbf{A}} > 0$ such that

$$\left\|\mathbf{A}_K^{-1} - \left(\mathbf{A}_K^{-1}\right)_J\right\| \leq \bar{C}_{\mathbf{A}} \mathrm{e}^{-\bar{\eta}_A J}. \tag{31}$$

4.2 The Constitutive Elements of an Adaptive Algorithm

We are going to present an adaptive algorithm which implements the following recursion: (i) compute a Galerkin approximation of the exact solution; (ii) compute the corresponding residual, actually a feasible (finite-dimensional) version of it, so that its norm can be taken as an error estimator; (iii) apply Dörfler's marking, also known as bulk-chasing, to the components of the residual in order to identify a set of new basis functions to be activated for the next Galerkin solve; (iv) expand this set using properties of the stiffness matrix of the problem; (v) compute the new Galerkin solution on the enriched finite-dimensional subspace; (vi) get rid of the negligible components of this solution by applying a coarsening procedure.

We anticipate that step (iv) guarantees an arbitrarily large error reduction, whereas step (vi) assures a quasi-optimal complexity count.

We us now introduce the specific procedures, which will enter the definition of our adaptive algorithm.

- $u_\Lambda := \mathbf{GAL}(\Lambda)$

  Given a finite subset $\Lambda \subset \mathbb{L}$, the output $u_\Lambda \in V_\Lambda$ is the solution of the Galerkin problem,

$$u_\Lambda \in V_\Lambda : \quad a(u_\Lambda, v_\Lambda) = \langle f, v_\Lambda \rangle \quad \forall v_\Lambda \in V_\Lambda. \tag{32}$$

- $r := \mathbf{RES}(v_\Lambda)$
  Given a function $v_\Lambda \in V_\Lambda$ for some finite index set Λ, the output r is, in an ideal algorithm, the residual $r(v_\Lambda) = f - Av_\Lambda$. In a feasible version, the output, say $\tilde{r}$ is a function with a finite expansion along the chosen basis, obtained by suitably approximating the right-hand side f and the image Av_Λ; it satisfies the inequality
  $$\|r - \tilde{r}\| \leq \gamma \|\tilde{r}\|$$
  for some fixed constant $\gamma \in (0, 1)$. In the following, we restrict ourselves to the ideal case where the residual is assumed to be computed exactly.
- $\Lambda^* := \mathbf{DÖRFLER}(r, \theta)$
  Given $\theta \in (0, 1)$ and an element $r \in H^{-1}(I)$, the output $\Lambda^* \subset \mathbb{L}$ is a finite set of minimal cardinality such that the inequality
  $$\|P_{\Lambda^*} r\| \geq \theta \|r\|\,, \tag{33}$$
  or equivalently
  $$\|r - P_{\Lambda^*} r\| \leq \sqrt{1 - \theta^2}\|r\|, \tag{34}$$
  is satisfied. In terms of expansion coefficients, condition (33) can be equivalently stated as
  $$\sum_{\lambda \in \Lambda^*} |\hat{r}_\lambda|^2 \geq \theta^2 \sum_{\lambda \in \mathbb{L}} |\hat{r}_\lambda|^2. \tag{35}$$
  Thus, the output set Λ^* of minimal cardinality can be immediately determined by a greedy algorithm, i.e., by rearranging the coefficients $\hat{r}_\lambda$ in non-increasing order of modulus and retaining the largest ones until (35) is fulfilled.
- $\Lambda^* := \mathbf{ENRICH}(\Lambda, J)$
  Given an integer $J \geq 0$ and a finite set $\Lambda \subset \mathbb{L}$, the output is the set
  $$\Lambda^* := \left\{\mu = (K, k) \in \mathbb{L} : \text{ there exists } \lambda = \left(K, k'\right) \in \Lambda \text{ such that } \left|k - k'\right| \leq J\right\}.$$
  Note that Λ is enriched element-by-element with respect to the fixed partition $\mathcal{T} = \{K\}$ of Ω.
- $\Lambda^* := \mathbf{E\text{-}DÖRFLER}(r, \theta)$
  The two previous procedures are combined as follows. Given $\theta \in (0, 1)$ and an element $r \in H^{-1}(I)$, the output $\Lambda^* \subset \mathbb{L}$ is defined by the sequence
  $$\begin{aligned} \widetilde{\Lambda} &:= \mathbf{DÖRFLER}(r, \theta) \\ \Lambda^* &:= \mathbf{ENRICH}(\widetilde{\Lambda}, J_\theta), \end{aligned} \tag{36}$$
  where, based on Proposition 4.2, J_θ is chosen as the smallest integer which satisfies
  $$\bar{C}_{\mathbf{A}} \mathrm{e}^{-\bar{\eta}_A J} \leq \sqrt{\frac{1 - \theta^2}{\alpha_* \alpha^*}} \tag{37}$$
  (see [15] for more details).

- $\Lambda := \mathbf{COARSE}(w, \varepsilon)$
  Given a function $w \in V_{\Lambda^*}$ for some finite index set Λ^*, and an accuracy $\varepsilon > 0$ which is known to satisfy $\|u - w\| \leq \varepsilon$, the output $\Lambda \subseteq \Lambda^*$ is a set of minimal cardinality such that

$$\|w - P_\Lambda w\| \leq 2\varepsilon, \tag{38}$$

  which obviously implies $\|u - P_\Lambda w\| \leq 3\varepsilon$.

4.3 An Adaptive Algorithm with Convergence Rate

We are ready to present our adaptive algorithm. Each iteration can be viewed as a prediction step, based on the inspection of the current residual and the application of (enriched) Dörfler marking, followed by a correction step, based on coarsening. For this reason, we call it **PC-ADLEG**—Predictor-Corrector ADaptive LEGendre algorithm.

Given two parameters $\theta \in (0, 1)$ and $tol \in [0, 1)$, let us define

Algorithm PC-ADLEG(θ, tol)
Set $r_0 := f$, $\Lambda_0 := \emptyset$, $n = -1$
do
$n \leftarrow n + 1$
$\widehat{\partial\Lambda}_n := \mathbf{E\text{-}DÖRFLER}(r_n, \theta)$
$\widehat{\Lambda}_{n+1} := \Lambda_n \cup \widehat{\partial\Lambda}_n$
$\widehat{u}_{n+1} := \mathbf{GAL}(\widehat{\Lambda}_n)$
$\Lambda_{n+1} := \mathbf{COARSE}(\widehat{u}_{n+1}, \frac{2}{\alpha_*}\sqrt{1-\theta^2}\|r_n\|)$
$u_{n+1} := \mathbf{GAL}(\Lambda_{n+1})$
$r_{n+1} := \mathbf{RES}(u_{n+1})$
while $\|r_{n+1}\| > tol$

The following convergence result can be proven, by adapting the arguments given in [15] for the single-element case.

Theorem 4.1 *Let* $0 < \theta < 1$ *be chosen so that*

$$\rho = \rho(\theta) = 6\frac{\alpha^*}{\alpha_*}\sqrt{1-\theta^2} < 1. \tag{39}$$

If the assumptions of Proposition 4.1 *are fulfilled, the sequence of errors* $u - u_n$ *generated for* $n \geq 0$ *by the algorithm satisfies the inequality*

$$|||u - u_{n+1}||| \leq \rho |||u - u_n||| .$$

Thus, for any $tol > 0$ *the algorithm terminates in a finite number of iterations, whereas for* $tol = 0$ *the sequence* u_n *converges to* u *in* $H^1(I)$ *as* $n \to \infty$.

Note that the rate of decay of the error can be brought as close to 0 as desired by choosing θ close enough to 1. This is a feature stemming from the Enrichment procedure, keeping into account the decay properties of the inverse of the stiffness matrices $\mathbf{A}_K$, $K \in \mathcal{T}$.

4.4 Nonlinear Approximation in Gevrey Spaces

In order to estimate the complexity of our algorithm, and evaluate its optimality, we have to make assumptions on the structure of the solution u. Precisely, we have to make assumptions on the minimal number of degrees of freedom (i.e., active basis functions) needed to build an approximation of u within a given tolerance. This is usually expressed as the condition that u belongs to a suitable *sparsity class*. Once this is done, we can compare the number of degrees of freedom activated by our algorithm at a certain iteration (actually, an estimate of this number) to the minimal number of degrees of freedom needed to obtain the same accuracy; optimality usually means that the two numbers are within a constant independent of the solution and the current iteration.

Sparsity classes typically involved in finite-order approximations such as wavelets or h-type finite elements describe an *algebraic* decay of the best approximation error vs the number of activated degrees of freedom. Hereafter, we will rather consider sparsity classes describing an *exponential* decay of that error; this choice is coherent with the nature of our discretization approach, which uses an infinite-order spectral-element method, or p-type finite element method, hence providing faster-than-algebraic decay of the error whenever the solution is piecewise smooth on the partition $\mathcal{T}$ of the domain.

The definition of sparsity class is based on the concept of *best N-term approximation error*, that we now recall. Given any nonempty finite index set $\Lambda \subset \mathbb{L}$ and the corresponding subspace $V_\Lambda \subset V$ of dimension $|\Lambda| = \operatorname{card} \Lambda$, the best approximation of v in V_Λ is the orthogonal projection of v upon V_Λ, i.e. the function $P_\Lambda v = \sum_{\lambda \in \Lambda} \hat{v}_\lambda \varphi_\lambda$, which satisfies

$$\|v - P_\Lambda v\| = \left(\sum_{\lambda \notin \Lambda} |\hat{v}_\lambda|^2 \right)^{1/2}.$$

For any integer $N \geq 1$, we minimize this error over all possible choices of Λ with cardinality N, thereby leading to the best N-term approximation error

$$E_N(v) = \inf_{\Lambda \subset \mathbb{L}, |\Lambda| = N} \|v - P_\Lambda v\|.$$

A way to construct a best N-term approximation v_N of v consists of rearranging the coefficients of v in decreasing order of modulus

$$|\hat{v}_{\lambda_1}| \geq \cdots \geq |\hat{v}_{\lambda_n}| \geq |\hat{v}_{\lambda_{n+1}}| \geq \cdots$$

and setting $v_N = P_{\Lambda_N} v$ with $\Lambda_N = \{\lambda_n : 1 \leq n \leq N\}$.

We are ready to give the following fundamental definition.

Definition 4.1 Given two real numbers $\eta > 0$ and $t \in (0, 1]$, we denote by $\mathcal{A}_G^{\eta,t}(\Omega, \mathcal{T})$ the set defined as

$$\mathcal{A}_G^{\eta,t}(\Omega, \mathcal{T}) := \left\{ v \in V = H_0^1(\Omega) : \|v\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})} := \sup_{N \geq 0} E_N(v) \mathrm{e}^{\eta N^t} < +\infty \right\}.$$

As shown in [14], the set $\mathcal{A}_G^{\eta,t}(\Omega, \mathcal{T})$ is not a vector space, since it may happen that u, v belong to this set, whereas $u + v$ does not; however, one can show that $u + v \in \mathcal{A}_G^{\bar{\eta},t}(\Omega, \mathcal{T})$ with $\bar{\eta} = 2^{-t}\eta$.

The quantity $\|v\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})}$ dictates the minimal number N_ε of basis functions needed to approximate v with accuracy ε. In fact, from the relations

$$E_{N_\varepsilon}(v) \leq \varepsilon < E_{N_\varepsilon - 1}(v) \leq \mathrm{e}^{-\eta(N_\varepsilon - 1)^t} \|v\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})}, \tag{40}$$

we obtain

$$N_\varepsilon \leq \frac{1}{\eta^{1/t}} \left(\log \frac{\|v\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})}}{\varepsilon} \right)^{1/t} + 1. \tag{41}$$

In order to motivate our definition, let us first assume that $\mathcal{T} = \{\Omega\}$, i.e., let us concentrate on a single element. Then, inspired by [7], one can introduce the following family of spaces of Gevrey type: given any $v \in V$, let $v = \sum_{k \geq 2} \hat{v}_k \phi_k$ be its expansion along the Babuška-Shen basis defined as in (22) relative to the interval Ω. Then, we set

$$A^{\eta,t}(\Omega) = \left\{ v \in V : \text{there exists a constant } C > 0 \text{ such that } |\hat{v}_k| \leq C \mathrm{e}^{-\eta k^t} \ \forall k \geq 2 \right\}.$$

It is well-known that for $t = 1$ we get analytic functions in a neighborhood of $\overline{\Omega}$. A slightly stronger family of spaces is represented by the Sobolev-Gevrey spaces (see [28]; see also [39]) defined as

$$G^{\eta,t}(\Omega) = \left\{ v \in V : \|v\|^2_{G^{\eta,t}(\Omega)} := \sum_{k=2}^{\infty} e^{2\eta k^t} |\hat{v}_k|^2 < +\infty \right\}. \tag{42}$$

We immediately observe that $G^{\eta,t}(\Omega) \subset A^{\eta,t}(\Omega)$. Furthermore, given any $v \in G^{\eta,t}(\Omega)$ and approximating it by the linear projection

$$P_N v = \sum_{k=2}^{N} \hat{v}_k \phi_k,$$

we immediately get

$$E_N(v) \leq \|v - P_N v\| \leq \mathrm{e}^{-\eta N^t} \|v\|_{G^{\eta,t}(\Omega)},$$

which implies $G^{\eta,t}(\Omega) \subset \mathcal{A}_G^{\eta,t}(\Omega, \mathcal{T})$. Thus, the latter space contains, in particular, analytic functions and Gevrey functions.

Let us now consider an arbitrary finite partition $\mathcal{T}$ of Ω. For any $K \in \mathcal{T}$, let $v \in H_0^1(K)$; its best N-term approximation error in K is defined as

$$E_{K,N}(v) := \inf_{\Lambda \subset \mathbb{L}(K), |\Lambda| = N} \|v - P_\Lambda v\|_{H_0^1(K)}.$$

Consequently, we can define the class $\mathcal{A}_G^{\eta,t}(K)$ by setting

$$\mathcal{A}_G^{\eta,t}(K) := \left\{ v \in H_0^1(K) : \|v\|_{\mathcal{A}_G^{\eta,t}(K)} := \sup_{N \geq 0} E_{K,N}(v) \mathrm{e}^{\eta N^t} < +\infty \right\}.$$

Given any $v \in H_0^1(\Omega)$, denote by $v_L \in V_L(\mathcal{T})$ its piecewise linear interpolant, and set $\tilde{v} = v - v_L$, so that $\tilde{v}_{|K} \in H_0^1(K)$ for all $K \in \mathcal{T}$. Now, assume that $v \in \mathcal{A}_G^{\eta,t}(\Omega, \mathcal{T})$, and let w be a best N-term approximation of v, i.e., a linear combination of at most N basis functions (we will write $|\operatorname{supp} w| \leq N$) such that

$$\|v - w\| \leq \mathrm{e}^{-\eta N^t} \|v\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})}.$$

Writing $v - w = (v - v_L) - (w - w_L) + (v_L - w_L) = \tilde{v} - \tilde{w} + z_L$ and using the orthogonality of the basis functions, we have

$$\|v - w\|^2 = \|\tilde{v} - \tilde{w} + z_L\|^2 \geq \|\tilde{v} - \tilde{w}\|^2 = \sum_{K \in \mathcal{T}} \|\tilde{v}_K - \tilde{w}_K\|^2_{H_0^1(K)},$$

where the appended K denotes restriction of a function to K; thus,

$$\|\tilde{v}_K - \tilde{w}_K\|_{H_0^1(K)} \leq \mathrm{e}^{-\eta N^t} \|v\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})} \quad \forall K \in \mathcal{T}$$

and since $|\operatorname{supp} w_K| \leq |\operatorname{supp} w| \leq N$, we deduce that $\tilde{v}_K \in \mathcal{A}_G^{\eta,t}(K)$ for all $K \in \mathcal{T}$.

On the other hand, let $v \in H_0^1(\Omega)$ be such that $\tilde{v}_{|K} \in \mathcal{A}_G^{\tilde{\eta},t}(K)$ for all $K \in \mathcal{T}$, for some $\tilde{\eta} > 0$ to be determined later on. Then, there exist a constant $C > 0$ and functions $\tilde{w}_K \in H_0^1(K)$ with $|\operatorname{supp} \tilde{w}_K| \leq N$ such that

$$\|\tilde{v}_K - \tilde{w}_K\|_{H_0^1(K)} \leq C \mathrm{e}^{-\eta N^t} \quad \forall K \in \mathcal{T}.$$

Denoting by $\tilde{w}$ the function in Ω which coincides with $\tilde{w}_K$ in each K, and setting $w = v_L + \tilde{w}$, we have

$$\|v - w\|^2 = \|\tilde{v} - \tilde{w}\|^2 = \sum_{K \in \mathcal{T}} \|\tilde{v}_K - \tilde{w}_K\|^2_{H_0^1(K)} \leq (Q+1) C^2 \mathrm{e}^{-2\tilde{\eta} N^t}.$$

Now, observe that $|\operatorname{supp} w| = |\operatorname{supp} v_I| + \sum_{K \in \mathcal{T}} |\operatorname{supp} \tilde{w}_K| \leq Q + (Q+1)N \leq (Q+2)N$. Choosing $\tilde{\eta} = (Q+2)^t \eta$ and letting $N \to \infty$, we conclude that $v \in \mathcal{A}_G^{\eta,t}(\Omega, \mathcal{T})$.

4.5 Complexity Analysis of the Algorithm

We are now ready to investigate complexity issues for the sequence of approximations $u_n = u_{\Lambda_n}$ generated by **PC-ADLEG**, under the assumption that the solution u belongs to a class $\mathcal{A}_G^{\eta,t}(\Omega, \mathcal{T})$ for some $\eta > 0$ and $t \in (0, 1]$.

At first we note that each set Λ_{n+1} of the active degrees of freedom produced by the algorithm is generated by the procedure **COARSE** with a suitable tolerance ε_n. A general result about coarsening (see, e.g., [52]) allows us to estimate its cardinality $|\Lambda_{n+1}| = N_{\varepsilon_n}$ according to (41). On the other hand, one can prove that $\|u - u_{n+1}\| \lesssim \varepsilon_n$. Thus, we obtain the following optimal result.

Theorem 4.2 *Suppose that $u \in \mathcal{A}_G^{\eta,t}$, for some $\eta > 0$ and $t \in (0, 1]$. Then, there exists a constant $C > 1$ such that the cardinality of the set Λ_{n+1} of the active degrees of freedom produced by **PC-ADLEG** satisfies the bound*

$$|\Lambda_{n+1}| \leq \frac{1}{\eta^{1/t}} \left(\log \frac{\|u\|_{\mathcal{A}_G^{\eta,t}}}{\|u - u_{n+1}\|} + \log C \right)^{1/t} + 1, \quad \forall n \geq 0.$$

Next, we focus on the cardinality of the intermediate set $\widehat{\Lambda}_{n+1}$, which depends on that of the incremental set $\widehat{\partial \Lambda}_{n+1}$; in turns, this can be bounded by $2J_\theta$ times the cardinality of the incremental set $\widetilde{\partial \Lambda}_{n+1}$ generated by **DÖRFLER** with residual r_n. Although under certain assumptions on θ it is possible to estimate such cardinality in terms of the sparsity class of the solution (see [52]), in the most general situation as the one we want to consider here, it is the sparsity class of the residual that influences the growth of degrees of freedom. Indeed, we recall that the step

$$\partial \Lambda := \mathbf{DÖRFLER}(r, \theta)$$

selects a set $\partial \Lambda$ of minimal cardinality in $\mathbb{L} \setminus \Lambda$ for which $\|r - P_{\partial \Lambda} r\| \leq \sqrt{1 - \theta^2}\|r\|$. In other words, it performs a best approximation of the residual for the accuracy $\varepsilon = \sqrt{1 - \theta^2}\|r\|$. Thus, if r belongs to a certain sparsity class $\mathcal{A}_G^{\bar{\eta},\bar{t}}(\Omega, \mathcal{T})$ for some $\bar{\eta} > 0$ and $\bar{t} > 0$, we have by (41)

$$|\partial \Lambda| \leq \frac{1}{\bar{\eta}^{1/\bar{t}}} \left(\log \frac{\|r\|_{\mathcal{A}_G^{\bar{\eta},\bar{t}}}}{\sqrt{1 - \theta^2}\|r\|} \right)^{1/\bar{t}} + 1. \tag{43}$$

Thus, it make sense to investigate the sparsity class of the residual. In a sparsity class of algebraic type, this is the same as the class of the solution (see again [52]). Unfortunately, in a sparsity class of exponential type such a property does not hold [14], and we have to expect the generic residual to be less sparse than the exact solution.

The best result we can expect is as follows.

Proposition 4.3 *Let $v \in \mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})$ for some $\eta > 0$ and $t \in (0,1]$. Assume that $\eta < \eta_A$, where η_A is the constant for which (28) holds. Let us set*

$$\bar{\eta} = \zeta(t)\eta, \qquad \bar{t} = \frac{t}{1+t},$$

where we define

$$\zeta(t) := \left(\frac{1+t}{2}\right)^{\frac{t}{1+t}} \quad \forall 0 < t \leq 1. \tag{44}$$

Then, one has $Av \in \mathcal{A}_G^{\bar{\eta},\bar{t}}(\Omega,\mathcal{T})$, with

$$\|Av\|_{\mathcal{A}_G^{\bar{\eta},\bar{t}}(\Omega,\mathcal{T})} \lesssim \|v\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})}. \tag{45}$$

Under the sparsity assumption on the solution u made in the previous theorem, this implies that $f = Au \in \mathcal{A}_G^{\bar{\eta},\bar{t}}(\Omega,\mathcal{T})$. On the other hand, it is possible to prove that any Galerkin solution produced by **PC-ADLEG** satisfies $\|u_n\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})} \lesssim \|u\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})}$, so that $Au_n \in \mathcal{A}_G^{\bar{\eta},\bar{t}}(\Omega,\mathcal{T})$. Keeping into account the remark after Definition 4.1, we obtain the following result.

Proposition 4.4 *Let $u \in \mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})$ for some $\eta > 0$ and $t \in (0,1]$. There exists $\tilde{\eta} \leq \eta$ such that $r_n = r(u_n) \in \mathcal{A}_G^{\tilde{\eta},\bar{t}}(\Omega,\mathcal{T})$ with*

$$\|r_n\|_{\mathcal{A}_G^{\tilde{\eta},\bar{t}}(\Omega,\mathcal{T})} \lesssim \|u\|_{\mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})}.$$

Using (43), we arrive at the following final estimate.

Theorem 4.3 *Suppose that $u \in \mathcal{A}_G^{\eta,t}(\Omega,\mathcal{T})$ for some $\eta > 0$ and $t \in (0,1]$ and that the assumptions of Proposition 4.1 are satisfied. Then, there exist positive constants $\tilde{\eta} \leq \eta$, $\bar{t} \leq t$ and C such that the cardinality of the intermediate sets $\widehat{\Lambda}_{n+1}$ activated in the predictor step of **PC-ADLEG** can be estimated as*

$$|\widehat{\Lambda}_{n+1}| \leq |\Lambda_n| + \frac{2J_\theta}{\tilde{\eta}^{1/\bar{t}}}\left(\log\frac{\|u\|_{\mathcal{A}_G^{\eta,t}}}{\|u-u_{n+1}\|} + \log C\right)^{1/\bar{t}} + 2J_\theta, \quad \forall n \geq 0.$$

Keeping into account the conditions on $\tilde{\eta}$ and $\bar{t}$, we expect the cardinality of $|\widehat{\Lambda}_{n+1}|$ to be asymptotically larger than the optimal one of $|\Lambda_{n+1}|$, estimated in Theorem 4.2. Precisely for this reason, a coarsening step has been added at the end of each adaptive iteration: coarsening brings complexity from the one dictated by the sparsity class of the residual back to the one associated with the exact solution. On the other hand, we consider such intermediate loss of optimality to be worth of being accepted, since it should be compensated by the fast convergence of our algorithm, guaranteed by the allowed aggressive policy of degree of freedom enrichment.

We mention that the sparsity class of the residual influences complexity even in other instances of the algorithm, not discussed here. For instance, this is the case when a feasible computation of the residual-based error estimator is considered: to avoid degradation of the contraction property of the algorithm, approximate finite-dimensional residuals should be sufficiently close to the exact ones, which can be obtained with a complexity related to the sparseness of the residuals themselves. We refer to [14] for more details.

References

1. Ainsworth, M., Oden, J.T.: A procedure for a posteriori error estimation for h-p finite element methods. Comput. Methods Appl. Mech. Eng. **101**(1–3), 73–96 (1992)
2. Ainsworth, M., Oden, J.T.: A unified approach to a posteriori error estimation using element residual methods. Numer. Math. **65**(1), 23–50 (1993)
3. Ainsworth, M., Senior, B.: Aspects of an adaptive hp-finite element method: adaptive strategy, conforming approximation and efficient solvers. Comput. Methods Appl. Mech. Eng. **150**(1–4), 65–87 (1997). Symposium on Advances in Computational Mechanics, vol. 2 (Austin, TX, 1997)
4. Ainsworth, M., Senior, B.: An adaptive refinement strategy for hp-finite element computations. Appl. Numer. Math. **26**(1–2), 165–178 (1998)
5. Babuška, I., Guo, B.Q.: Regularity of the solutions of elliptic problems with piecewise analytic data. SIAM J. Numer. Anal. **20**, 763–781 (1989)
6. Babuška, I., Guo, B.Q.: Approximation properties of the h-p version of the finite element method. Comput. Methods Appl. Mech. Eng. **133**(3–4), 319–346 (1996)
7. Baouendi, M.S., Goulaouic, C.: Régularité analytique et itérés d'opérateurs elliptiques dégénérés; applications. J. Funct. Anal. **9**, 208–248 (1972)
8. Bernardi, C.: Indicateurs d'erreur en h-N version des éléments spectraux. Modél. Math. Anal. Numér. **30**(1), 1–38 (1996)
9. Binev, P., Dahmen, W., DeVore, R.: Adaptive finite element methods with convergence rates. Numer. Math. **97**(2), 219–268 (2004)
10. Braess, D., Pillwein, V., Schöberl, J.: Equilibrated residual error estimates are p-robust. Comput. Methods Appl. Mech. Eng. **198**(13–14), 1189–1197 (2009)
11. Bürg, M., Dörfler, W.: Convergence of an adaptive hp finite element strategy in higher space-dimensions. Appl. Numer. Math. **61**(11), 1132–1146 (2011)
12. Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A.: Spectral Methods. Fundamentals in Single Domains. Scientific Computation. Springer, Berlin (2006)
13. Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A.: Spectral Methods. Evolution to Complex Geometries. Scientific Computation. Springer, Berlin (2007)
14. Canuto, C., Nochetto, R.H., Verani, M.: Adaptive Fourier-Galerkin Methods, pp. 1–48 (2011, submitted). arXiv:1201.5648v1
15. Canuto, C., Nochetto, R.H., Verani, M.: Contraction and optimality properties of adaptive Legendre-Galerkin methods: the 1-dimensional case, pp. 1–26 (2011, submitted). arXiv:1206.5524v1
16. Cascon, J.M., Kreuzer, C., Nochetto, R.H., Siebert, K.G.: Quasi-optimal convergence rate for an adaptive finite element method. SIAM J. Numer. Anal. **46**(5), 2524–2550 (2008)
17. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods for elliptic operator equations—convergence rates. Math. Comput. **70**, 27–75 (1998)
18. Cohen, A., DeVore, R., Nochetto, R.H.: Convergence rates for AFEM with H^{-1} data. Found. Comput. Math. **12**(5), 671–718 (2012)

19. Costabel, M., Dauge, M., Nicaise, S.: Analytic regularity for linear elliptic systems in polygons and polyhedra. Math. Models Methods Appl. Sci. **22**(8), 1250015 (2012). doi:10.1142/S0218202512500157
20. Dahmen, W., Scherer, K.: Best approximation by piecewise polynomials with variable knots and degrees. J. Approx. Theory **26**(1), 1–13 (1979)
21. Dahmen, W., Scherer, K.: On optimal global error bounds obtained by scaled local error estimates. Numer. Math. **36**, 151–176 (1981)
22. Demkowicz, L., Oden, J.T., Rachowicz, W., Hardy, O.: Toward a universal h-p adaptive finite element strategy. I. Constrained approximation and data structure. Comput. Methods Appl. Mech. Eng. **77**(1–2), 79–112 (1989)
23. Demkowicz, L., Rachowicz, W., Devloo, Ph.: A fully automatic hp-adaptivity. J. Sci. Comput. **17**(1–4), 127–155 (2002)
24. DeVore, R., Scherer, K.: Variable knot, variable degree spline approximation to x^β. In: Quantitative Approximation. Proc. Internat. Sympos., Bonn, 1979, pp. 121–131. Academic Press, New York (1980).
25. Dörfler, W.: A convergent adaptive algorithm for Poisson's equation. SIAM J. Numer. Anal. **33**(3), 1106–1124 (1996)
26. Dörfler, W., Heuveline, V.: Convergence of an adaptive hp finite element strategy in one space dimension. Appl. Numer. Math. **57**(10), 1108–1124 (2007)
27. Eibner, T., Melenk, J.M.: An adaptive strategy for hp-FEM based on testing for analyticity. Comput. Mech. **39**(5), 575–595 (2007)
28. Foias, C., Temam, R.: Gevrey class regularity for the solutions of the Navier-Stokes equations. J. Funct. Anal. **87**(2), 359–369 (1989)
29. Gantumur, T., Harbrecht, H., Stevenson, R.: An optimal adaptive wavelet method without coarsening of the iterands. Math. Comput. **76**(258), 615–629 (2007)
30. Gui, W., Babuška, I.: The h, p and h-p versions of the finite element method in 1 dimension. II. The error analysis of the h- and h-p versions. Numer. Math. **49**(6), 613–657 (1986)
31. Gui, W., Babuška, I.: The h, p and h-p versions of the finite element method in 1 dimension. III. The adaptive h-p version. Numer. Math. **49**(6), 659–683 (1986)
32. Guo, B., Babuška, I.: The hp-version of the finite element method I: the basic approximation results. Comput. Mech. **1**, 21–41 (1986)
33. Guo, B., Babuška, I.: The hp-version of the finite element method II: general results and applications. Comput. Mech. **1**, 203–226 (1986)
34. Guo, B., Babuška, I.: Regularity of the solutions for elliptic problems on nonsmooth domains in $\mathbf{R}^3$. II. Regularity in neighbourhoods of edges. Proc. R. Soc. Edinb., Sect. A **127**(3), 517–545 (1997)
35. Heuveline, V., Rannacher, R.: Duality-based adaptivity in the hp-finite element method. J. Numer. Math. **11**(2), 95–113 (2003)
36. Houston, P., Senior, B., Süli, E.: hp-discontinuous Galerkin finite element methods for hyperbolic problems: error analysis and adaptivity. Int. J. Numer. Methods Fluids **40**(1–2), 153–169 (2002). ICFD Conference on Numerical Methods for Fluid Dynamics (Oxford, 2001)
37. Houston, P., Süli, E.: A note on the design of hp-adaptive finite element methods for elliptic partial differential equations. Comput. Methods Appl. Mech. Eng. **194**(2–5), 229–243 (2005)
38. Kurtz, J., Demkowicz, L.: A fully automatic hp-adaptivity for elliptic PDEs in three dimensions. Comput. Methods Appl. Mech. Eng. **196**(37–40), 3534–3545 (2007)
39. Lions, J.-L., Magenes, E.: Problèmes aux Limites Non Homogènes et Applications, vol. III. Dunod, Paris (1970)
40. Melenk, J.M., Wohlmuth, B.I.: On residual-based a posteriori error estimation in hp-FEM. Adv. Comput. Math. **15**(1–4), 311–331 (2002). 2001
41. Mitchell, W.F., McCain, M.: A comparison of hp-adaptive strategies for elliptic partial differential equations. In: NIST Report, pp. 1–39 (2011, submitted)
42. Morin, P., Nochetto, R.H., Siebert, K.G.: Data oscillation and convergence of adaptive FEM. SIAM J. Numer. Anal. **38**(2), 466–488 (2000) (electronic)

43. Nochetto, R.H., Siebert, K.G., Veeser, A.: Theory of adaptive finite element methods: an introduction. In: Multiscale, Nonlinear and Adaptive Approximation, pp. 409–542. Springer, Berlin (2009)
44. Oden, J.T., Demkowicz, L., Rachowicz, W., Westermann, T.A.: Toward a universal h-p adaptive finite element strategy. II. A posteriori error estimation. Comput. Methods Appl. Mech. Eng. **77**(1–2), 113–180 (1989)
45. Oden, J.T., Patra, A., Feng, Y.: An hp adaptive strategy. In: Noor, A.K. (ed.) Adaptive Multilevel and Hierarchical Computational Strategies. ASME Publication, vol. 157, pp. 23–46 (1992)
46. Rachowicz, W., Oden, J.T., Demkowicz, L.: Toward a universal h-p adaptive finite element strategy. III. Design of h-p meshes. Comput. Methods Appl. Mech. Eng. **77**(1–2), 181–212 (1989)
47. Schmidt, A., Siebert, K.G.: A posteriori estimators for the h-p version of the finite element method in 1D. Appl. Numer. Math. **35**(1), 43–66 (2000)
48. Schötzau, D., Schwab, C., Wihler, T.: hp-dGFEM for elliptic problems in polyhedra I: stability and quasi-optimality on geometric meshes. In: SAM Report ETHZ (2012)
49. Schötzau, D., Schwab, C., Wihler, T.: hp-dGFEM for elliptic problems in polyhedra II: exponential convergence. SAM Report ETHZ, (2012)
50. Schwab, C.: p- and hp-finite element methods. In: Theory and Applications in Solid and Fluid Mechanics. Numerical Mathematics and Scientific Computation. The Clarendon Press/Oxford University Press, New York (1998)
51. Stevenson, R.: Optimality of a standard adaptive finite element method. Found. Comput. Math. **7**(2), 245–269 (2007)
52. Stevenson, R.: Adaptive wavelet methods for solving operator equations: an overview. In: Multiscale, Nonlinear and Adaptive Approximation, pp. 543–597. Springer, Berlin (2009)
53. Valenciano, J., Owens, R.G.: A new adaptive modification strategy for numerical solutions to elliptic boundary value problems. Appl. Numer. Math. **32**(3), 305–329 (2000)
54. Wihler, T.P.: An hp-adaptive strategy based on continuous Sobolev embeddings. J. Comput. Appl. Math. **235**(8), 2731–2739 (2011)

A Theory and Challenges for Coarsening in Microstructure

Katayun Barmak, Eva Eggeling, Maria Emelianenko, Yekaterina Epshteyn, David Kinderlehrer, Richard Sharp, and Shlomo Ta'asan

Abstract Cellular networks are ubiquitous in nature. Most engineered materials are polycrystalline microstructures composed of a myriad of small grains separated by grain boundaries, thus comprising cellular networks. The grain boundary character distribution (GBCD) is an empirical distribution of the relative length (in 2D) or area (in 3D) of interface with a given lattice misorientation and normal. During the coarsening, or growth, process, an initially random grain boundary arrangement

Dedicated to the memory of Enrico Magenes.

Research supported by NSF DMR0520425, DMS 0405343, DMS 0305794, DMS 0806703, DMS 0635983, DMS 0915013, DMS 1056821, DMS 1216433, OISE 0967140, DMS 1112984.

K. Barmak
Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA
e-mail: kb2612@columbia.edu

E. Eggeling
Visual Computing, Fraunhofer Austria Research GmbH, 8010 Graz, Austria
e-mail: eva.eggeling@fraunhofer.at

M. Emelianenko
Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030, USA
e-mail: memelian@gmu.edu

Y. Epshteyn
Department of Mathematics, The University of Utah, Salt Lake City, UT 84112, USA
e-mail: epshteyn@math.utah.edu

D. Kinderlehrer (✉) · S. Ta'asan
Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: davidk@cmu.edu

S. Ta'asan
e-mail: shlomo@andrew.cmu.edu

R. Sharp
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA
e-mail: rsharp@gmail.com

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*, Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_12,

reaches a steady state that is strongly correlated to the interfacial energy density. In simulation, if the given energy density depends only on lattice misorientation, then the steady state GBCD and the energy are related by a Boltzmann distribution. This is among the simplest non-random distributions, corresponding to independent trials with respect to the energy. Why does such simplicity emerge from such complexity?

Here we an describe an entropy based theory which suggests that the evolution of the GBCD satisfies a Fokker-Planck Equation, an equation whose stationary state is a Boltzmann distribution. The properties of the evolving network that characterize the GBCD must be identified and appropriately upscaled or 'coarse-grained'. This entails identifying the evolution of the statistic in terms of the recently discovered Monge-Kantorovich-Wasserstein implicit scheme. The undetermined diffusion coefficient or temperature parameter is found by means of a convex optimization problem reminiscent of large deviation theory.

1 Introduction

Cellular networks are ubiquitous in nature. They exhibit behavior on many different length and time scales and are generally metastable. Most technologically useful materials are polycrystalline microstructures composed of a myriad of small monocrystalline grains separated by grain boundaries, and thus comprise cellular networks. The energetics and connectivity of the grain boundary network plays a crucial role in determining the properties of a material across a wide range of scales. A central problem is to develop technologies capable of producing an arrangement of grains that provides for a desired set of material properties. Traditionally the focus has been on distributions of geometric features, like cell size, and a preferred distribution of grain orientations, termed texture. Attaining these gives the configuration order in a statistical sense. More recent mesoscale experiment and simulation permit harvesting large amounts of information about both geometric features and crystallography of the boundary network in material microstructures, [1, 2, 43, 58, 59]. This has led us to the notion of the Grain Boundary Character Distribution.

The grain boundary character distribution (GBCD) is an empirical distribution of the relative length (in 2D) or area (in 3D) of interface with a given lattice misorientation and grain boundary normal.

A first discovery is that during the growth process, an initially random grain boundary arrangement reaches a steady state that is strongly correlated to the interfacial energy density. In simulation, a stationary GBCD is always found. Moreover there is consistency between experimental GBCD's and simulated GBCD's. The boundary network of a cellular structure is naturally ordered.

A second discovery is that if the given interfacial energy depends only on lattice misorientation, then the steady state GBCD and the density are related by a Boltzmann distribution. This is among the simplest non-random distributions, corresponding to independent trials with respect to the density. Such straightforward dependence between the character distribution and the interfacial energy offers evidence that the GBCD is a material property. It is a leading candidate to characterize

texture of the boundary network [43]. Why does such simplicity emerge from such complexity?

Here we describe our recent work developing an entropy based theory that suggests that the evolving GBCD satisfies a Fokker-Planck Equation, [13, 14], cf. also [12, 15], to which we refer for a more complete exposition. Coarsening in polycrystalline systems is a complicated process involving details of material structure, chemistry, arrangement of grains in the configuration, and environment. In this context, we consider just two competing global features, as articulated by C.S. Smith [60]: cell growth according to a local evolution law and space filling constraints. We shall impose curvature driven growth for the local evolution law, cf. Mullins [54]. Space filling requirements are managed by critical events, rearrangements of the network involving deletion of small contracting cells and facets. The properties of this system that characterize the GBCD must be identified and appropriately upscaled or 'coarse-grained'. For a perspective on these issues, we recommend the article by R.V. Kohn [45].

The general platform for this investigation is large scale computation. Numerical simulations are well established as a major tool in the analysis of many physical systems, see for example [22, 23, 25, 26, 29, 30, 46, 48–50, 52, 56, 61, 62, 64]. However, the idea of large scale computation as the essential method for the modeling and comprehension of large complex systems is relatively new. Porous media and groundwater flow is an important case of this, see for example [4–7, 33]. For coarsening of cellular systems, it is a natural approach as well. The laboratory is the venue to assess the validity of the local evolution law. Once this law is adopted, we appeal to simulation, since we cannot control all the other elements present in the experimental system, many of which are unknown. On the other hand, *in silico* we may exercise, or at least we may attempt to exercise, precise control of the variables appropriate to the evolution law and the constraint.

There are many large scale metastable material systems, for example, magnetic hysteresis, [19], and second phase coarsening, [51, 66]. In these, the theory is based on mesoscopic or macroscopic variables simply abstracting the role of the smaller scale elements of the system. There is no general 'multiscale' framework for upscaling from the local behavior of individual cells to behavior of the network when they interact and change their character. So we must attempt to tease the system level information from the many coupled elements of which it consists. This information will be available primarily from the dissipation relation (2.6) which is implied by the balance of forces at triple junctions (2.3), due to Herring, [34, 35]. Lax resolution of the Herring Condition gives rise to an unreliable GBCD.

Our strategy is to introduce a simplified coarsening model that is driven by the boundary conditions and reflects the dissipation relation of the grain growth system. This will be more accessible to analysis. It resembles an ensemble of inertia-free spring-mass-dashpots. For this simpler network, we learn how entropic or diffusive behavior at the large scale emerges from a dissipation relation at the scale of local evolution. The cornerstone is a novel implementation of the iterative scheme for the Fokker-Planck Equation in terms of the system free energy and a Kantorovich-Rubinstein-Wasserstein metric [39], cf. also [38], which will be summarized later in the presentation.

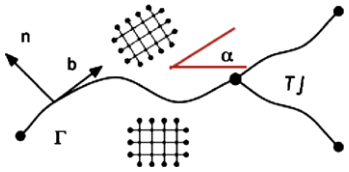


Fig. 1 An arc Γ with normal n, tangent b, and lattice misorientation α, illustrating lattice elements

The network level nonequilibrium nature of the scheme leaves undetermined the diffusion constant in the Fokker-Planck Equation, or equivalently the 'temperature parameter' of the Boltzmann Distribution we are seeking. We employ the Kullback-Leibler relative entropy, cf. (5.2), and find a convex duality problem for this parameter. It has a statistical interpretation, or information theory interpretation, in terms of an optimal prefix code, cf. e.g. [57], and moreover has evident connections to large deviations. This suggests that had we simply asked to identify an optimal distribution via a known statistical method, we would have been led *full circle* to entropy methods.

2 Reprise of Mesoscale Theory

Our point of departure is the common denominator theory for the mesoscale description of microstructure evolution. This is growth by curvature, the Mullins Equation (2.2) below, for the evolution of curves or arcs individually or in a network, which we employ for our local law of evolution. Boundary conditions must be imposed where the arcs meet. This condition is the Herring Condition, (2.3), which is the natural boundary condition at equilibrium for the Mullins Equation. Since their introduction by Mullins, [54], and Herring, [34, 35], a large and distinguished body of work has grown about these equations. Most relevant to here are [20, 32, 41, 55]. Curvature driven growth has old origins, dating at least to Burke and Turnbull [21]. Let α denote the misorientation between two grains separated by an arc Γ, as noted in Fig. 1, with normal $n = (\cos\theta, \sin\theta)$, tangent direction b and curvature κ. Let $\psi = \psi(\theta, \alpha)$ denote the energy density on Γ. So

$$\Gamma : x = \xi(s,t), \quad 0 \leqq s \leqq L, \ t > 0, \tag{2.1}$$

with

$$b = \frac{\partial \xi}{\partial s} \quad \text{(tangent)} \quad \text{and} \quad n = Rb \quad \text{(normal)},$$
$$v = \frac{\partial \xi}{\partial t} \quad \text{(velocity)} \quad \text{and} \quad v_n = v \cdot n \quad \text{(normal velocity)},$$

where R is a positive rotation of $\pi/2$. The Mullins Equation of evolution is

$$v_n = (\psi_{\theta\theta} + \psi)\kappa \quad \text{on } \Gamma. \tag{2.2}$$

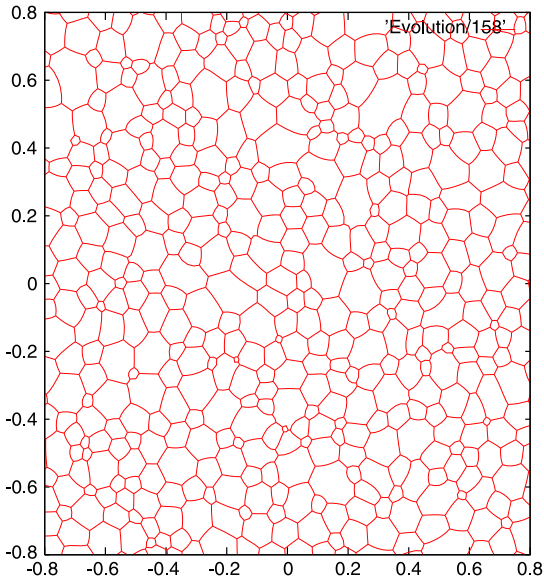


Fig. 2 Example of an instant during the simulated evolution of a cellular network. This is from a small simulation with constant energy density and periodic conditions at the border of the configuration

We assume that only triple junctions are stable and that the Herring Condition holds at triple junctions. This means that whenever three curves, $\{\Gamma^{(1)}, \Gamma^{(2)}, \Gamma^{(3)}\}$, meet at a point p the force balance, (2.3) below, holds:

$$\sum_{i=1,\ldots,3} \left(\psi_\theta n^{(i)} + \psi b^{(i)}\right) = 0. \tag{2.3}$$

It is easy to check that the instantaneous rate of change of energy of Γ is

$$\frac{d}{dt}\int_\Gamma \psi|b|ds = -\int_\Gamma v_n^2 ds + v \cdot (\psi_\theta n + \psi b)|_{\partial\Gamma}. \tag{2.4}$$

Consider a network of grains bounded by $\{\Gamma_i\}$ subject to some condition at the border of the region they occupy, like fixed end points or periodicity, cf. Fig. 2. The important features of the algorithm used in the current simulation are given briefly in the next Sect. 3. For the description of the previous algorithms the reader can consult [42, 44]. The typical simulation consists in initializing a configuration of cells and their boundary arcs, usually by a modified Voronoi tessellation, and then solving the system (2.2), (2.3), eliminating facets when they have negligible length and cells when they have negligible area, cf. Sect. 3. The total energy of the system is given by

$$E(t) = \sum_{\{\Gamma_i\}} \int_{\Gamma_i} \psi|b|ds. \tag{2.5}$$

Owing exactly to the Herring Condition (2.3), the instantaneous rate of change of the energy

$$\begin{aligned}\frac{d}{dt}E(t) &= -\sum_{\{\Gamma_i\}}\int_{\Gamma_i} v_n^2 ds + \sum_{TJ} v\cdot\sum(\psi_\theta n + \psi b)\\ &= -\sum_{\{\Gamma_i\}}\int_{\Gamma_i} v_n^2 ds\\ &\leqq 0,\end{aligned} \tag{2.6}$$

rendering the network dissipative for the energy in any instant absent of critical events. Indeed, in an interval $(t_0, t_0+\tau)$ where there are no critical events, we may integrate (2.6) to obtain a local dissipation equation

$$\sum_{\{\Gamma_i\}}\int_{t_0}^{t_0+\tau}\int_{\Gamma_i} v_n^2 ds dt + E(t_0+\tau) = E(t_0) \tag{2.7}$$

which bears a strong resemblance to the simple dissipation relation for an ensemble of inertia free springs with friction. In the simulation, the facet interchange and cell deletion are arranged so that (2.6) is maintained.

Suppose, for simplicity, that the energy density is independent of the normal direction, so $\psi = \psi(\alpha)$. It is this situation that will concern us here. Then (2.2) and (2.3) may be expressed

$$v_n = \psi\kappa \quad \text{on } \Gamma, \tag{2.8}$$

$$\sum_{i=1,\ldots,3} \psi b^{(i)} = 0 \quad \text{at } p, \tag{2.9}$$

where p denotes a triple junction. Equation (2.9) is the same as the Young wetting law.

For this situation we define the grain boundary character distribution, GBCD,

$$\rho(\alpha, t) = \text{relative length of arc of misorientation } \alpha \text{ at time } t,$$

$$\text{normalized so that } \int_\Omega \rho d\alpha = 1. \tag{2.10}$$

3 Discussion of the Simulation

Our simulation method involves a boundary tracking approach which is in contrast to level set methods, for example recent work [24], and phase field methods, for example [28, 40], used by other groups. We approximate only the network of grain boundaries while other methods involve the interior of the grains as well. The location of grain boundaries is implicit in these methods. The advantage of our method

is the flexibility in applying a selected energy density on curves and the boundary conditions at triple junctions. The evolution in our approach is based on a variational approach for solving numerically the system (2.2), (2.3) for the network while managing the critical events. It must be designed so it is robust and reliable statistics can be harvested. Owing to the size and complexity of the network there are number of challenges in the designing of the method. These include

- management of the data structure of cells, facets, and triple junctions, dynamic because of critical events;
- management of the computational domain;
- initialization of the computation;
- maintaining the triple junction boundary condition (2.3) while
- resolving Eqs. (2.2) with sufficient accuracy

We will address some of these issues below. We also need some diagnostics to understand the accuracy of the physical model and of the numerical scheme. Questions of numerical accuracy can be addressed by mesh refinement and the convergence of the approximate solutions. For physical accuracy of the model we look at certain properties of the solutions. For example, it is known that the average area of cells grows linearly even in very casual simulations of coarsening, although more careful diagnostics show that the Herring Condition (2.3) in these efforts fails. As noted in the introduction, this will lead to an unreliable determination of the GBCD.

In view of the dissipation inequality (2.6) the evolution of the grain boundary system may be viewed as a modified steepest descent for the energy. Therefore, the cornerstone of our scheme which assures its stability is the discrete dissipation inequality for the total grain boundary energy which holds when the discrete Herring Condition is satisfied. In general, discrete dissipation principles ensure the stability and convergence of numerical schemes to the continuous solution. The design of our numerical scheme is based on a weak formulation, a variational principle which avoids the additional complexity of higher order spaces. In particular, there is no explicit use of curvature which is the case for direct discretization of Eqs. (2.2), (2.3).

The simulation of the grain network is done in three steps by evolving first the grain boundaries, according to Mullin's equation (2.2), and then updating the triple points according to Herring's boundary condition (2.3), imposed at the triple junctions, and finally managing the rearrangement events. In our numerical simulations, grain boundaries are defined by the set of nodal points and are approximated using linear elements. In the algorithm, we define a global mesh size, h, and uniformly discretized grain boundaries with local mesh size (distance between neighboring nodal points) which depends on h. Due to the frequency of critical events, we have used a first order method in time, namely the Forward Euler method. Increasing the order of time discretization to 2 by using a predictor corrector method did not affect the distribution functions, which is the focus of this study.

Resolution of the Herring Condition: To satisfy the Herring Condition (2.3) one has to solve the nonlinear equation to determine the new position of the triple junction

[44]. We use the Newton method with line search [37] to approximate the new position for the triple junction. As the initial guess for Newton's method, we determine the position of the triple point by defining the velocity of the triple junction to be proportional to the total line stress at that point with coefficient of the proportionality equal to the mobility. This is also dissipative for the network. The Newton algorithm stops if it exceeds a certain tolerance on the number of the iterations. If the Newton algorithm converges, the Herring Condition (2.3) is satisfied to machine precision accuracy at the new position of the triple junction. If the Newton algorithm fails to converge at some triple junctions (this happens when we work with very small cells) we use our initial guess to update the triple junction position.

Critical events: As grain growth proceeds, critical events occur. When grain boundaries (GB) shrink below a certain size, they trigger one or more of the following processes (i) short GB removal, (ii) splitting of unstable junctions (where more then three GB meet), (iii) fixing double GB (GB that share two vertices).

Removal of short GB: A short GB whose length is decreasing is removed. If its length is increasing, it is not removed.

Splitting unstable vertices: When a GB disappears, new vertices may appear where more than three edges meet. These are unstable junctions which are split by introducing a new vertex and a new GB of short length. This step reduces the number of edges meeting at the unstable junctions. This process continues until all vertices are triple junctions. Details of each split are designed to maximally decrease the energy.

4 A Simplified Coarsening Model with Entropy and Dissipation

The coarsening process is irreversible because of its dissipative nature. Even in an interlude when there are no rearrangement events, (2.7) shows that a configuration cannot evolve to a former state from a later one. This could be a source of entropy for the system. In our investigation, we view the principal source of entropy to be configurational since we observe the evolution of an 'upscaled' ensemble represented by a single statistic, the misorientation α, neglecting the remaining information. This is also a source of irreversibility since we have forgotten information. We return to this shortly.

A significant difficulty in developing a theory for the GBCD, and understanding texture development in general, lies in the lack of understanding of consequences of rearrangement events or critical events, facet interchange and grain deletion, on network level properties. For example, in Fig. 3, the average area of five-faceted grains during a growth experiment on an *Al* thin film and the average area of five-faceted cells in a typical simulation both increase with time. Now the von Neumann-Mullins Rule is that the area A_n of a cell with n-facets satisfies

$$A_n'(t) = c(n-6), \tag{4.1}$$

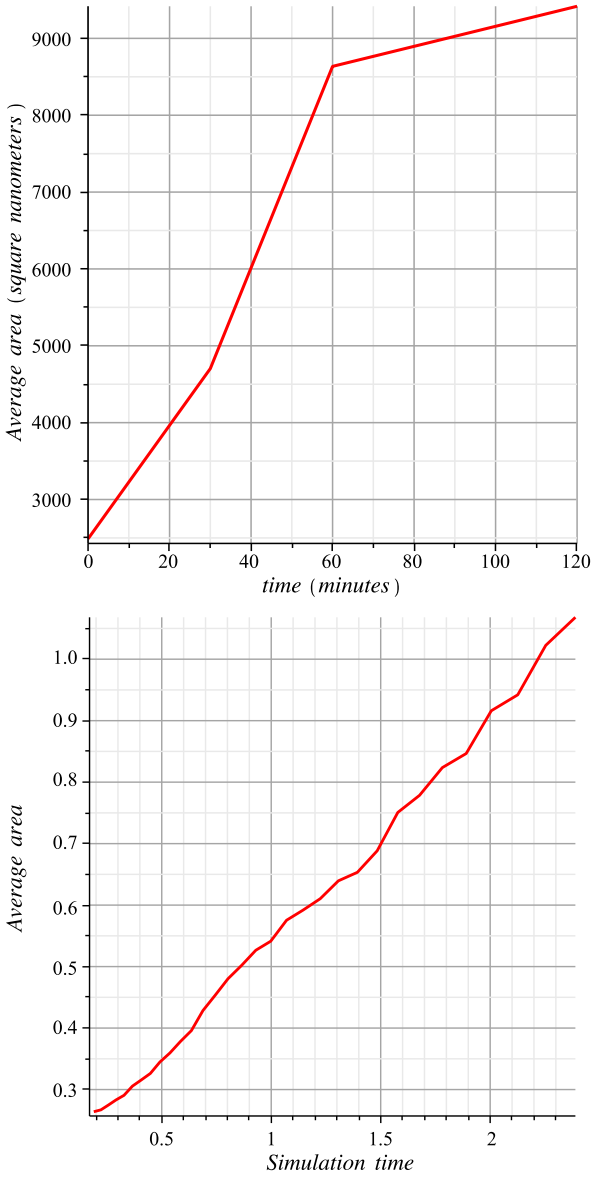


Fig. 3 The average area of five-sided cell populations during coarsening in two different cellular systems showing that the von Neumann-Mullins $n - 6$-Rule (4.1) does not hold at the scale of the network. (*Top*) In an experiment on *Al* thin film, [8], and (*bottom*) a typical simulation (arbitrary units)

when $\psi = \text{const.}$ and triple junctions meet at angles of $2\pi/3$, [31, 53, 65]. This is thought to hold approximately when anisotropy is small. The von Neumann-Mullins Rule does not fail in the example above, of course, but cells observed at later times had 6, 7, 8, ... facets at earlier times. Thus in the network setting, changes which rearrange the network play a major role.

To address these issues, we will examine a much simpler 1D model which retains kinetics and critical events but neglects curvature driven growth of the boundaries.

In our view, there are two important features of the coarsening system: the evolution of the network by steepest descent of the surface energy and the irreversible change/disappearance of the grain boundaries at certain discrete times, which is necessary because the entire configuration is confined. Elaborating on the latter in the two-dimensional setting of Fig. 2: at most times the evolution is smooth, but once in a while a pair of neighboring triple points collides and the grain boundary that joins them disappears forever.

We have used this model to develop a statistical theory for critical events, [9–11]. It has been found to have its own GBCD as well, [12–15], which we shall now review.

Our main idea in [12–15] is that the GBCD statistic for the simplified model resembles the solution of a Fokker-Planck Equation via the mass transport implicit scheme, [39]. In [12–15] the simplified model is formulated as a gradient flow which results in a dissipation inequality analogous to the one found for the coarsening grain network. Because of this simplicity, it will be possible to 'upscale' the network level system description to a higher level GBCD description that accommodates irreversibility. A more useful dissipation inequality is obtained by modifying the viscous term to be a mass transport term, which now brings us to the realm of the Kantorovich-Rubinstein-Wasserstein implicit scheme. As this changes the ensemble, there is an entropic contribution, which we take to be proportional to configurational entropy. This then suggests the Fokker-Planck paradigm.

However, we do not know that the statistic solves the Fokker-Planck PDE but we can ask if it shares important aspects of Fokker-Planck behavior. We give evidence for this by asking for the unique 'temperature-like' parameter, the factor noted above, the relative entropy achieves a minimum over long time. The empirical stationary distribution and Boltzmann distribution with the special value of 'temperature' are in excellent agreement. This gives an explanation for the stationary distribution and the kinetics of evolution. At this point of our investigations, we do not know that the two dimensional network has the detailed dissipative structure of the simplified model, but we are able to produce evidence that the same argument employing the relative entropy does suggest the correct kinetics and stationary distribution.

4.1 Formulation

The simplified coarsening model, driven by the boundary conditions, reflects the dissipation relation of the grain growth system. It resembles an ensemble of inertia-free spring-mass-dashpots. It is an abstraction of the role of triple junctions in the presence of the rearrangement events.

Let $I \subset \mathbf{R}$ be an interval of length L partitioned by points $x_i, i = 1, \ldots, n$, where $x_i < x_{i+1}, i = 1, \ldots, n-1$ and x_{n+1} identified with x_1. For each interval $[x_i, x_{i+1}]$, $i = 1, \ldots, n$ select a random misorientation number $\alpha_i \in (-\pi/4, \pi/4]$. The intervals $[x_i, x_{i+1}]$ correspond to grain boundaries (but not the 1D "grain") with misorientations α_i and the points x_i represent the triple junctions. Choose an energy

density $\psi(\alpha) \geqq 0$ and introduce the energy

$$E = \sum_{i=1,\ldots,n} \psi(\alpha_i)(x_{i+1} - x_i). \tag{4.2}$$

To have consistency with the evolution of the 2D cellular network, we impose gradient flow kinetics with respect to (4.2), which is just the system of ordinary differential equations

$$\begin{aligned} &\frac{dx_i}{dt} = -\frac{\partial E}{\partial x_i}, \quad i = 1, \ldots, n, \text{ that is} \\ &\frac{dx_i}{dt} = \psi(\alpha_i) - \psi(\alpha_{i-1}), \quad i = 2, \ldots, n, \quad \text{and} \quad \frac{dx_1}{dt} = \psi(\alpha_1) - \psi(\alpha_n). \end{aligned} \tag{4.3}$$

The velocity v_i of the ith boundary is

$$v_i = \frac{dx_{i+1}}{dt} - \frac{dx_i}{dt} = \psi(\alpha_{i-1}) - 2\psi(\alpha_i) + \psi(\alpha_{i+1}). \tag{4.4}$$

The grain boundary velocities are constant until one of the boundaries collapses. That segment is removed from the list of current grain boundaries and the velocities of its two neighbors are changed due to the emergence of a new junction. Each such deletion event rearranges the network and, therefore, affects its subsequent evolution just as in the two dimensional cellular network. Actually, since the interval velocities are constant, this gradient flow is just a sorting problem. At any time, the next deletion event occurs at smallest positive value of

$$\frac{x_i - x_{i+1}}{v_i}.$$

The length $l_i(t)$ of the ith interval is linear in t until it reaches 0 or until a collision event, when it becomes linear with a different slope. In any event, it is continuous, so $E(t), t > 0$, the sum of such functions multiplied by factors, is continuous.

At any time t between deletion events,

$$\frac{dE}{dt} = -\sum \frac{dx_i}{dt}^2 \leqq 0. \tag{4.5}$$

Next consider for the 1D system (4.3), a time interval $(t_0, t_0 + \tau)$ with no critical events for now. Then we obtain a grain growth analog of the spring-mass-dashpot-like local dissipation inequality,

$$\sum_{i=1\ldots n} \int_0^\tau \frac{dx_i^2}{dt} dt + E(t_0 + \tau) = E(t_0). \tag{4.6}$$

With an appropriate interpretation of the sum, (4.6) holds for all t_0 and almost every τ sufficiently small. The dissipation equality (4.6) can also be rewritten in terms of

grain boundary velocities as follows:

$$\frac{1}{4}\sum_{i=1\ldots n}\int_0^{\tau} v_i^2 dt + E(t_0+\tau) \leqq E(t_0). \tag{4.7}$$

The energy of the system at time $t_0+\tau$ is determined by its state at time t_0. Vice versa, changing the sign on the right hand side of (4.3) allows us to begin with the state at time $t_0+\tau$ and return to the state of time t_0: the system is reversible in an interval of time absent of rearragement events. This is no longer the situation after such an event. At the later time, we have no knowledge about which interval, now no longer in the inventory, was deleted.

As explained in [12–14], we can introduce now the idea of GBCD for the simplified 1D model. Let us consider a new ensemble based on the misorientation parameter α where we take Ω: $-\frac{\pi}{4} \leqq \alpha \leqq \frac{\pi}{4}$, for later ease of comparison with the two dimensional network for which we are imposing "cubic" symmetry, i.e., "square" symmetry in the plane. The *GBCD* or character distribution in this context is, as expected, the histogram of lengths of intervals sorted by misorientation α scaled to be a probability distribution on Ω. To be precise, we let

$$\begin{aligned} l_i(\alpha,t) &= x_{i+1}(t) - x_i(t) \\ &= \text{length of the } i\text{th interval, where explicit note has been taken of} \\ &\quad \text{its misorientation parameter } \alpha. \end{aligned}$$

Partition Ω into m subintervals of length $h = \frac{\pi}{2}\frac{1}{m}$ and define

$$\rho(\alpha,t) := \sum_{\alpha' \in ((k-1)h,kh]} l_i\left(\alpha',t\right)\cdot\frac{1}{Lh}, \quad \text{for } (k-1)h < \alpha \leqq kh. \tag{4.8}$$

For this definition of the statistic,

$$\int_{\Omega}\rho(\alpha,t)d\alpha = 1.$$

One may express (4.7) in terms of the character distribution (4.8), which amounts to

$$\mu_0\int_{t_0}^{t_0+\tau}\int_{\Omega}\left|\frac{\partial\rho}{\partial t}(\alpha,t)\right|^2 d\alpha dt + \int_{\Omega}\psi(\alpha)\rho(\alpha,t_0+\tau)d\alpha \leqq \int_{\Omega}\psi(\alpha)\rho(\alpha,t_0)d\alpha, \tag{4.9}$$

where $\mu_0 > 0$ is some constant.

The expression (4.9) is in terms of the new misorientation level ensemble, upscaled from the local level of the original system. We now introduce, as discussed earlier, the modeling assumption, consistent with the lack of reversibility when rearrangement/or critical events occur and add an entropic contribution to (4.9). We consider a standard configurational entropy,

$$+\int_{\Omega}\rho\log\rho d\alpha, \tag{4.10}$$

although this is not the only choice. Minimizing (4.10) favors the uniform state, which would be the situation were $\psi(\alpha) =$ constant. A tantalizing clue to the development of texture will be whether or not this entropy strays from its minimum during the simulation.

Given that (4.9) holds, we assume now that there is some $\lambda > 0$ such that for any t_0 and τ sufficiently small that

$$\mu_0 \int_{t_0}^{t_0+\tau} \int_\Omega \left(\frac{\partial \rho}{\partial t}\right)^2 d\alpha dt + \int_\Omega (\psi\rho + \lambda\rho \log \rho) d\alpha \Big|_{t_0+\tau} \leqq \int_\Omega (\psi\rho + \lambda\rho \log \rho) d\alpha \Big|_{t_0} . \tag{4.11}$$

$E(t)$ was analogous to an internal energy or the energy of a microcanonical ensemble and now

$$F(\rho) = F_\lambda(\rho) = E(t) + \lambda \int_\Omega \rho \log \rho d\alpha \tag{4.12}$$

is a free energy. The value of the parameter λ is unknown and will be determined in the Validation Sect. 5.

4.2 The Mass Transport Paradigm

The kinetics of the simplified problem will be understood by interpreting the dissipation principle for the GBCD in terms of a mass transport implicit scheme. In fact, (4.11) fails as a proper dissipation principle because the first term

$$\mu_0 \int_{t_0}^{t_0+\tau} \int_\Omega \left(\frac{\partial \rho}{\partial t}\right)^2 d\alpha dt \tag{4.13}$$

does not represent lost energy due to frictional or viscous forces. For a deformation path $f(\alpha, t)$, $0 \leqq t \leqq \tau$, of probability densities, this quantity is

$$\int_0^\tau \int_\Omega v^2 f d\alpha dt \tag{4.14}$$

where f, v are related by the continuity equation and initial and terminal conditions

$$\begin{aligned} &f_t + (vf)_\alpha = 0 \quad \text{in } \Omega \times (0, \tau), \quad \text{and} \\ &f(\alpha, 0) = \rho(\alpha, 0), \qquad f(\alpha, \tau) = \rho(\alpha, \tau), \end{aligned} \tag{4.15}$$

by analogy with fluids [47], p. 53 et seq., and elementary mechanics. (We have set $t_0 = 0$ for convenience.)

On the other hand, by a result of Benamou and Brenier [18], given two probability densities f^*, f on Ω, the Wasserstein distance $d(f, f^*)$ between them is given by

$$\begin{aligned}
&\frac{1}{\tau} d\big(f, f^*\big)^2 = \inf \int_0^\tau \int_\Omega v^2 f \, d\xi \, dt \\
&\text{over deformation paths } f(\xi, t) \text{ subject to} \\
&f_t + (vf)_\xi = 0 \quad \text{(continuity equation)}, \\
&f(\xi, 0) = f^*(\xi), \qquad f(\xi, \tau) = f(\xi) \text{ (initial and terminal conditions)}.
\end{aligned} \tag{4.16}$$

Let us briefly review the notion of Kantorovich-Rubinstein-Wasserstein metric, or simply Wasserstein metric. The reader can consult [3, 63] for more detailed exposition of the subject.

Let $D \subset \mathbf{R}$ be an interval, perhaps infinite, and f^*, f a pair of probability densities on D (with finite variance). The quadratic Wasserstein metric or 2-Wasserstein metric is defined to be

$$\begin{aligned}
&d\big(f, f^*\big)^2 = \inf_P \int_{D\times D} |x - y|^2 dp(x, y), \\
&P = \text{joint distributions for } f, f^* \text{ on } \bar{D} \times \bar{D},
\end{aligned} \tag{4.17}$$

i.e., the marginals of any $p \in P$ are f, f^*. The metric induces the weak-$*$ topology on $C(\bar{D})'$. If f, f^* are strictly positive, there is a transfer map which realizes p, essentially the solution of the Monge-Kantorovich mass transfer problem for this situation. This means that there is a strictly increasing

$$\begin{aligned}
&\phi : D \to D \quad \text{such that} \\
&\int_D \zeta(y) f(y) dy = \int_D \zeta\big(\phi(x)\big) f^*(x) dx, \quad \zeta \in C(\bar{D}), \quad \text{and} \\
&d\big(f, f^*\big)^2 = \int_D \big|x - \phi(x)\big|^2 f^* dx.
\end{aligned} \tag{4.18}$$

In this one dimensional situation, as was known to Frechét, [27],

$$\begin{aligned}
&\phi(x) = F^{*-1}\big(F(x)\big), \quad x \in D, \quad \text{where} \\
&F^*(x) = \int_{-\infty}^x f^*\big(x'\big) dx' \quad \text{and} \quad F(x) = \int_{-\infty}^x f\big(x'\big) dx'
\end{aligned} \tag{4.19}$$

are the distribution functions of f^*, f. In one dimension there is only one transfer map. The conditions (4.16) are in 'Eulerian' form. Likewise there is the 'Lagrangian' form which follows by rewriting (4.16) using the transfer function formu-

lation in (4.18),

$$\begin{aligned}&\frac{1}{\tau}d\big(f,f^*\big)^2=\inf\int_0^\tau\int_D\phi_t^2 f^*dx\\&\text{over transfer paths }\phi(x,t)\text{ from }D\text{ to }D\text{ with}\\&\phi(x,0)=x\quad\text{and}\quad\phi(x,\tau)=\phi(x).\end{aligned}\tag{4.20}$$

Therefore, our goal is to replace (4.13) with (4.14). Since the associated metrics induce different topologies, an estimate must involve additional terms. Assume that our statistic $\rho(\alpha,t)$ satisfies

$$\rho(\alpha,t)\geqq\delta>0\quad\text{in }\Omega,t>0.\tag{4.21}$$

This is a necessary assumption for our estimates below. In fact, to proceed with the implicit scheme introduced later, it is sufficient to require (4.21) just for the initial data $\rho_0(\alpha)$ since this property is inherited by the iterates. We now use the representation (4.16) and we use the deformation path given by ρ itself to calculate that for some $c_\Omega>0$,

$$\begin{aligned}&\frac{1}{\tau}d\big(\rho,\rho^*\big)^2\leqq\int_0^\tau\int_\Omega v^2\rho dxdt\leqq\frac{c_\Omega}{\min_\Omega\rho}\int_0^\tau\int_\Omega\frac{\partial\rho}{\partial t}(x,t)^2dxdt,\\&\rho^*(x)=\rho(x,0)\quad\text{and}\quad\rho(x)=\rho(x,\tau),\end{aligned}\tag{4.22}$$

where 0 represents an arbitrary starting time and τ a relaxation time.

Thus there is a $\mu>0$ such that for any relaxation time $\tau>0$,

$$\frac{\mu}{2}\int_0^\tau\int_\Omega v^2\rho d\alpha dt+F_\lambda(\rho)\leqq F_\lambda\big(\rho^*\big).\tag{4.23}$$

We next replace (4.23) by a minimum principle, arguing that the path given by $\rho(\alpha,t)$ is the one most likely to occur and the minimizing path has the highest probability. For this step, let $\rho^*=\rho(\cdot,t_0)$ and $\rho=\rho(\cdot,t+\tau)$. Observe that from (4.16),

$$\begin{aligned}&\frac{1}{\tau}d\big(\rho,\rho^*\big)^2=\inf\int_0^\tau\int_\Omega v^2 f d\alpha dt\\&\text{over deformation paths }f(\alpha,t)\text{ subject to}\\&f_t+(vf)_\alpha=0\quad\text{(continuity equation)},\\&f(\xi,0)=\rho^*(\alpha),\qquad f(\alpha,\tau)=\rho(\alpha,\tau)\quad\text{(initial and terminal conditions)},\end{aligned}\tag{4.24}$$

where d is the Wasserstein metric. So we may express the minimum principle in the form

$$\frac{\mu}{2\tau}d\big(\rho,\rho^*\big)^2+F_\lambda(\rho)=\inf_{\{\eta\}}\left\{\frac{\mu}{2\tau}d\big(\eta,\rho^*\big)^2+F_\lambda(\eta)\right\}.\tag{4.25}$$

For each relaxation time $\tau > 0$ we determine iteratively the sequence $\{\rho^{(k)}\}$ by choosing $\rho^* = \rho^{(k-1)}$ and $\rho^{(k)} = \rho$ in (4.25) and set

$$\rho^{(\tau)}(\alpha, t) = \rho^{(k)}(\alpha) \quad \text{in } \Omega \text{ for } k\tau \leqq t < (k+1)\tau. \tag{4.26}$$

We then anticipate recovering the GBCD ρ as

$$\rho(\alpha, t) = \lim_{\tau \to 0} \rho^{(\tau)}(\alpha, t), \tag{4.27}$$

with the limit taken in a suitable sense. It is known that ρ obtained from (4.27) is the solution of the Fokker-Planck Equation, [39],

$$\mu \frac{\partial \rho}{\partial t} = \frac{\partial}{\partial \alpha}\left(\lambda \frac{\partial \rho}{\partial \alpha} + \psi' \rho\right) \quad \text{in } \Omega, 0 < t < \infty. \tag{4.28}$$

We might point out here, as well, that a solution of (4.28) with periodic boundary conditions and nonnegative initial data is positive for $t > 0$.

5 Validation of the Scheme

We now begin the validation step of our model. The procedure which leads to the implicit scheme, based on the dissipation inequality (4.7), holds for the entire system but does not identify individual intermediate 'spring-mass-dashpots'. The consequence is that we cannot set the temperature-like parameter σ, but in some way must decide if one exists. Introduce the notation for the Boltzmann distribution with parameter λ

$$\rho_\lambda(\alpha) = \frac{1}{Z_\lambda} e^{-\frac{1}{\lambda}\psi(\alpha)}, \quad \alpha \in \Omega, \text{ with } Z_\lambda = \int_\Omega e^{-\frac{1}{\lambda}\psi(\alpha)} d\alpha. \tag{5.1}$$

With validation we would gain qualitative properties of solutions of (4.28):

- $\rho(\alpha, t) \to \rho_\sigma(\alpha)$ as $t \to \infty$, and
- this convergence is exponentially fast.

The Kullback-Leibler relative entropy for (4.28) is given by

$$\begin{aligned} &\Phi_\lambda(\eta) = \Phi(\eta \| \rho_\lambda) = \int_\Omega \eta \log \frac{\eta}{\rho_\lambda} d\alpha \quad \text{where} \\ &\eta \geqq 0 \quad \text{in } \Omega, \qquad \int_\Omega \eta d\alpha = 1, \end{aligned} \tag{5.2}$$

with ρ_λ from (5.1). By Jensen's Inequality it is always nonnegative. In terms of the free energy (4.12) and (5.1), (5.2) is given by

$$\Phi_\lambda(\eta) = \frac{1}{\lambda} F_\lambda(\eta) + \log Z_\lambda. \tag{5.3}$$

(Note: In our earlier work [13, 14], we defined relative entropy to be λ times (5.2).) A solution ρ of (4.28) has the property that

$$\Phi_\lambda(\rho) \to 0 \quad \text{as } t \to \infty. \tag{5.4}$$

Therefore, we seek to identify the particular $\lambda = \sigma$ for which Φ_σ defined by the GBCD statistic ρ tends monotonically to the minimum of all the $\{\Phi_\lambda\}$ as t becomes large. We then ask if the terminal, or equilibrium, empirical distribution ρ is equal to ρ_σ. Note that since

$$f(x, y) = x \log x - x \log y, \quad x, y > 0,$$

is convex, $\Phi(\eta \| \rho_\lambda)$ is a convex function of (η, ρ_λ). We assign a time $t = T_\infty$ and seek to minimize (5.2) at T_∞. With

$$\psi_\lambda = \frac{\psi}{\lambda} + \log Z_\lambda, \tag{5.5}$$

this minimization is a convex duality type of optimization problem, namely, to find the σ for which

$$\int_\Omega \{\psi_\sigma \rho + \rho \log \rho\} d\alpha = \inf_{\{\psi_\lambda\}} \int_\Omega \{\psi_\lambda \rho + \rho \log \rho\} d\alpha. \tag{5.6}$$

Note that

$$\int_\Omega e^{-\psi_\lambda} d\alpha = 1$$

which gives the minimization in (5.6) the form of finding an optimal prefix code, e.g. [57]. Here the potential ψ_λ, the code, is minimized in a family rather than the unknown density ρ itself, which is the given alphabet. For practical purposes, note that

$$\Phi(\rho \| \rho_\lambda) = \int_\Omega \rho \log \rho d\alpha + \frac{1}{\lambda} \int_\Omega \psi \rho d\alpha + \log Z_\lambda \tag{5.7}$$

is a strictly convex non-negative function of the 'inverse temperature' $\beta = \frac{1}{\lambda}$, $\beta > 0$, and thus admits a unique minimum.

The information theory interpretation is that we are minimizing the information loss among trial encodings of the alphabet represented by the statistic ρ. In this sense we see that asking for an optimal distribution ρ_σ to represent our statistic ρ, necessarily introduces (relative) entropy in our considerations, returning us, as it were, full circle.

From a given simulation, we harvest the GBCD statistic. It is a trial. The convexity of $\Phi(\rho \| \rho_\lambda)$ suggests that we can average trials. For trials $\{\rho_1, \ldots, \rho_N\}$,

$$\Phi\left(\frac{1}{N} \sum_{i=1,\ldots,N} \rho_i \| \rho_\lambda\right) \leqq \frac{1}{N} \sum_{i=1,\ldots,N} \Phi(\rho_i \| \rho_\lambda). \tag{5.8}$$

So we can seek the optimal $\lambda = \sigma$ by optimizing with the averaged trial. We shall illustrate this for the validation process for the two dimensional simulation.

5.1 An Example of the Simplified Problem

For the simplified coarsening model, we consider

$$\psi(\alpha) = 1 + 2\alpha^2 \quad \text{in } \Omega = \left(-\frac{\pi}{4}, \frac{\pi}{4}\right), \tag{5.9}$$

and shall identify a unique such parameter, which we label σ, by seeking the minimum of the relative entropy (5.2), namely by inspection of plots of (5.6) and (5.7), and then comparing ρ with the found ρ_σ. This ψ the development to second order of $\psi(\alpha) = 1 + 0.5 \sin^2 2\alpha$ used in the 2D simulation. Moreover, since the potential is quadratic, it represents a version of the Ornstein-Uhlenbeck process. We agree that $T_\infty = T(80\ \%) = 6.73$ represents time equals infinity. This is the time at which 80 % of the segments have been deleted and corresponds to the stationary configuration in the two-dimensional simulation. For the simplified critical event model we are considering, it is clear that by computing for a sufficiently long time, all cells will be gone. This time may be quite long. For comparison, $T(90\ \%) = 30$ and $T(95\ \%) = 103$. There may be additional criteria for choosing a T in the neighborhood of $T(80\ \%)$ and we may wish to discuss this later. The results are reported in Fig. 4.

6 The Entropy Method for the GBCD

6.1 Quadratic Interfacial Energy Density

We shall apply the method of Sect. 5 to the GBCD harvested from the 2D simulation. We consider first a typical simulation with the energy density

$$\psi(\alpha) = 1 + \epsilon(\sin 2\alpha)^2, \quad -\frac{\pi}{4} \leqq \alpha \leqq \frac{\pi}{4}, \ \epsilon = 1/2. \tag{6.1}$$

Figure 5, initialized with 10^4 cells and normally distributed misorientation angles and terminated when 2000 cells remain. At this stage, the simulation is essentially stagnant. Five trials were executed and we consider the average of ρ of the empirical GBCD's. Possible 'temperature' parameters λ and ρ_λ in (5.1) for the density (6.1) are constructed. This ρ_λ then defines a trial relative entropy via (5.2). We now identify the parameter σ, which turns out to be $\sigma \approx 0.1$, and the value of the relative entropy $\Phi_\sigma(T_\infty) \approx 0.01$, which is about 10 % of its initial value, Fig. 6. From Fig. 7 (top), we see that this relative entropy Φ_σ has exponential decay until it reaches

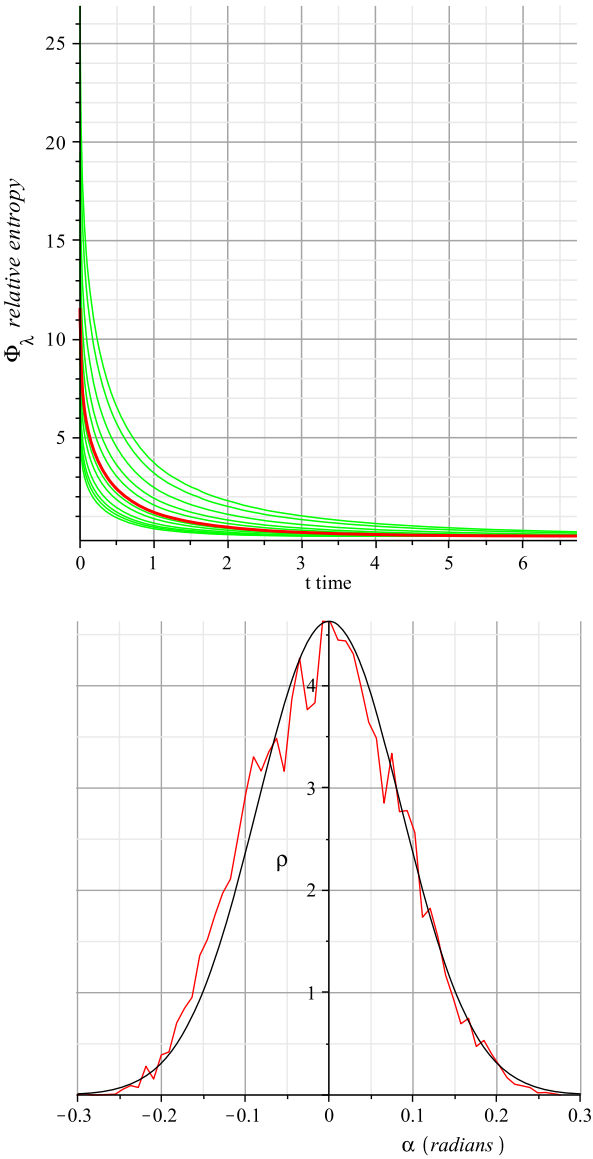


Fig. 4 Graphical results for the simplified coarsening model. (*Top*) Relative entropy plots for selected values of λ with Φ_σ noted in *red*. The value of $\sigma = 0.0296915$. (*Bottom*) Empirical distribution at time $t = T = T_\infty$ in *red* compared with ρ_σ in *black* (Color figure online)

time about $t = 1.5$, after which it remains constant. The averaged empirical GBCD is compared with the Boltzmann distribution in Fig. 7 (bottom). The solution itself then tends exponentially in L^1 to its limit ρ_σ by the Kullback-Leibler Inequality.

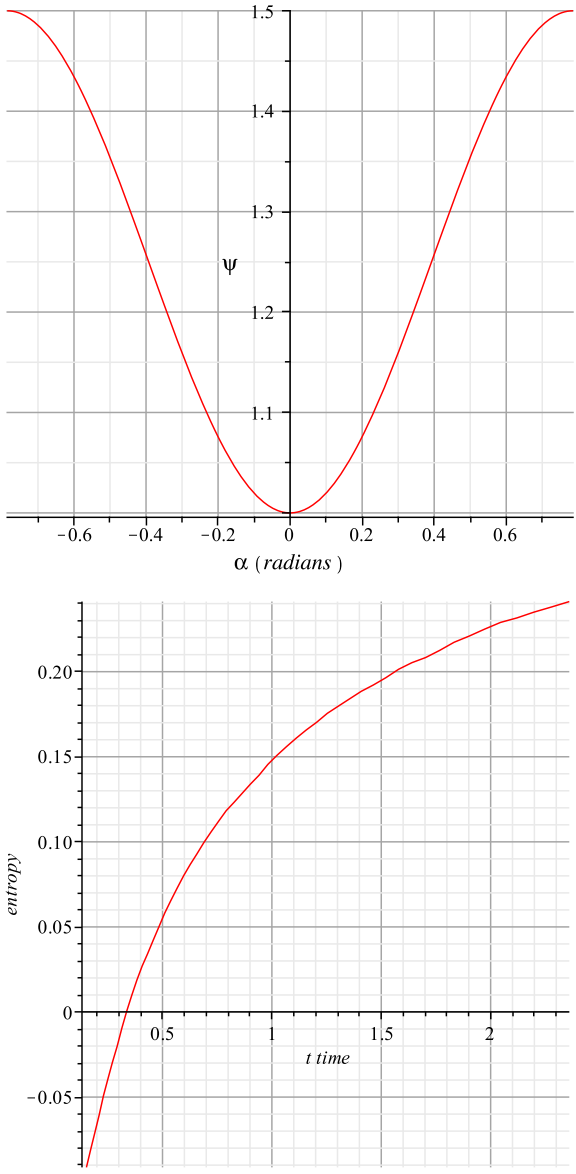


Fig. 5 (*Top*) The energy density $\psi(\alpha) = 1 + \epsilon \sin^2 2\alpha$, $|\alpha| < \pi/4$, $\epsilon = \frac{1}{2}$. (*Bottom*) The entropy of $\rho(\alpha, t)$ as a function of time t is increasing, suggesting the development of order in the configuration

6.2 Quartic Interfacial Energy Density

Our second example is a quartic energy, Fig. 8,

$$\psi(\alpha) = 1 + \epsilon(\sin 2\alpha)^4, \quad -\frac{\pi}{4} \leqq \alpha \leqq \frac{\pi}{4}, \ \epsilon = 1/2. \tag{6.2}$$

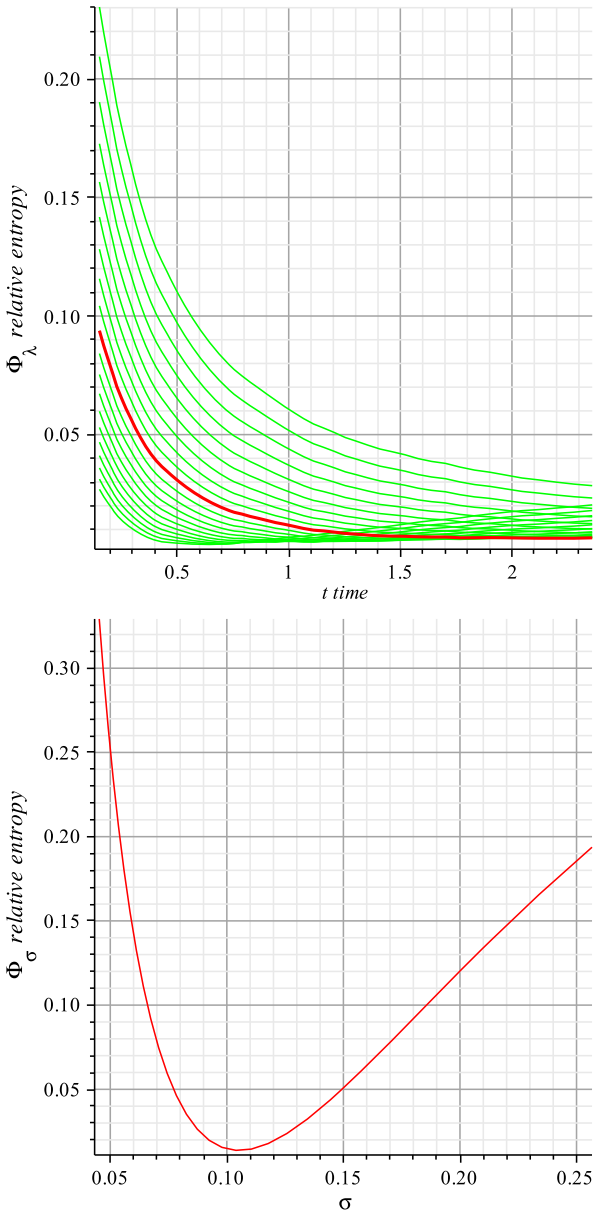


Fig. 6 In these plots, the GBCD ρ is averaged over 5 trials. (*Top*) The relative entropy of the grain growth simulation with energy density (6.1) for a sequence of Φ_λ vs. t with the optimal choice $\sigma \approx 0.1$ noted in *red*. (*Bottom*) Relative entropy for an indicated range of values of temperature parameter λ at the terminal time $t = T_\infty = 2.3$. The minimum value of the relative entropy is ≈ 0.01 (Color figure online)

Again, a configuration of 10^4 cells is initialized with normally distributed misorientations and, this time, the computation proceeds until about 1000 cells remain. The relative entropy and the equilibrium Boltzmann statistic stabilize when 2000 cells remain. Seven trials were executed and we consider the average of ρ of seven empirical GBCD's. Results are summarized in Fig. 9 and Fig. 10.

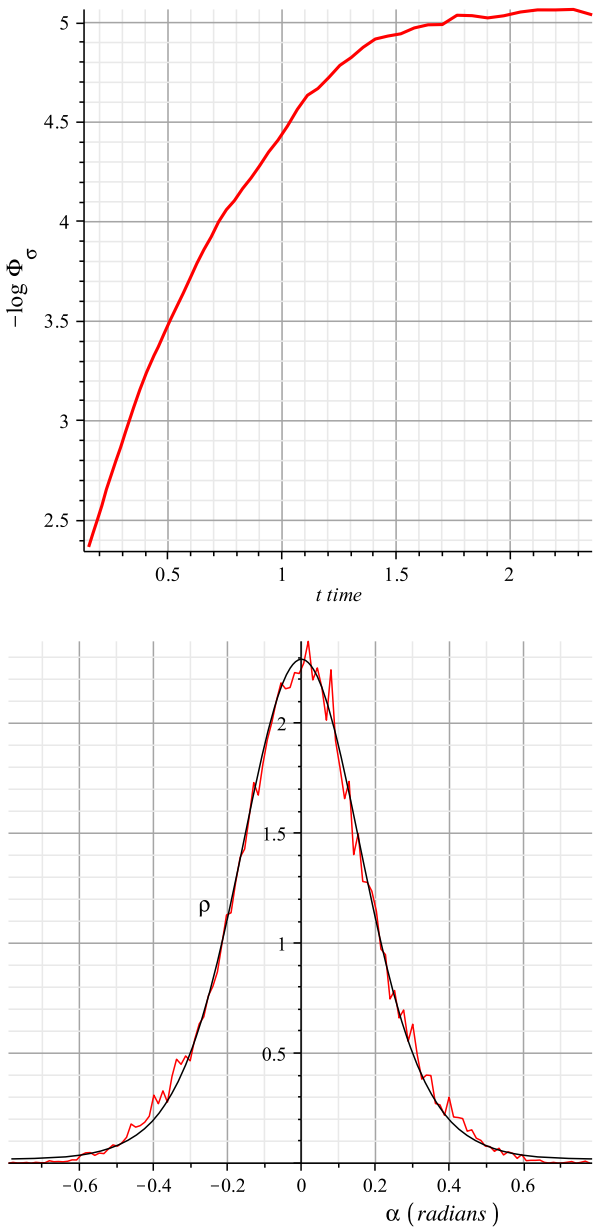


Fig. 7 In these plots, the GBCD is averaged over 5 trials. (*Top*) Plot of $-\log \Phi_\sigma$ vs. t with energy density (6.1). It is approximately linear until it becomes constant showing that Φ_σ decays exponentially. (*Bottom*) GBCD ρ (*red*) and Boltzmann distribution ρ_σ (*black*) for the potential ψ of (6.1) with parameter $\sigma \approx 0.1$ as predicted by our theory (Color figure online)

6.3 Remarks on a Theory for the Diffusion Coefficient σ or the Temperature-Like Parameter

The network level nonequilibrium nature of the iterative scheme introduced in our theory Sects. 4–5, leaves free a temperature-like parameter σ. However, as we

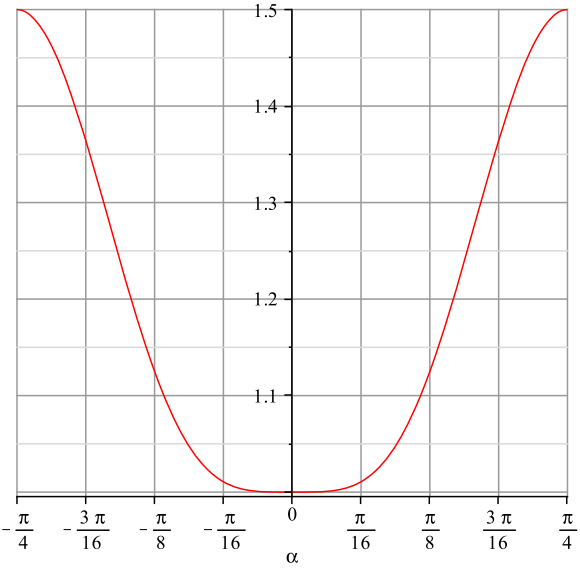


Fig. 8 The energy density $\psi(\alpha) = 1 + \epsilon \sin^4 2\alpha$, $|\alpha| < \pi/4$, $\epsilon = \frac{1}{2}$

showed in Sect. 5, we can uniquely identify σ. But can we a priori determine or control this temperature-like parameter? There are different approaches to this question, none of which have been especially successful at this point. One possible approach is to consider a different theory that is developed for the simplified model based on the kinetic equations description in [11]. However, this particular description [11] would have to be improved, since it does not produce a very good result for σ at this point. However, this method would still have only an empirical flavor: the value of σ will be obtained once the solution of kinetic equations is computed. Another direction to consider here is based on the statistical analysis of the data obtained from many trials and to understand the possible connection to branching processes.

7 Closing Comments

Engineering the microstructure of a material is a central task of materials science and its study gives rise to a broad range of basic science issues, as has been long recognized. Central to these issues is the coarsening of the cellular structure. Here we have outlined an entropy based theory of the GBCD which is an upscaling of cell growth according to the two most basic properties of a coarsening network: a local evolution law and space filling contraints. The theory accommodates the irreversibility conferred by the critical events or topological rearrangements which arise during coarsening. It adds to the body of evidence that the evolution of the boundary network is the primary origin of texture development. It accounts both for the GBCD and its kinetics.

There are many known environments where the kinetics of growth do not seem to follow this sort of pattern. Let us briefly consider one, stagnation in the evolution of

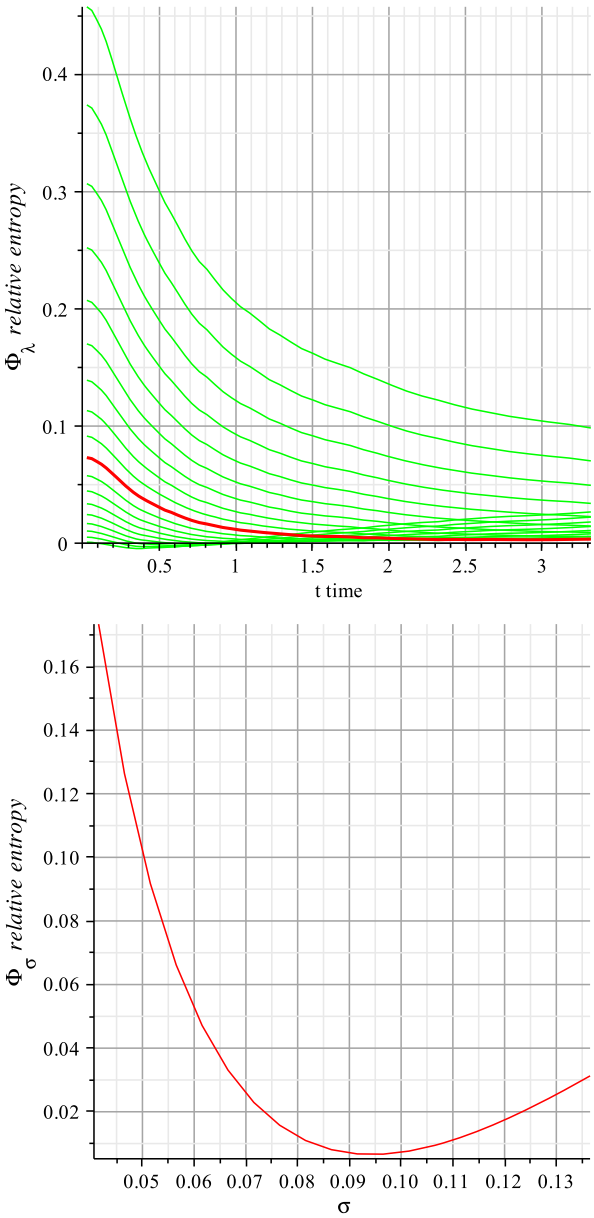


Fig. 9 In these plots, the GBCD ρ is averaged over 7 trials. (*Top*) The relative entropy of the grain growth simulation with energy density (6.2) for a sequence of Φ_λ vs. t with the optimal choice $\sigma \approx 0.095$ noted in *red*. (*Bottom*) Relative entropy for an indicated range of values of temperature parameter σ at the terminal time $t = T_\infty = 3$. The minimum value of the relative entropy is ≈ 0.007 (Color figure online)

metallic (Cu and Al) thin films, important for the metallization of semiconductors, [16, 17]. Stagnation means that the growth process appears to stop even though the material remains in the furnace. Some progress is found in [36]. A striking feature of these films is a nearly exact log-normal distribution of the relative grain diameters based on a study of 27 samples consisting of 35,000 grains prepared in different

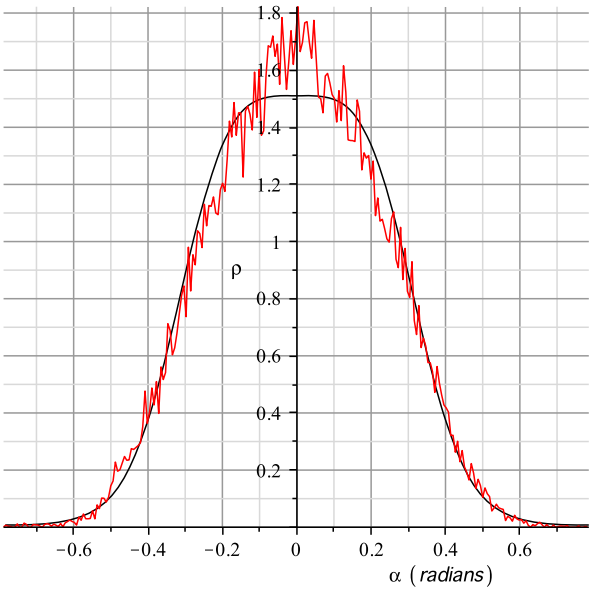


Fig. 10 Comparison of the empirical distribution at time $t = T_\infty = 3$, when 80 % of the cells have been deleted, with ρ_σ, the Boltzmann distribution of (5.1), with σ extracted from Fig. 9. The GBCD ρ is averaged over 7 trials

experiments in a wide variety of conditions. The grain diameter is, basically, the square root of its area. This distribution is not found in any simulation of coarsening known to us. One possible starting point for an investigation is the well known Kolmogorov "rock crushing" problem, which has a representation as a scaled branching process.

The stagnation isssue is, of course, just a hint of the variety of challenges we encounter in this exciting field.

Acknowledgements Much of this research was done while E. Eggeling, Y. Epshteyn and R. Sharp were postdoctoral associates at the Center for Nonlinear Analysis at Carnegie Mellon University. We are grateful to our colleagues G. Rohrer, A.D. Rollett, R. Schwab, and R. Suter for their collaboration.

References

1. Adams, B.L., Kinderlehrer, D., Mullins, W.W., Rollett, A.D., Ta'asan, S.: Extracting the relative grain boundary free energy and mobility functions from the geometry of microstructures. Scr. Mater. **38**(4), 531–536 (Jan 13 1998)
2. Adams, B.L., Kinderlehrer, D., Livshits, I., Mason, D., Mullins, W.W., Rohrer, G.S., Rollett, A.D., Saylor, D., Ta'asan, S., Wu, C.: Extracting grain boundary energy from triple junction measurement. Interface Sci. **7**, 321–338 (1999)
3. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows in Metric Spaces and in the Space of Probability Measures, 2nd edn. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel (2008)
4. Arbogast, T.: Implementation of a locally conservative numerical subgrid upscaling scheme for two-phase Darcy flow. Comput. Geosci. **6**(3–4), 453–481 (2002). Locally conservative numerical methods for flow in porous media

5. Arbogast, T., Lehr, H.L.: Homogenization of a Darcy-Stokes system modeling vuggy porous media. Comput. Geosci. **10**(3), 291–302 (2006)
6. Balhoff, M.T., Thomas, S.G., Wheeler, M.F.: Mortar coupling and upscaling of pore-scale models. Comput. Geosci. **12**(1), 15–27 (2008)
7. Balhoff, M., Mikelić, A., Wheeler, M.F.: Polynomial filtration laws for low Reynolds number flows through porous media. Transp. Porous Media **81**(1), 35–60 (2010)
8. Barmak, K.: unpublished
9. Barmak, K., Emelianenko, M., Golovaty, D., Kinderlehrer, D., Ta'asan, S.: On a statistical theory of critical events in microstructural evolution. In Proceedings CMDS 11, pp. 185–194. ENSMP Press, Paris (2007)
10. Barmak, K., Emelianenko, M., Golovaty, D., Kinderlehrer, D., Ta'asan, S.: Towards a statistical theory of texture evolution in polycrystals. SIAM J. Sci. Comput. **30**(6), 3150–3169 (2007)
11. Barmak, K., Emelianenko, M., Golovaty, D., Kinderlehrer, D., Ta'asan, S.: A new perspective on texture evolution. Int. J. Numer. Anal. Model. **5**(Sp. Iss. SI), 93–108 (2008)
12. Barmak, K., Eggeling, E., Emelianenko, M., Epshteyn, Y., Kinderlehrer, D., Ta'asan, S.: Geometric growth and character development in large metastable systems. Rend. Mat, Ser. VII **29**, 65–81 (2009)
13. Barmak, K., Eggeling, E., Emelianenko, M., Epshteyn, Y., Kinderlehrer, D., Sharp, R., Ta'asan, S.: Critical events, entropy, and the grain boundary character distribution. Phys. Rev. B **83**(13), 134117 (Apr 2011)
14. Barmak, K., Eggeling, E., Emelianenko, M., Epshteyn, Y., Kinderlehrer, D., Sharp, R., Ta'asan, S.: An entropy based theory of the grain boundary character distribution. Discrete Contin. Dyn. Syst. **30**(2), 427–454 (2011)
15. Barmak, K., Eggeling, E., Emelianenko, M., Epshteyn, Y., Kinderlehrer, D., Sharp, R., Ta'asan, S.: Predictive theory for the grain boundary character distribution. In: Materials Science Forum, vols. 715–716, pp. 279–285. Trans Tech Publications, Durnten-Zurich (2012)
16. Barmak, K., Eggeling, E., Sharp, R., Roberts, S., Shyu, T., Sun, T., Yao, B., Ta'asan, S., Kinderlehrer, D., Rollett, A., Coffey, K.: Grain growth and the puzzle of its stagnation in thin films: A detailed comparison of experiments and simulations. In: Materials Science Forum, vols. 715–716, pp. 473–479. Trans Tech Publications, Durnten-Zurich (2012)
17. Barmak, K., Eggeling, E., Sharp, R., Ta'asan, S., Kinderlehrer, D., Rollett, A., Coffey, K.: Grain growth and the puzzle of its stagnation in thin films: a detailed comparison of experiments and simulations (2012, submitted), cf. Center for Nonlinear Analysis, Carnegie Mellon University, preprint 10-CNA-012
18. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer. Math. **84**(3), 375–393 (2000)
19. Bertotti, G.: Hysteresis in Magnetism. Academic Press, San Diego (1998)
20. Bronsard, L., Reitich, F.: On three-phase boundary motion and the singular limit of a vector-valued Ginzburg-Landau equation. Arch. Ration. Mech. Anal. **124**(4), 355–379 (1993)
21. Burke, J.E., Turnbull, D.: Recrystallization and grain growth. Prog. Met. Phys. **3**(C), 220–244 (1952). IN11–IN12, 245–266, IN13–IN14, 267–274, IN15, 275–292
22. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. Studies in Mathematics and Its Applications, vol. 4. North-Holland, Amsterdam (1978)
23. DeSimone, A., Kohn, R.V., Müller, S., Otto, F., Schäfer, R.: Two-dimensional modelling of soft ferromagnetic films. R. Soc. Lond. Proc., Ser. A, Math. Phys. Eng. Sci. **457**(2016), 2983–2991 (2001)
24. Elsey, M., Esedoḡlu, S., Smereka, P.: Diffusion generated motion for grain growth in two and three dimensions. J. Comput. Phys. **228**(21), 8015–8033 (2009)
25. Epshteyn, Y., Rivière, B.: On the solution of incompressible two-phase flow by a p-version discontinuous Galerkin method. Commun. Numer. Methods Eng. **22**, 741–751 (2006)
26. Epshteyn, Y., Rivière, B.: Fully implicit discontinuous finite element methods for two-phase flow. Appl. Numer. Math. **57**, 383–401 (2007)

27. Fréchet, M.: Sur la distance de deux lois de probabilite. C. R. Acad. Sci., Sér. 1 Math. **244**(6), 689–692 (1957)
28. Garcke, H., Nestler, B., Stoth, B.: A multiphase field concept: numerical simulations of moving phase boundaries and multiple junctions. SIAM J. Appl. Math. **60**(1), 295–315 (2000) (electronic)
29. Godunov, S.K.: A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. Mat. Sb. (N.S.) **47**(89), 271–306 (1959)
30. Godunov, S.K., Ryaben'kii, V.S.: Difference Schemes. Studies in Mathematics and Its Applications, vol. 19. North-Holland, Amsterdam (1987). An introduction to the underlying theory, translated from the Russian by E.M. Gelbard
31. Gomer, R., Smith, C.S. (eds.): Structure and Properties of Solid Surfaces. The University of Chicago Press, Chicago (1952). Proceedings of a conference arranged by the National Research Council and held in September 1952, in Lake Geneva, Wisconsin, USA
32. Gurtin, M.: Thermomechanics of Evolving Phase Boundaries in the Plane. Clarendon, Oxford (1993)
33. Helmig, R.: Multiphase Flow and Transport Processes in the Subsurface. Springer, Berlin (1997)
34. Herring, C.: Surface tension as a motivation for sintering. In: Kingston, W.E. (ed.) The Physics of Powder Metallurgy, pp. 143–179. McGraw-Hill, New York (1951)
35. Herring, C.: The use of classical macroscopic concepts in surface energy problems. In: Gomer, R., Smith, C.S. (eds.) Proceedings of a Conference Arranged by the National Research Council and Held in September 1952, in Lake Geneva, Wisconsin, USA, pp. 5–81 (1952)
36. Holm, E.A., Foiles, S.M.: Grain growth stagnation caused by the grain boundary roughening transition. In: Materials Science Forum, vols. 715–716, p. 415. Trans Tech Publications, Durnten-Zurich (2012)
37. Iserles, A.: A First Course in the Numerical Analysis of Differential Equations. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge (1996)
38. Jordan, R., Kinderlehrer, D., Otto, F.: Free energy and the Fokker-Planck equation. Physica D **107**(2–4), 265–271 (Sep 1, 1997)
39. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker-Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (Jan 1998)
40. Kim, S.G., Kim, D.I., Kim, W.T., Park, Y.B.: Computer simulations of two-dimensional and three-dimensional ideal grain growth. Phys. Rev. E **74**, 061605 (2006)
41. Kinderlehrer, D., Liu, C.: Evolution of grain boundaries. Math. Models Methods Appl. Sci. **11**(4), 713–729 (Jun 2001)
42. Kinderlehrer, D., Lee, J., Livshits, I., Rollett, A., Ta'asan, S.: Mesoscale simulation of grain growth. In: Recrystallization and Grain Growth, Pts 1 and 2, vol. 467–470, pp. 1057–1062 (2004)
43. Kinderlehrer, D., Livshits, I., Rohrer, G.S., Ta'asan, S., Yu, P.: Mesoscale simulation of the evolution of the grain boundary character distribution. In: Recrystallization and Grain Growth, Pts 1 and 2, vols. 467–470, pp. 1063–1068 (2004)
44. Kinderlehrer, D., Livshits, I., Ta'asan, S.: A variational approach to modeling and simulation of grain growth. SIAM J. Sci. Comput. **28**(5), 1694–1715 (2006)
45. Kohn, R.V.: Irreversibility and the statistics of grain boundaries. Physics **4**, 33 (Apr 2011)
46. Kohn, R.V., Otto, F.: Upper bounds on coarsening rates. Commun. Math. Phys. **229**(3), 375–395 (2002)
47. Landau, L.D., Lifshitz, E.M.: Fluid mechanics. In: Course of Theoretical Physics, vol. 6. Pergamon Press, London (1959). Translated from the Russian by J.B. Sykes and W.H. Reid
48. Lax, P.D.: Weak solutions of nonlinear hyperbolic equations and their numerical computation. Commun. Pure Appl. Math. **7**, 159–193 (1954)
49. Lax, P.D.: Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 11. Society for Industrial and Applied Mathematics, Philadelphia (1973)

50. Li, B., Lowengrub, J., Rätz, A., Voigt, A.: Geometric evolution laws for thin crystalline films: modeling and numerics. Commun. Comput. Phys. **6**(3), 433–482 (2009)
51. Lifshitz, I.M., Slyozov, V.V.: The kinetics of precipitation from supersaturated solid solutions. J. Phys. Chem. Solids **19**(1–2), 35–50 (1961)
52. Lowengrub, J.S., Rätz, A., Voigt, A.: Phase-field modeling of the dynamics of multicomponent vesicles: spinodal decomposition, coarsening, budding, and fission. Phys. Rev. E **79**(3), 031926 (2009), 13 pages
53. Mullins, W.W.: 2-Dimensional motion of idealized grain growth. J. Appl. Phys., **27**(8), 900–904 (1956)
54. Mullins, W.W.: Solid Surface Morphologies Governed by Capillarity, pp. 17–66. American Society for Metals, Metals Park (1963)
55. Mullins, W.W.: On idealized 2-dimensional grain growth. Scr. Metall. **22**(9), 1441–1444 (Sep 1988)
56. Otto, F., Rump, T., Slepčev, D.: Coarsening rates for a droplet model: rigorous upper bounds. SIAM J. Math. Anal. **38**(2), 503–529 (2006)
57. Rissanen, J.: Complexity and information in data. In: Entropy. Princeton Ser. Appl. Math., pp. 299–312. Princeton University Press, Princeton (2003)
58. Rohrer, G.S.: Influence of interface anisotropy on grain growth and coarsening. Annu. Rev. Mater. Res. **35**, 99–126 (2005)
59. Rollett, A.D., Lee, S.-B., Campman, R., Rohrer, G.S.: Three-dimensional characterization of microstructure by electron back-scatter diffraction. Annu. Rev. Mater. Res. **37**, 627–658 (2007)
60. Smith, C.S.: Grain shapes and other metallurgical applications of topology. In: Gomer, R., Smith, C.S. (eds.) Proceedings of a Conference Arranged by the National Research Council and Held in September 1952, in Lake Geneva, Wisconsin, USA, pp. 65–108 (1952)
61. Stewart, H.B., Wendroff, B.: Two-phase flow: models and methods. J. Comput. Phys. **56**(3), 363–409 (1984)
62. Toselli, A., Widlund, O.: Domain Decomposition Methods—Algorithms and Theory. Springer Series in Computational Mathematics, vol. 34. Springer, Berlin (2005)
63. Villani, C.: Topics in Optimal Transportation. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence (2003)
64. Von Neumann, J., Richtmyer, R.D.: A method for the numerical calculation of hydrodynamic shocks. J. Appl. Phys. **21**, 232–237 (1950)
65. Von Neumann, J., Discussion remark concerning paper of C.S. Smith "Grain shapes and other metallurgical applications of topology". In: Gomer, R., Smith, C.S. (eds.) Proceedings of a Conference Arranged by the National Research Council and Held in September, 1952, in Lake Geneva, Wisconsin, USA, pp. 108–110 (1952)
66. Wagner, C.: Theorie der Alterung von Niederschlagen durch Umlosen (Ostwald-reifung). Z. Elektrochem. **65**(7–8), 581–591 (1961)

A Generalized Empirical Interpolation Method: Application of Reduced Basis Techniques to Data Assimilation

Yvon Maday and Olga Mula

Abstract This paper, written as a tribute to Enrico Magenes, a giant that has kindly and warmly supported generations of young researchers, introduces a generalization of the empirical interpolation method (EIM) and the reduced basis method (RBM) in order to allow their combination with data mining and data assimilation. The purpose is to be able to derive sound information from data and reconstruct information, possibly taking into account noise in the acquisition, that can serve as an input to models expressed by partial differential equations. The approach combines data acquisition (with noise) with domain decomposition techniques and reduced basis approximations.

1 Introduction

The representation of some physical or mechanical quantities, representing a scalar or vectorial function that depends on space, time or both, can be elaborated through at least two—possibly—complementary approaches: the first one, called explicit hereafter, is based on the measurement of some instances of the quantity of interest that consists in getting its value at some points from which, by interpolation or extrapolation, the quantity is approximated in other points than where the measurements have been performed. The second approach, called implicit hereafter, is more elaborated. It is based on a model, constructed by expertise, that implicitly characterizes the quantity as a solution to some problem fed with input data. The model

In memory of Enrico Magenes.

Y. Maday (✉) · O. Mula
Laboratoire Jacques-Louis Lions, UMR 7598, UPMC Univ Paris 06, 75005, Paris, France
e-mail: maday@ann.jussieu.fr

Y. Maday
Division of Applied Mathematics, Institut Universitaire de France and Brown University, Providence, RI, USA

O. Mula
CEA Saclay—DEN/DANS/DM2S/SERMA/LLPR, 91191 Gif-Sur-Yvette Cedex, France
e-mail: olga.mulahernandez@cea.fr

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*,
Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_13,

can e.g. be a parameter dependent partial differential equation, the simulation of which allows to get an approximation of the quantity of interest, and, actually, many more outputs than the sole value of the quantity of interest. This second approach, when available, is more attractive since it allows to have a better understanding of the working behavior of the phenomenon that is under consideration. In turn, it facilitates optimization, control or decision making.

Nevertheless for still a large number of problems, the numerical simulation of this model is indeed possible—though far too expensive to be performed in a reasonable enough time. The combined efforts of numerical analysts, specialists of algorithms and computer scientists, together with the increase of the performances of the computers allow to increase every days the domains of application where numerical simulation can be used, to such an extent that it is possible now to rigorously adapt the approximation, degrade the models, degrade the simulation, or both in an intelligent way without sacrificing the quality of the approximation where it is required.

Among the various ways to reduce the problem's complexity stand approaches that use the smallness of the Kolmogorov n-width [5] of the manifold of all solutions considered when the parameters varies continuously in some range. This idea, combined with the Galerkin method is at the basis of the reduced basis method and the Proper Orthogonal Decomposition (POD) methods to solve parameter dependent partial differential equations. These approximation methods allow to build the solution to the model associated to some parameter as a linear combination of some precomputed solutions associated to some well chosen parameters. The precomputations can be lengthy but are performed off-line, the online computation has a very small complexity, based on the smallness of the Kolmogorov n-width. We refer to [9, 10] for an introduction to these approaches.

Another possibility, rooted on the same idea, is the empirical interpolation method (EIM) that allows, from values of the quantity at some interpolating points, to build a linear combination of again preliminary fully determined quantities associated to few well chosen instances of the parameter. The linear combination is determined in such a way that it takes the same values at the interpolating points as the quantity we want to represent. This concept generalizes the classical—e.g. polynomial or radial basis—interpolation procedure and is recalled in the next section. The main difference is that the interpolating function may even, a priori, not be known but depend on the quantity we want to represent.

In this paper we first aim at generalizing further this EIM concept by replacing the pointwise evaluations of the quantity by more general measures, mathematically defined as linear forms defined on a superspace of the manifold of appropriate functions. We consider that this generalization, named Generalized Empirical Interpolation Method (GEIM), represents already an improvement with respect to classical interpolation reconstructions.

Bouncing on this GEIM, we propose a coupled approach based on the domain decomposition of the computational domain into two parts: one small domain Ω_1 where the Kolmogorov n-width of the manifold is not small and where the parametrized PDE will be simulated and the other subdomain Ω_2, much larger

but with a small Kolmogorov n-width because for instance the solution is driven over Ω_2 by the behavior of the solution over Ω_1. The idea is then to first construct (an approximation of) the solution from the measurements using the GEIM. In turn this reconstruction, up to the interface between Ω_1 and Ω_2, provides the necessary boundary conditions for solving the model over Ω_1.

This is not the first attempt to use the small Kolmogorov n-width for another aim than the POD or reduced basis technique which are both based on a Galerkin approach. In [2] e.g. the smallness of the Kolmogorov width is used to post-process a coarse finite element approximation and get an improved accuracy.

The problems we want to address with this coupled approach, stem from, e.g., actual industrial process or operations that work on a day-to-day basis; they can be observed with experimental sensors that provide sound data and are able to characterize part of their working behavior. We think that the numerical simulation and data mining approaches for analyzing real life systems are not enough merged in order to (i) complement their strength and (ii) cope for their weaknesses. This paper is a contribution in this direction.

In the last section, we evoke the problem of uncertainty and noises in the acquisition of the data, since indeed, the data are most often polluted by noises. Due to this, statistical data acquisition methods are used to filter out the source signals so that an improved knowledge is accessible. In many cases though, and this is more and more the case now, the data are far too numerous to all be taken into account, most of them are thus neglected because people do not know how to analyze them, in particular when the measures that are recorded are not directly related to some directly understandable quantity.

2 Generalized Empirical Interpolation Method

The rationale of all our approach relies on the possibility to approximately represent a given set, portion of a regular manifold (here the set of solution to some PDE), as a linear combination of very few computable elements. This is linked to the notion of n-width following Kolmogorov [5]:

Definition 2.1 Let F be a subset of some Banach space $\mathcal{X}$ and Y_n be a generic n-dimensional subspace of $\mathcal{X}$. The angle between F and Y_n is

$$E(F; Y_n) := \sup_{x \in F} \inf_{y \in Y_n} \|x - y\|_{\mathcal{X}}.$$

The *Kolmogorov n-width* of F in $\mathcal{X}$ is given by

$$\begin{aligned} d_n(F, \mathcal{X}) &:= \inf\big\{E(F; Y_n) : Y_n \text{ a } n\text{-dimensional subspace of } \mathcal{X}\big\} \\ &= \inf_{Y_n} \sup_{x \in F} \inf_{y \in Y_n} \|x - y\|_{\mathcal{X}}. \end{aligned} \tag{1}$$

The n-width of F thus measures to what extent the set F can be approximated by an n-dimensional subspace of $\mathcal{X}$.

We assume from now on that F and $\mathcal{X}$ are composed of functions defined over a domain $\Omega \subset \mathbb{R}^d$, where $d = 1, 2, 3$ and that F is a compact set of $\mathcal{X}$.

2.1 Recall of the Empirical Interpolation Method

We begin by describing the construction of the empirical interpolation method [1, 3, 6] that allows us to define simultaneously the set of generating functions recursively chosen in F together with the associated interpolation points. It is based on a greedy selection procedure as outlined in [7, 10, 12]. With $\mathcal{M}$ being some given large number, we assume that the dimension of the vectorial space spanned by F: span(F) is of dimension $\geq \mathcal{M}$.

The first generating function is $\varphi_1 = \arg\max_{\varphi \in F} \|\varphi(\cdot)\|_{L^\infty(\Omega)}$, the associated interpolation point satisfies $x_1 = \arg\max_{x \in \overline{\Omega}} |\varphi_1(x)|$, we then set $q_1 = \varphi_1(\cdot)/\varphi_1(x_1)$ and $B^1_{11} = 1$. We now construct, by induction, the nested sets of interpolation points $\Xi_M = \{x_1, \ldots, x_M\}$, $1 \leq M \leq M_{\max}$, and the nested sets of basis functions $\{q_1, \ldots, q_M\}$, where $M_{\max} \leq \mathcal{M}$ is some given upper bound fixed *a priori*. For $M = 2, \ldots, M_{\max}$, we first solve the interpolation problem: Find

$$\mathcal{I}_{M-1}\big[\varphi(\cdot)\big] = \sum_{j=1}^{M-1} \alpha_{M-1,j}[\varphi] q_j, \tag{2}$$

such that

$$\mathcal{I}_{M-1}\big[\varphi(\cdot)\big](x_i) = \varphi(x_i), \quad i = 1, \ldots, M-1, \tag{3}$$

that allows to define the $\alpha_{M-1,j}[\varphi]$, $1 \leq j \leq M$, as it can be proven indeed that the $(M-1) \times (M-1)$ matrix of running entry $q_j(x_i)$ is invertible, actually it is lower triangular with unity diagonal.

We then set

$$\forall \varphi \in F, \quad \varepsilon_{M-1}(\varphi) = \big\|\varphi - \mathcal{I}_{M-1}[\varphi]\big\|_{L^\infty(\Omega)}, \tag{4}$$

and define

$$\varphi_M = \arg\max_{\varphi \in F} \varepsilon_{M-1}(\varphi), \tag{5}$$

and

$$x_M = \arg\max_{x \in \overline{\Omega}} \big|\varphi_M(x) - \mathcal{I}_{M-1}[\varphi_M](x)\big|, \tag{6}$$

we finally set $r_M(x) = \varphi_M(x) - \mathcal{I}_{M-1}[\varphi_M(x)]$, $q_M = r_M / r_M(x_M)$ and $B^M_{ij} = q_j(x_i)$, $1 \leq i, j \leq M$.

The Lagrangian functions—that can be used to build the interpolation operator $\mathcal{I}_M$ in $X_M = \text{span}\{\varphi_i, 1 \le i \le M\} = \text{span}\{q_i, 1 \le i \le M\}$ over the set of points $\Xi_M = \{x_i, 1 \le i \le M\}$—verify for any given M, $\mathcal{I}_M[u(\cdot)] = \sum_{i=1}^M u(x_i) h_i^M(\cdot)$, where $h_i^M(\cdot) = \sum_{j=1}^M q_j(\cdot)[B^M]_{ji}^{-1}$ (note indeed that $h_i^M(x_j) = \delta_{ij}$).

The error analysis of the interpolation procedure classically involves the Lebesgue constant $\Lambda_M = \sup_{x\in\Omega} \sum_{i=1}^M |h_i^M(x)|$.

Lemma 2.2 *For any $\varphi \in F$, the interpolation error satisfies*

$$\left\|\varphi - \mathcal{I}_M[\varphi]\right\|_{L^\infty(\Omega)} \le (1 + \Lambda_M) \inf_{\psi_M \in X_M} \|\varphi - \psi_M\|_{L^\infty(\Omega)}. \tag{7}$$

The last term in the right hand side of the above inequality is known as the best fit of φ by elements in X_M.

2.2 The Generalization

Let us assume now that we do not have access to the values of $\varphi \in F$ at points in Ω easily, but, on the contrary, that we have a dictionary of linear forms $\sigma \in \Sigma$—assumed to be continuous in some sense, e.g. in $L^2(\Omega)$ with norm 1—the application of which over each $\varphi \in F$ is easy. Our extension consists in defining $\tilde{\varphi}_1, \tilde{\varphi}_2, \ldots, \tilde{\varphi}_M$ and a family of associated linear forms $\sigma_1, \sigma_2, \ldots, \sigma_M$ such that the following generalized interpolation process (our GEIM) is well defined:

$$\mathcal{J}_M[\varphi] = \sum_{j=1}^M \beta_j \tilde{\varphi}_j, \quad \text{such that} \quad \forall i = 1, \ldots, M, \quad \sigma_i\big(\mathcal{J}_M[\varphi]\big) = \sigma_i(\varphi). \tag{8}$$

Note that the GEIM reduces to the EIM when the dictionary is composed of Dirac masses, defined in the dual space of $\mathcal{C}^0(\Omega)$.

As explained in the introduction, our generalization is motivated by the fact that, in practice, measurements provide outputs from function φ that are some averages—or some moments—of φ over the actual size of the mechanical device that takes the measurement.

Among the questions raised by GEIM:

- is there an optimal selection for the linear forms σ_i within the dictionary Σ?
- is there a constructive optimal selection for the functions $\tilde{\varphi}_i$?
- given a set of linearly independent functions $\{\tilde{\varphi}_i\}_{i\in[1,M]}$ and a set of continuous linear forms $\{\sigma_i\}_{i\in[1,M]}$, does the interpolant (in the sense of (8)) exist?
- is the interpolant unique?
- how does the interpolation process compares with other approximations (in particular orthogonal projections)?
- Under what hypothesis can we expect the GEIM approximation to converge rapidly to φ?

In what follows, we provide answers to these questions either with rigorous proofs or with numerical evidences.

The construction of the generalized interpolation functions and linear forms is done recursively, following the same procedure as in the previous subsection, based on a greedy approach, both for the construction of the interpolation linear forms $\tilde{\varphi}_i$ and the associated forms selected in the dictionary Σ: The first interpolating function is, e.g.:

$$\tilde{\varphi}_1 = \arg\sup_{\varphi\in F} \|\varphi\|_{L^2(\Omega)},$$

the first interpolating linear form is:

$$\sigma_1 = \arg\sup_{\sigma\in\Sigma} \big|\sigma(\varphi_1)\big|.$$

We then define the first basis function as: $\tilde{q}_1 = \frac{\tilde{\varphi}_1}{\sigma_1(\tilde{\varphi}_1)}$. The second interpolating function is:

$$\tilde{\varphi}_2 = \arg\sup_{\varphi\in F} \big\|\varphi - \sigma_1(\varphi)\tilde{q}_1\big\|_{L^2(\Omega)}.$$

The second interpolating linear form is:

$$\sigma_2 = \arg\sup_{\sigma\in\Sigma} \big|\sigma\big(\tilde{\varphi}_2 - \sigma_1(\tilde{\varphi}_2)\tilde{q}_1\big)\big|,$$

and the second basis function is defined as:

$$\tilde{q}_2 = \frac{\tilde{\varphi}_2 - \sigma_1(\tilde{\varphi}_2)q_1}{\sigma_2(\tilde{\varphi}_2 - \sigma_1(\tilde{\varphi}_2)q_1)},$$

and we proceed by induction: assuming that we have built the set of interpolating functions $\{\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_{M-1}\}$ and the set of associated interpolating linear forms $\{\sigma_1, \sigma_2, \ldots, \sigma_{M-1}\}$, for $M > 2$, we first solve the interpolation problem: find $\{\widetilde{\alpha_j^{M-1}(\varphi)}\}_j$ such that

$$\forall i = 1, \ldots, M-1, \quad \sigma_i(\varphi) = \sum_{j=1}^{M-1} \widetilde{\alpha_j^{M-1}}(\varphi)\sigma_i(\tilde{q}_j),$$

and then compute:

$$\mathcal{I}_{M-1}[\varphi] = \sum_{j=1}^{M-1} \widetilde{\alpha_j^{M-1}}(\varphi)\tilde{q}_j.$$

We then evaluate

$$\forall\varphi\in F, \quad \varepsilon_M(\varphi) = \big\|\varphi - \mathcal{I}_{M-1}[\varphi]\big\|_{L^2(\Omega)},$$

and define:

$$\tilde{\varphi}_M = \arg\sup_{\varphi\in F} \varepsilon_{M-1}(\varphi)$$

and: $\sigma_M = \arg\sup_{\sigma\in\Sigma} |\sigma(\tilde{\varphi}_M - \mathcal{J}_{M-1}[\tilde{\varphi}_M])|$. The next basis function is then

$$\tilde{q}_M = \frac{\tilde{\varphi}_M - \mathcal{J}_{M-1}[\tilde{\varphi}_M]}{\sigma_M(\tilde{\varphi}_M - \mathcal{J}_{M-1}[\tilde{\varphi}_M])}.$$

We finally define the matrix $\widetilde{B^M}$ such that $\tilde{B}^M_{ij} = \sigma_i(\tilde{q}_j)$, and set $\widetilde{X}_M \equiv \text{span}\{\tilde{q}_j, j \in [1, M]\} = \text{span}\{\tilde{\varphi}_j, \ j \in [1, M]\}$. It can be proven as in [7, 10, 12].

Lemma 2.3 *For any $M \leq M_{\max}$, the set $\{\tilde{q}_j, j \in [1, M]\}$ is linearly independent and $\tilde{X}_M$ is of dimension M. The matrix B^M is lower triangular with unity diagonal (hence invertible) with other entries $\in [-1, 1]$. The generalized empirical interpolation procedure is well-posed in $L^2(\Omega)$.*

In order to quantify the error of the interpolation procedure, like in the standard interpolation procedure, we introduce the Lebesgue constant in the L^2 norm: $\Lambda_M = \sup_{\varphi\in F} \frac{\|\mathcal{J}_M[\varphi]\|_{L^2(\Omega)}}{\|\varphi\|_{L^2(\Omega)}}$ i.e. the L^2-norm of $\mathcal{J}_M$. A similar result as in the previous subsection holds.

Lemma 2.4 $\forall\varphi \in F$, *the interpolation error satisfies*:

$$\left\|\varphi - \mathcal{J}_M[\varphi]\right\|_{L^2(\Omega)} \leq (1 + \Lambda_M) \inf_{\psi_M\in\tilde{X}_M} \|\varphi - \psi_M\|_{L^2(\Omega)}.$$

A (very pessimistic) upper-bound for Λ_M is:

$$\Lambda_M \leq 2^{M-1} \max_{i\in[1,M]} \|q_i\|_{L^2(\Omega)}.$$

Proof The first part is standard and relies on the fact that, for any $\psi \in \tilde{X}_N$ then $\mathcal{J}_M(\psi_M) = \psi_M$. It follows that

$$\begin{aligned}
&\forall\psi_M \in \tilde{X}_M,\\
&\left\|\varphi - \mathcal{J}_M[\varphi]\right\|_{L^2(\Omega)} = \left\|[\varphi - \psi_M] - \mathcal{J}_M[\varphi - \psi_M]\right\|_{L^2(\Omega)}\\
&\qquad\qquad \leq (1 + \Lambda_M)\|\varphi - \psi_M\|_{L^2(\Omega)}.
\end{aligned}$$

Let us now consider a given $\varphi \in F$ and its interpolant $\mathcal{J}_M[\varphi] = \sum_{i=1}^M \widetilde{\alpha_i^M}(\varphi)\tilde{q}_i$ in dimension M. The constants $\widetilde{\alpha_i^M}(\varphi)$ come from the generalized interpolation problem: $\forall j \in [1, M]$, $\sigma_j(\varphi) = \sum_{i=1}^{j-1} \widetilde{\alpha_i^M}(\varphi)\sigma_j(\tilde{q}_i) + \widetilde{\alpha_j^M}(\varphi) j(\psi)$. We infer the recurrence relation for the constants:

$$\begin{aligned}
&\forall j \in [1, M],\\
&\widetilde{\alpha_j^M}(\varphi) = \sigma_j(\psi) - \sum_{i=1}^{j-1} \alpha_i(\psi)\sigma_j(q_i).
\end{aligned}$$

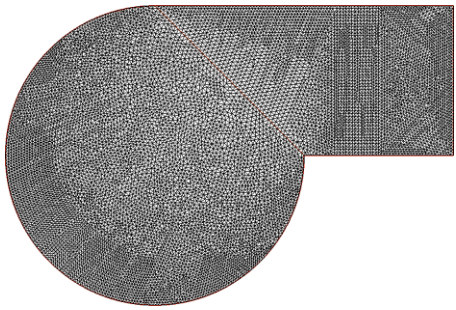


Fig. 1 The domain Ω and its mesh

Based on the properties of the entries in matrix $\tilde{B}^M$ stated in Lemma 2.3, we can obtain, by recurrence, an upper bound for each $\widetilde{\alpha_j^M}(\varphi)$: $\forall j \in [1, M]$, $|\widetilde{\alpha_j^M}(\varphi)| \leq (2^{j-1})\|\varphi\|_{L^2(\Omega)}$. Then, $\forall \varphi \in F$, $\forall M \leq M_{\max}$: $\|\mathcal{J}_M(\varphi)\|_{L^2(\Omega)} \leq [\sum_{i=1}^{M}(2^{j-1})\|q_i\|_{L^2(\Omega)}]\|\varphi\|_{L^2(\Omega)}$. Therefore: $\Lambda_M \leq 2^{M-1}\max_{i\in[1,M]}\|q_i\|_{L^2(\Omega)}$. Note that the norms of the rectified basis function q_i verify $\|q_i\|_{L^2(\Omega)} \geq 1$ from the hypothesis done on the norm of the σ_i. □

2.3 Numerical Results

The results that we present here to illustrate the GEIM are based on data acquired in silico using the finite element code Freefem [4] on the domain represented in Fig. 1.

We consider over the domain $\Omega \in \mathbb{R}^2$ the Laplace problem:

$$\begin{aligned} -\Delta\varphi &= f, \quad \text{in } \Omega \\ f &= 1 + \big(\alpha \sin(x) + \beta \cos(\gamma\pi y)\big)\chi_1(x, y) \end{aligned} \tag{9}$$

complemented with homogeneous Dirichlet boundary conditions. Here α, β and γ are 3 parameters freely chosen in given intervals in $\mathbb{R}$ that modulate the forcing term on the right hand side. We assume that the forcing term only acts on a part of Ω named Ω_1 ($\Omega_1 = \text{support}(\chi_1)$) and we denote as Ω_2 the remaining part $\Omega_2 = \Omega \setminus \overline{\Omega}_1$.

The easy observation is that the solution φ, depends on the parameters α, β, γ: we plot in Fig. 2 one of the possible solutions.

We also note that the restriction $\varphi_{|\Omega_2}$ to Ω_2 is indirectly dependent on these coefficients and thus is a candidate for building a set (when the parameters vary) of small Kolmogorov width. This can be guessed if we look at the numerical simulations obtained choices for α, β, γ (see Fig. 3).

For the GEIM, we use moments computed from the restriction of the solution $\varphi(\alpha, \beta, \gamma)$ over Ω_2 multiplied by localized functions with small compact support

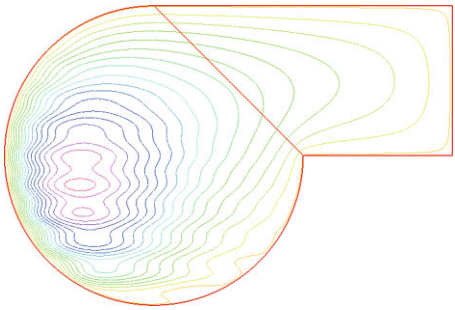


Fig. 2 One of the solutions, we note that the effect of the forcing is mainly visible on domain Ω_1 on the *left hand side*

over Ω_2. The reconstructed solutions with the GEIM based on only 5 interpolating functions is 10^{14} time better than the reconstructed function with 1 interpolating function illustrating the high order of the reconstruction's convergence.

In the next example, we choose a similar problem but the shape of domain Ω_2 a further parameter (see Fig. 4).

In order to get an idea of the Kolmogorov width of the set $\{\varphi_{|\Omega_2}(\alpha, \beta, \gamma, \Omega_2)\}$, we perform two Singular Value Decompositions (one in L^2, the other in H^1) over 256 values (approximated again with Freefem) and plot the decay rate of the eigenvalues ranked in decreasing order: the results are shown in Fig. 5.

We note that after about 9 eigenvalues, the finite element error dominates the decay rate of the true eigenvalues. The GEIM is built up again with captors represented as local weighted averages over Ω_2. The interpolation error is presented on the next figure (Fig. 6) and we note that the decay rate, measured both in L^2 and H^1 is again quite fast. In order to compare with the best fit represented by the projection, in L^2 or in H^1, we use the SVD eigenvectors associated with the first M eigenvalues and compare it with $\mathcal{J}_M$, for various values of M. This is represented on Fig. 7.

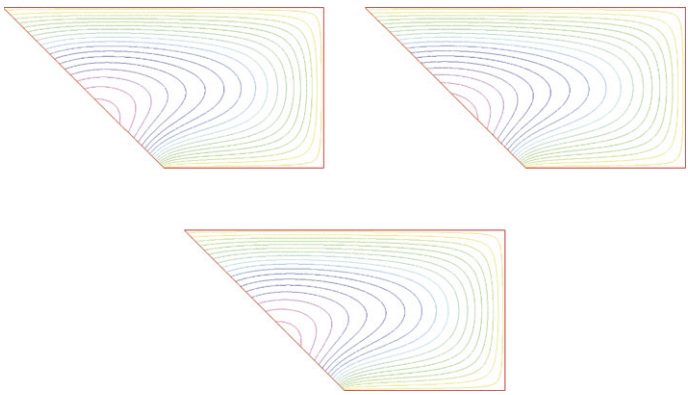


Fig. 3 Three generic solutions restricted on the sub-domain Ω_2

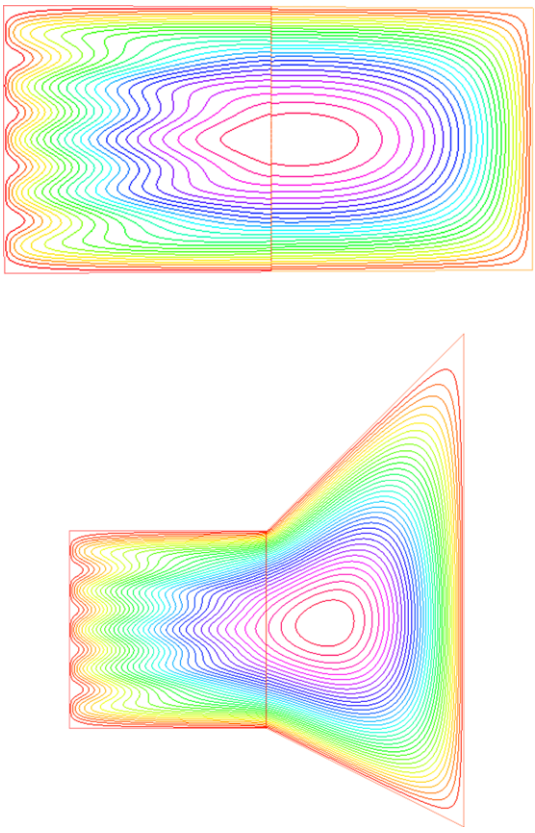


Fig. 4 Two generic solutions when shape of the sub-domain Ω_2 varies

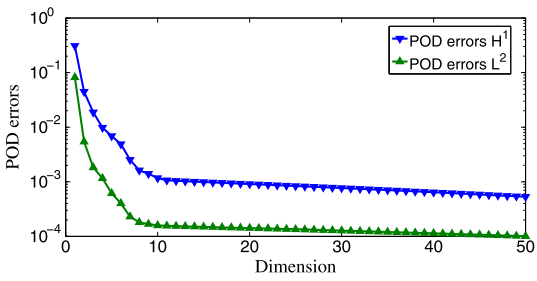


Fig. 5 Two SVD (in L^2 and in H^1) of the set of solutions over Ω_2

The very good comparison allow to expect that the Lebesgue constant is much better than what is announced in Lemma 2.4. A computational estimation of Λ_M (represented in Fig. 8) has been carried out:

$$\widetilde{\Lambda_M} = \max_{i\in[1,256]} \frac{\|\mathcal{I}_M[u_i]\|_{L^2(\Omega)}}{\|u_i\|_{L^2(\Omega)}}.$$

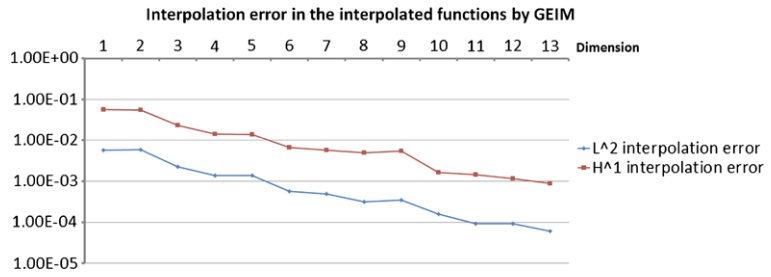


Fig. 6 The worse GEIM error with respect to M

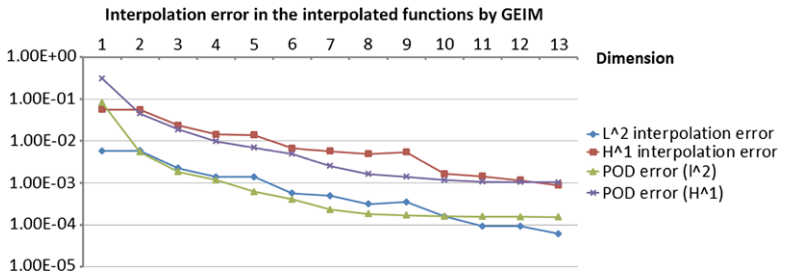


Fig. 7 Evolution of the GEIM error versus the best fit error, both in L^2 and in H^1-norms

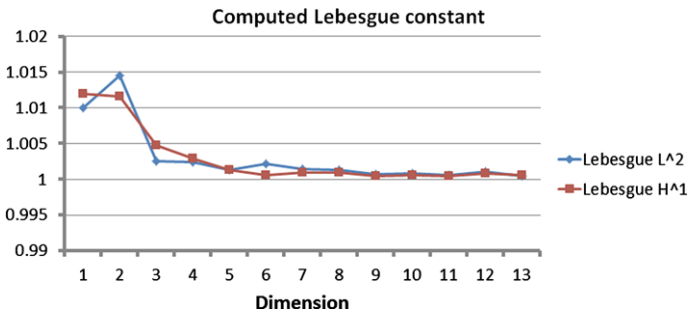


Fig. 8 Evolution of the Lebesgue constant, i.e. the norm of the GEIM operator, both in L^2 and in H^1

3 Coupling of Deterministic and Assimilation Methods

3.1 The Framework

Imagine that we want to supervise a process in **real-time** for which we have a parameter dependent PDE. Assume that the computation of the solution over the full domain Ω is too expensive but we are in a situation where the domain Ω can be decomposed, as before, into two non overlapping subdomains Ω_1 and Ω_2 and that

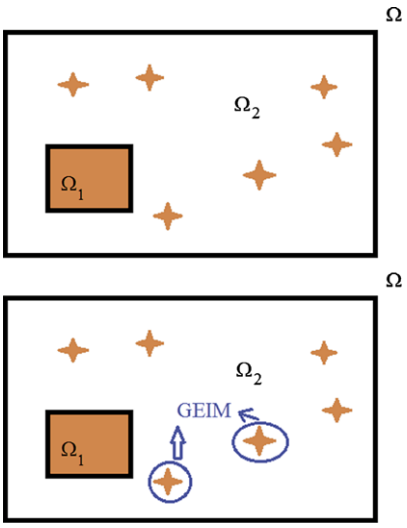


Fig. 9 Schematic representation of the reconstruction over Ω_2

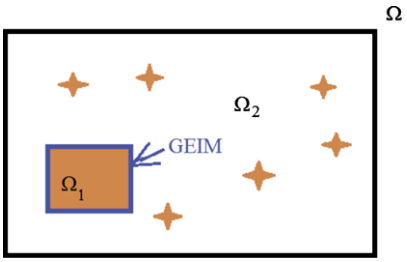


Fig. 10 Schematic representation of the recovery over Ω_1 thanks to the knowledge of the interface condition

- Ω_1 is small subdomain but the set of the restriction of the parameter dependent solutions has a large Kolmogorov width.
- Ω_2 is a big subdomain but the set of the restriction of the parameter dependent solutions has a small Kolmogorov n-width.

In addition assume that it is possible to get outputs from sensors based in Ω_2. The GEIM allows to reconstruct accurately the current solution associated to some parameters over Ω_2 and thus is able to build the boundary condition necessary over the interface between Ω_1 and Ω_2 that with the initially given boundary condition over $\partial\Omega$ to be the necessary boundary condition over $\partial\Omega_1$ that complement the original PDE set now over Ω_1 and not Ω as illustrated in Figs. 9 and 10.

3.2 The Combined Approach—Numerical Results

We take over the numerical frame of the previous section and go further. We want to apply the GEIM to have a knowledge of the solution $\varphi_{|\Omega_2}$ and want to use the trace

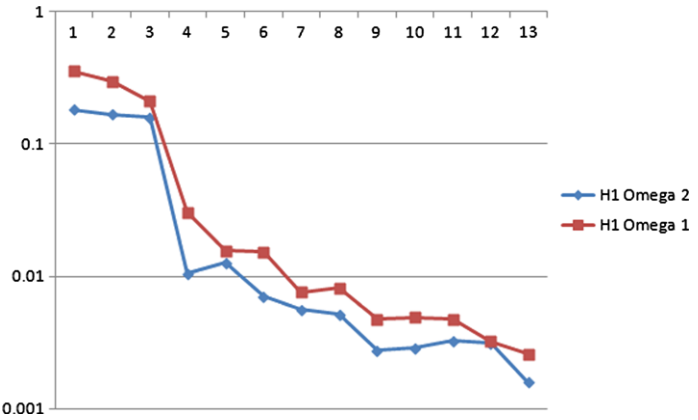


Fig. 11 Reconstructed analysis—error in H^1-norm over Ω_1 and Ω_2

of the reconstruction on the interface to provide the boundary condition, over $\partial\Omega_1$ to the problem

$$-\Delta\varphi = f, \quad \text{in } \Omega_1$$
$$f = 1 + \big(\alpha \sin(x) + \beta \cos(\gamma\pi y)\big)\chi_1(x, y)$$

derived from (9).

The results are presented in Fig. 11 where both the H^1 error on $\varphi_{|\Omega_1}$ and $\varphi_{|\Omega_2}$ are presented as a function of M being the number of interpolation data that are used to reconstruct $\varphi_{|\Omega_2}$. This illustrates that the use of the small Kolmogorov width of the set $\{\varphi_{|\Omega_2}\}$ as the parameters vary (including the shape of Ω_2) can help in determining the value of the full φ all over Ω.

4 About Noisy Data

In practical applications, data are measured with an intrinsic noise due to physical limitations of the sensors. In some sense, the noisy data acquired from the sensors are exact acquisitions from a noisy function that we consider to be a Markovian random field with spacial values locally dependent (on the support of the sensor) and globally independent (from one sensor to the others). An extension of the previous development needs therefore to be done in order to take this fact under consideration.

Let us assume that all the sensors are subject to the same noise, i.e. provide averages—or some moments—computed, not from φ, but from a random process $\varphi_\varepsilon \simeq \mathcal{N}(\varphi, \varepsilon^2)$. The norm of the GEIM operator being equal to Λ_M the GEIM-reconstruction forms a random process $\mathcal{J}_M[\varphi_\varepsilon] \simeq \mathcal{N}(\mathcal{J}_M[\varphi], \Lambda_M^2\varepsilon^2)$ due to linearity.

Even though the Lebesgue constant seems to be small in practice, we would like to use all the data that are available in order to get a better knowledge of φ. For the

definition of $\mathcal{J}_M$ we indeed only use M data selected out of a large set of all data. For this purpose, let us consider that, with some greedy approaches, we have determined P independent series of M different captors $\{\sigma_1^{(p)}, \sigma_2^{(p)}, \ldots, \sigma_M^{(p)}\}, \forall 1 \leq p \leq P$. For each of these series, the GEIM applied to φ is noisy and each application provides $\mathcal{J}_M^p[\varphi_\varepsilon] \simeq \mathcal{N}(\mathcal{J}_M^p[\varphi], {\Lambda_M^p}^2 \varepsilon^2)$. We shall use these P reconstructions by averaging them and expect to improve the variance of the reconstruction.

Let $\lambda^{-1} = \frac{1}{P}\sum_{p=1}^{P} \frac{1}{\Lambda_N^p}$. Since the P realizations: $\{\mathcal{J}_M^p[\varphi_\varepsilon]\}_p$ are independent, then the random variable $\overline{\mathcal{J}_M^P}(\varepsilon) = \frac{\lambda}{P}\sum_{p=1}^{P} \frac{\mathcal{J}_M^p[\varphi_\varepsilon]}{\Lambda_N^{(p)}}$ follows a Gaussian Markov random field of parameters $\mathcal{N}(\mathcal{J}_N(\varphi), \frac{\varepsilon^2\lambda^2}{P})$. A realization of this random process could be chosen for an improved estimate of $\mathcal{J}_M(\varphi)$. Indeed, the law of the error follows $\mathcal{N}(0, \frac{\varepsilon^2\lambda^2}{P})$ and its variance can be less than the size of the initial noise on the captors (ε) provided that $\Lambda_N^{(p)} < \sqrt{P}, \forall 1 \leq p \leq P$, which, from the numerical experiments, seems to be the case.

5 Conclusions

We have presented a generalization of the Empirical Interpolation Method, based on ad hoc interpolating functions and data acquired from sensors of the functions to be represented as those that can arise from data assimilation. We think that the GEIM is already interesting per se as it allows to select in a greedy way the most informative sensors one after the other. It can also propose, in case this is feasible, to build better sensors in order to complement a given family of existing ones and/or detect in which sense some of them are useless because redundant. Finally we also explain how noise on the data can be filtered out.

The coupled use of GEIM with reduced domain simulation is also proposed based on domain decomposition technique leading to a small portion where numerical simulation is performed and a larger one based on data assimilation.

We think that the frame presented here can be used as an alternative to classical Bayesian or frequentistic statistic where the knowledge developed on the side for building mathematical models and their simulations can be fully used for data mining (we refer also to [8] and [11] for recent contributions in this direction).

Acknowledgements This work was supported in part by the joint research program MANON between CEA-Saclay and University Pierre et Marie Curie-Paris 6. We want to thank G. Biot from LSTA and G. Pagès from LPMA for constructive discussions on the subject.

References

1. Barrault, M., Nguyen, N.C., Maday, Y., Patera, A.T.: An "empirical interpolation" method: Application to efficient reduced-basis discretization of partial differential equations. C. R. Acad. Sci. Paris, Sér. I **339**, 667–672 (2004)

2. Chakir, R., Maday, Y.: A two-grid finite-element/reduced basis scheme for the approximation of the solution of parametric dependent PDE. C. R. Math. **347**, 435–440 (2009). doi:10.1016/j.crma.2009.02.019
3. Grepl, M.A., Maday, Y., Nguyen, N.C., Patera, A.T.: Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. M2AN (Math. Model. Numer. Anal.) **41**(3), 575–605 (2007)
4. http://www.freefem.org
5. Kolmogoroff, A.: Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse. Ann. Math. (2) **37**, 107–110 (1936)
6. Maday, Y., Nguyen, N.C., Patera, A.T., Pau, G.S.H.: A general multipurpose interpolation procedure: the magic points. Commun. Pure Appl. Anal. **8**(1), 383–404 (2009)
7. Nguyen, N.C., Veroy, K., Patera, A.T.: Certified real-time solution of parametrized partial differential equations. In: Yip, S. (ed.) Handbook of Materials Modeling, pp. 1523–1558. Springer, Berlin (2005)
8. Patera, A.T., Rønquist, E.M.: Regression on parametric manifolds: estimation of spatial fields, functional outputs, and parameters from noisy data. C. R. Acad. Sci. Paris, Sér. I **350**(9–10), 543–547 (2012)
9. Patera, A.T., Rozza, G.: Reduced basis approximation and a posteriori error estimation for parametrized partial differential equations, Version 1.0. Copyright MIT 2006 to appear in (tentative rubric). MIT Pappalardo Graduate Monographs in Mechanical Engineering. http://augustine.mit.edu/methodology/methodology_bookPartI.htm
10. Prud'homme, C., Rovas, D., Veroy, K., Maday, Y., Patera, A.T., Turinici, G.: Reliable real-time solution of parametrized partial differential equations: reduced-basis output bound methods. J. Fluids Eng. **124**(1), 70–80 (March 2002)
11. Rozza, G., Manzoni, A., Negri, F.: Reduction strategies for PDE-constrained optimization problems in haemodynamics. MATHICSE Technical Report, Nr. 26.2012 (July 2012)
12. Veroy, K., Prud'homme, C., Rovas, D.V., Patera, A.T.: A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In: Proceedings of the 16th AIAA Computational Fluid Dynamics Conference (AIAA Paper 2003-3847) (June 2003)

Analysis and Numerics of Some Fractal Boundary Value Problems

Umberto Mosco

Abstract We describe some recent results for boundary value problems with fractal boundaries. Our aim is to show that the numerical approach to boundary value problems, so much cherished and in many ways pioneering developed by Enrico Magenes, takes on a special relevance in the theory of boundary value problems in fractal domains and with fractal operators. In this theory, in fact, the discrete numerical analysis of the problem precedes, and indeed give rise to, the asymptotic continuous problem, reverting in a sense the process consisting in deriving discrete approximations from the PDE itself by finite differences or finite elements. As an illustration of this point, in this note we describe some recent results on: the approximation of a fractal Laplacian by singular elliptic partial differential operators, by Vivaldi and the author; the asymptotic of degenerate Laplace equations in domains with a fractal boundary, by Capitanelli-Vivaldi; the fast heat conduction on a Koch interface, by Lancia-Vernole and co-authors. We point out that this paper has an illustrative purpose only and does not aim at providing a survey on the subject.

1 Introduction

Very simple fractals, as the Koch curve or the Sierpiński gasket, are non-differentiable sets, therefore they do not allow the explicit writing of an intrinsic differential operator. Moreover, if the fractal is the (compact) boundary of an open domain of the plane, the boundary trace spaces may be difficult to characterize. Merging classical boundary value problems—even of the simple kind as those related to the Laplace equation or the heat equation—with the theory of fractal sets and fractal operators is a challenging task. In fact, fractal sets and operators are the result of an *asymptotic* process, the one induced by the infinite iterations of a family of contractive similarities. Such an asymptotic feature of fractals introduces an additional approximation

In memory of Enrico Magenes.

U. Mosco (✉)
Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280, USA
e-mail: mosco@wpi.edu

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*, Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_14,

level to those already inherent in the numerical approximation of the classical PDE at hand.

The construction of fractal sets goes back to the early years of the twentieth century. The construction of the Laplace and heat equations on a large class of fractals was achieved in the late 1980's, first by probabilistic methods, then analytically. We refer to [25] for a brief description of this early work.

A problem related to heat conduction in a planar domain relies on the idea that the insertion in the domain of a highly conductive material path connecting two points of the boundary can efficiently act as a preferential fast absorbing trail for the heat stream. An early model of the kind, with an infinitely conductive layer, was produced by Cannon and Meyer in 1971, [5], in connection with so called fractured oil wells. A related singular homogenization problem was later considered by Pam Huy-Sanchez Palencia in 1974, [31], see also [2, 24].

In the simplest version of this model, the domain is a rectangle, the infinitely conductive pattern is the segment connecting the middle points A and B of two opposite sides of the rectangle, and the segment is approximated by thin highly conductive rectangles of transversal size ε. In the two regions above and below this thin layer, the two-dimensional heat equation, with a normalized conductivity coefficient and with a prescribed source term, is assumed to govern the slow diffusion of heat. At the same time, the fast diffusion of heat within the ε-layer is described by the two-dimensional heat equation, this time with a conductivity coefficient of the order $1/\varepsilon$. The boundary condition for the temperature is zero on the boundary of the rectangle. In the limit as $\varepsilon \to 0$, the thin layer shrinks to the transversal segment. In this process, the temperature converges to a limit temperature, given by the *two-dimensional* heat equation in each one of the open domains separated by the segment and by the *one-dimensional* heat equation along the segment itself. These two equations—*both* of second order—are coupled by a transmission condition across the segment. This condition stipulates that the *jump* of the external normal derivatives from each side of the segment acts as a *source term* for the one-dimensional tangential heat equation within the segment; moreover, the tangential diffusion has boundary values zero at the end points A and B.

We note, incidentally, that the natural Sobolev space taking into account this homogeneous boundary condition for the tangential equation is the so-called *Lions-Magenes* space $H_{0,0}^{1/2}$. A fractal analogue of the Lions-Magenes space occurs in the problems that we now describe.

A big innovation into the transmission model was indeed carried out in 2002 by M.R. Lancia, [16]. The segment connecting the points A and B of the boundary of the domain was replaced by a *fractal Koch curve*, connecting again A with B. The rationale for this new model is clear. By increasing the length of the preferential pattern that conveys the heat stream towards the two selected points A and B, and actually making the length of this path *infinite* in the limit, we expect that the cooling effect of the layer will be increased.

As mentioned previously, in Lancia's fractal model the transmission problem with the fractal Koch curve is obtained in the limit of a sequence of transmission

problems for the approximating pre-fractal polygonal curves as the number of iteration increases to infinity. This model opens two related orders of problems. One is the rigorous analytic formulation of the second order transmission condition, in suitable fractional Besov spaces. This study was first carried out by Lancia in [16] and we refer to this paper and to [18] for the technical details. The second problem is the analytical and numerical study of the approximating pre-fractal equations. We report on this study in Sect. 3.

We also report on some recent results of a joint research by Vivaldi and the author, [29, 30], that is indeed related to the second order transmission problems discussed so far. The object of this study is a sequence of second order elliptic operators

$$A^n_{\varepsilon_n} u = -\operatorname{div}\big(a^n_{\varepsilon_n}(x, y)\nabla u\big) \tag{1}$$

in divergence form in a bounded domain Ω of $\mathbb{R}^2$, with discontinuous coefficients $a^n_{\varepsilon_n}$. The coefficients $a^n_{\varepsilon_n}$ develop an increasing number of singularities on an array of thin fibers $\Sigma^n_{\varepsilon_n}$ obtained by the iterated action of a given family of contractive similarities. The geometry and the singularity of the conductivity coefficients are initially prescribed on an array of thin hexagonal fibers connecting the essential fixed points of the similarities. The parameter $n = 1, 2, \ldots$ indicates the level of iteration of the similarities, the parameters ε_n refer to the transversal thickness of the fiber at each nth-iteration. The detailed geometry of the fibers and the expression of the singular coefficients $a^n_{\varepsilon_n}$ are described in Sect. 1. One of the main objective of this study is to prove the convergence of the spectral measures for the elliptic operators to the spectral measure of a limit self-adjoint operator. The limit operator is the intrinsic Laplace operator of the fractal set defined by the given family of similarities. We outline our main results in Sect. 1.

A related topic has been the object of a recent paper by Capitanelli and Vivaldi, [7]. They consider the domain bounded by a square snowflake type domain, bounded by four Koch curves, and the approximated domains obtained by replacing each side of the pre-fractal Koch curves by a quadrilateral thin fiber of the kind mentioned before. Differently than in the previous case, the conductivity of the fiber is now assumed to vanish with the iterations. On such increasingly insulating boundary layers, a homogeneous Dirichlet condition is imposed to the solution of a Laplace equation. A characterization of the limit boundary value problem in the snowflake domain is then given, that depends on the relative size of the thickness and the conductivity of the boundary layer. These results are outlined in Sect. 2.

There is no space in this note to report on some other aspects of the numerics of fractals. However, we wish to mention a new kind of interesting problems that deserve further research. The objective of this study is to approximate the two dimensional Laplace operator (and related PDEs) in an open domain of the plane, with a sequence of curvilinear one-dimensional Laplacians (and related ODEs), taken along a sequence of fractal curves homeomorphic to the segment [0, 1], that asymptotically fill the whole open domain. Such a dynamical dimensional blow up is of theoretical interest in itself. It seems also interesting from the numerical point of view and, in the applications, as a model for the study of *invasive interfaces* that infiltrate the whole space.

In conclusion, we may observe that introducing fractal constructions into the classic theory of PDEs opens a vast new field of study, both theoretically and numerically. With the very simple examples object of this note, and with other recent contributions—in particular, the works by Vacca [32], Bagnerini-Buffa-Vacca [3], Wasyk [33], Evans [10], Liang [9, 23], and the work by Achdou-Sabot-Tchou, [1] that explores different but related problems—this new field has been only scratched.

Enough however to unveil promising new directions in applied analysis and PDEs, and to point out how fundamental is to keep analysis and numerics in tight contact one each other. A point of view this one to which the author was already exposed in early years of his scientific life in Rome by his adviser Gaetano Fichera, and that he is happy to take now also as one of the most illuminating aspects of the scientific legacy of Enrico Magenes.

2 Elliptic Operators with Fractal Singularities

In this section we describe the recent work carried out by Vivaldi and the author in [29, 30]. In these papers, a singular elliptic operator is submitted to the iterated action of a family of similarities and the convergence of the spectrum is investigated.

We begin by introducing the similarities. We consider a family $\Psi = \{\psi_1, \ldots, \psi_N\}$ of $N \geq 2$ contractive similarities in $\mathbb{R}^2$, with distinct fixed-points, with a common contractive factor α^{-1}, $\alpha > 1$; a *similarity*, or *similitude*, in a Euclidean space is a map obtained by composition of translations, orthogonal transformations, and homotheties. The set of *essential fixed-points* of these maps will be denoted by Γ; a point $b_r \in \mathbb{R}^2$ is an *essential fixed-point* for the family Ψ if $b_r = \psi_i(b_r) = \psi_j(b_s)$ for some $i \in \{1, \ldots, N\}$, $j \neq i$, $j \in \{1, \ldots, N\}$ and b_s a fixed-point of a map of Ψ.

Here, for simplicity, we assume that Ψ is the so-called *Koch family* of similarities, that is, the family $\Psi = \{\psi_1, \ldots, \psi_4\}$ of the following $N = 4$ similitudes, each one contractive with a factor α^{-1}, $\alpha = 3$:

$$\psi_1(z) = \frac{z}{3}, \qquad \psi_2(z) = \frac{z}{3}\, e^{i\pi/3} + \frac{1}{3},$$

$$\psi_3(z) = \frac{z}{3} e^{-i\pi/3} + \frac{1}{2} + \frac{i \sin \pi/3}{3}, \qquad \psi_4(z) = \frac{z+2}{3},$$

where $z = x + iy \in \mathbb{C}$. The set of the essential fixed-points of this family is $\Gamma = \{A, B\}$, where $A = (0, 0)$ and $B = (1, 0)$. The third vertex of the equilateral triangle based on the side A, B is the point $C = (1/2, \sqrt{3}/2)$.

We now define a reference fiber in the Cartesian plane $\mathbb{R}^2$. This fiber is a thin hexagon which has the segment connecting the points $A = (0, 0)$ and $B = (1, 0)$ as its longitudinal axis. The middle point of the segment AB is denoted by $AB/2$. The fiber is symmetric with respect to the *x-axis* and the vertical line $x = 1/2$, therefore, it suffices to describe the geometry of the fiber in the region $y \geq 0$, $x \leq 1/2$. We consider the right triangle with vertices A, $AB/2$, Q_0 which makes the angle $\pi/12$ at

A. Thus, $Q_0 = (1/2, \varepsilon_0)$, where $\varepsilon_0 = h_0/2$, $h_0 = \tan(\pi/12)$. For every $0 < \varepsilon \le \varepsilon_0$, we consider the two points $Q_1(\varepsilon) = (\varepsilon/h_0, \varepsilon)$ and $Q_0(\varepsilon) = (1/2, \varepsilon)$ and the quadrilateral A, $AB/2$, $Q_0(\varepsilon)$, $Q_1(\varepsilon)$. We then define the set $\Sigma^0_{0,2\varepsilon}$ to be the thin hexagon obtained by reflection of this quadrilateral across the *x-axis*, followed by a symmetry across the vertical axis $x = 1/2$. The vertices of $\Sigma^0_{0,2\varepsilon}$, listed clockwise, are the points A, $Q_1(\varepsilon)$, $Q_2(\varepsilon)$, B, $Q_3(\varepsilon)$, $Q_4(\varepsilon)$, where now $Q_2(\varepsilon) = (1 - \varepsilon/h_0, \varepsilon)$, $Q_3(\varepsilon) = (1 - \varepsilon/h_0, -\varepsilon)$, $Q_4(\varepsilon) = (\varepsilon/h_0, -\varepsilon)$. The perimeter of the hexagon $\Sigma^0_{0,2\varepsilon}$ gives the external profile of our fiber. Inside the hexagon $\Sigma^0_{0,2\varepsilon}$, we now insert a smaller hexagon $\Sigma^0_{0,\varepsilon}$. The construction of this hexagon is similar to that of $\Sigma^0_{0,2\varepsilon}$, by replacing the triangle A, $AB/2$, Q_0 with the smaller right triangle with vertices A, $AB/2$, P_0, where $P_0 = (1/2, \varepsilon_0/2)$. The angle of this triangle at A is $\arctan(h_0/2)$. The vertices of the hexagon $\Sigma^0_{0,\varepsilon}$, again listed clockwise, are now the points A, $P_1(\varepsilon)$, $P_2(\varepsilon)$, B, $P_3(\varepsilon)$, $P_4(\varepsilon)$, where now $P_1(\varepsilon) = (\varepsilon/h_0, \varepsilon/2)$, $P_2(\varepsilon) = (1 - \varepsilon/h_0, \varepsilon/2)$, $P_3(\varepsilon) = (1 - \varepsilon/h_0, -\varepsilon/2)$, $P_4(\varepsilon/h_0, -\varepsilon/2)$.

With the notation set before, the reference fiber is given by the two co-axial thin hexagons $\Sigma^0_{0,\varepsilon} \subseteq \Sigma^0_{0,2\varepsilon}$, of largest transversal size ε and 2ε, respectively. The two hexagons meet at the common vertices A and B, and $\Sigma^0_{0,\varepsilon} \setminus \{A, B\}$ is contained in the interior of $\Sigma^0_{0,2\varepsilon}$.

In the case at hand, the Koch family of similarities, the single reference fiber $\Sigma^0_{0,2\varepsilon}$ of our construction connects the two essential fixed-points $\Gamma = \{A, B\}$, which, in this special case, are the only essential fixed points of the family Ψ (the pair-wise connection of all essential fixed points of the family by means of a fiber is a requirement of this theory).

Our next step is to submit the fiber to the iterated action of the family Ψ. We first set a useful notation. For each integer $n \ge 0$, we consider arbitrary n-tuples of indices $i|n = (i_1, i_2, \ldots, i_n) \in \{1, \ldots, N\}^n$ and define $\psi_{i|n} = \psi_{i_1} \circ \psi_{i_2} \circ \cdots \circ \psi_{i_n}$ if $n > 0$, with $\psi_{i|n}$ the identity map if $n = 0$; for every set $\mathcal{O} \subseteq \mathbb{R}^2$, we define $\mathcal{O}^{i|n} = \psi_{i|n}(\mathcal{O})$. With this notation at hand, for every $n \ge 0$, we construct the array of fibers obtained by defining

$$\Sigma^n_{2\varepsilon} = \bigcup_{i|n} \Sigma^{i|n}_{2\varepsilon}, \qquad \Sigma^{i|n}_{2\varepsilon} = \psi_{i|n}\big(\Sigma^0_{2\varepsilon}\big), \tag{2}$$

$$\Sigma^n_{\varepsilon} = \bigcup_{i|n} \Sigma^{i|n}_{\varepsilon}, \qquad \Sigma^{i|n}_{\varepsilon} = \psi_{i|n}\big(\Sigma^0_{\varepsilon}\big) = \bigcup_{b_r \ne b_s \in \Gamma} \Sigma^{i|n}_{\varepsilon}(b_r, b_s), \tag{3}$$

where $\Sigma^0_{\varepsilon} = \Sigma^0_{0,\varepsilon}$, $\Sigma^0_{2\varepsilon} = \Sigma^0_{0,2\varepsilon}$.

The space $\mathbb{R}^2$—actually, a bounded open domain $\Omega \subset \mathbb{R}^2$, with $\overline{\Omega}$ containing $\overline{\Sigma}^0_{0,2\varepsilon}$, that contains the fibers $\Sigma^n_{2\varepsilon}$ for every ε and every n and which will be specified later on—is now converted into a physical *composite body*. This is done by defining a discontinuous conductivity matrix $a^n_{\varepsilon}(\xi, \eta)\mathrm{Id}$, $(\xi, \eta) \in \Omega$, again by the iterated action of the similarity family. The whole iteration process is externally governed by two sequences of constants, $\zeta_n > 0$ and $\gamma_n > 0$. The limit values assigned

to ζ_n and γ_n as $n \to +\infty$ affect the nature and the properties of the asymptotic effective medium.

In order to observe boundary effects, we choose Ω in a way that Γ belongs to the boundary $\partial\Omega$ of Ω, namely Ω is now the triangle with vertices $D = (1/2, -\sqrt{3}/2)$, $E = (3/2, \sqrt{3}/2)$, $F = (-1/2, \sqrt{3}/2)$. The domain Ω contains the interior of the triangle of vertices A, B, C, and the vertices A, B, C belong to $\partial\Omega$.

The matrix a_ε^n Id—for given $0 < \varepsilon \le \varepsilon_0$ and $n \ge 0$—is defined at every $(\xi, \eta) \in \Omega$ by

$$a_\varepsilon^n(\xi,\eta)\,\mathrm{Id} = \begin{cases} \zeta_n \mathbf{1}_{\Omega\setminus\Sigma_{2\varepsilon}^n}(\xi,\eta)\,\mathrm{Id} + \mathbf{1}_{\Sigma_{2\varepsilon}^n\setminus\Sigma_\varepsilon^n}(\xi,\eta)\,\mathrm{Id} \\ \quad + 1/2\sigma_n \sum_{b_r\neq b_s\in\Gamma} w_\varepsilon^n(\xi,\eta)\mathbf{1}_{\Sigma_\varepsilon^n(b_r,b_s)}(\xi,\eta)\,\mathrm{Id}\,. \end{cases} \tag{4}$$

In this expression, Id is the 2-dimensional identity matrix and $\mathbf{1}_S$ the indicatrix function of a set $S \subset \mathbb{R}^2$, that is, $\mathbf{1}_S(\xi,\eta) = 1$ if $(\xi,\eta) \in S$, $\mathbf{1}_S(\xi,\eta) = 0$ if $(\xi,\eta) \notin S$.

The constants σ_n, which will be specified later on, are scaling factors associated with Ψ. At each iteration n, $\zeta_n > 0$ is a material constant that takes into account how the conductivity of the surrounding medium evolves with n. Since in (4) the conductivity of the coating region $\Sigma_{2\varepsilon}^n \setminus \Sigma_\varepsilon^n$ has been normalized to 1, the constant ζ_n can be interpreted as a *viscosity* coefficient, that expresses the relative strength of the conductivity of the space that surrounds the fiber $\Sigma_{2\varepsilon}^n$ with respect to the conductivity of the fiber Σ_ε^n itself. For every $n \ge 0$ and $i|n$, the conductivity of the fiber $\Sigma_\varepsilon^{i|n}(b_r, b_s)$ is given by

$$w_\varepsilon^n(\xi,\eta)\mathbf{1}_{\Sigma_\varepsilon^{i|n}(b_r,b_s)}(\xi,\eta) = \gamma_n\alpha^n w_\varepsilon^0 \circ \psi_{i|n}^{-1}(\xi,\eta)\mathbf{1}_{\Sigma_\varepsilon^0(b_r,b_s)}(\xi,\eta). \tag{5}$$

This expression is obtained from the conductivity $w_\varepsilon^0(x,y)$ of the reference fiber $\Sigma_\varepsilon^0(b_r,b_s)$ by applying the map $(\xi,\eta) = \psi_{i|n}(x,y)$. The function $w_\varepsilon^0(x,y)$ is defined on the inner fiber $\Sigma_{0,\varepsilon}^0$ as follows:

$$w_\varepsilon^0(x,y) = \begin{cases} \frac{2+h_0^2}{4|P-P^\perp|} & \text{if } (x,y) \in \mathcal{T} \\ \frac{1}{2|P-P^\perp|} & \text{if } (x,y) \in \mathcal{R} \end{cases} \tag{6}$$

where $\mathcal{R}$ is the central rectangle in $\Sigma_{0,\varepsilon}^0$ with vertices P_1, P_2, P_3, P_4, and $\mathcal{T}$ is the union of the two isosceles triangles A, P_1, P_4 and P_2, B, P_3. For every $(x,y) \in \Sigma_{0,\varepsilon}^0$, we consider the point $P^\perp = (x,0)$ on the longitudinal axis of $\Sigma_{0,\varepsilon}^0$ and we define $P = (x, y_P)$ to be the intersection of the vertical line through $P^\perp = (x,0)$ with the boundary $\partial\Sigma_{0,\varepsilon}^0$ of $\Sigma_{0,\varepsilon}^0$ in the half plane $y \ge 0$. This boundary is the polygonal line connecting the vertices A, $P_1(\varepsilon)$, $P_2(\varepsilon)$, B. Then $|P - P^\perp|$ is the (Euclidean) distance between P and $P^\perp$ in $\mathbb{R}^2$.

At this stage, we are confronted with two asymptotic limits. For fixed n, the limit as $\varepsilon \to 0$ gives vanishing thickness to the fibered neighborhood of the pre-fractal polygonal curve. The limit as $n \to +\infty$ leads to the fractal set included in Ω. We proceed diagonally, by suitably choosing for each n a value $\varepsilon_n > 0$, such that $\varepsilon_n \to 0$ as $n \to +\infty$. Then, we take the single limit as $n \to +\infty$.

We consider the sequence of operators $A_n = A_{\varepsilon_n}$, where A_{ε_n} are the operators given in (1). The operators A_n are defined as self-adjoint operators in the space $L^2(\Omega)$ (with Neumann boundary condition on $\partial\Omega$). Our main goal is to show the convergence of the spectral measures $P_n(d\lambda)$ of the operators A_n to the spectral measure of a suitable self-adjoint asymptotic operator, as $n \to +\infty$. We rely on variational and convergence tools from [24]. In particular, by a general result in [24], we obtain the convergence of the spectral measures of the operators A_n to the spectral measure of a limit operator A as a consequence of the M-convergence of the (extended-valued) energy forms associated with these operators.

For every n, and for the specified value of $\varepsilon_n > 0$, the energy form of the operator A_{ε_n} in $L^2(\Omega)$ is the functional

$$F_n[u] = F^n_{\varepsilon_n}[u] = \begin{cases} \int_\Omega a^n_{\varepsilon_n}(x,y)|\nabla u|^2 dxdy & \text{if } u \in D[F^n_{\varepsilon_n}] \\ +\infty & \text{if } u \in L^2(\Omega) \setminus D[F^n_{\varepsilon_n}] \end{cases}$$

where $a^n_{\varepsilon_n}\mathrm{Id}$ is the coefficient matrix defined in (4) for $\varepsilon = \varepsilon_n$, and the domain $D[F^n_{\varepsilon_n}] \subset L^2(\Omega)$ is the completion of $C^1(\overline{\Omega})$ in the norm

$$\|u\|_{D[F^n_{\varepsilon_n}]} = \left\{ \int_\Omega |u|^2 dxdy + \int_\Omega |\nabla u|^2 a^n_{\varepsilon_n} dxdy \right\}^{\frac{1}{2}}. \tag{7}$$

As explained before, we let, simultaneously, the iteration parameter n go to $+\infty$ and the transversal size ε of the fibers go to 0, by choosing $\varepsilon = \varepsilon_n$ to be infinitesimal as $n \to +\infty$. We must also choose the scaling constants σ_n. These scaling laws can be expressed in terms of a single parameter $\delta > 0$. The value of δ is given by the ratio

$$\delta := \frac{d_H}{d_S}$$

of the Hausdorff dimension d_H of $\mathcal{G}$ and the spectral dimension d_S of $\mathcal{G}$. Here $\mathcal{G}$ is the invariant (self-similar) set defined by the family Φ. The constant δ is an effective metric parameter that depends on the fractal $\mathcal{G}$. For the Koch curve, $N = 4$, $\alpha = 3$, $\delta = \ln 4/\ln 3$. For the Sierpiński gasket, $N = 3$, $\alpha = 2$, $\delta = \ln 5/\ln 4$. Note that in both cases $\delta > 1$. We then define

$$\rho = \frac{\alpha^{2\delta}}{N}$$

and take

$$\varepsilon_n = \left(\frac{\rho}{N}\right)^n \omega_n, \quad (\text{with } \omega_n \to 0 \text{ as } n \to +\infty), \quad \sigma_n = \left(\frac{\rho}{\alpha}\right)^n. \tag{8}$$

We also assume that the material constants ζ_n, γ_n remain finite and non-vanishing through the iteration process:

$$\lim \zeta_n = \zeta^* \in (0, +\infty), \qquad \lim \gamma_n = \gamma^* \in (0, +\infty) \tag{9}$$

as $n \to +\infty$. In [30] the following result is obtained, which extends previous results from [27] and [28]:

Theorem 2.1 *With the value of $\delta > 0$ specified before, under the assumptions (8) and (9) the sequence of functionals F_n M-converges in $L^2(\Omega)$ to the functional*

$$F[u] = \begin{cases} \zeta^* \int_\Omega |\nabla u|^2 dxdy + \gamma^* \mathcal{E}_{\mathcal{G}}[u|_{\mathcal{G}}] & \text{if } u \in H^1(\Omega), u|_{\mathcal{G}} \in D[\mathcal{E}_{\mathcal{G}}] \\ +\infty \quad \text{if } u \in L^2(\Omega) \setminus \{u : u \in H^1(\Omega), u|_{\mathcal{G}} \in D[\mathcal{E}_{\mathcal{G}}]\} \end{cases} \tag{10}$$

where $\mathcal{E}_{\mathcal{G}}[u|_{\mathcal{G}}]$ is the energy functional on the fractal $\mathcal{G}$.

In this statement, $H^1(\Omega) \subset L^2(\Omega)$ is the Sobolev space obtained as the completion of $C^1(\overline{\Omega})$ in the norm

$$\|u\|_{H^1(\Omega)} = \left\{ \int_\Omega |u|^2 dxdy + \int_\Omega |\nabla u|^2 dxdy \right\}^{\frac{1}{2}}$$

and $u|_{\mathcal{G}}$ is the trace of $u \in H^1(\Omega)$ on $\mathcal{G}$, defined, *e.g.*, as in [13, 14]. For u in $C(\mathcal{G})$ the energy functional $\mathcal{E}_{\mathcal{G}}[u]$ is obtained as the increasing limit

$$\mathcal{E}_{\mathcal{G}}[u] = \lim_{n \to +\infty} \mathcal{E}^n_{\mathcal{G}}[u] \tag{11}$$

of the discrete energy forms

$$\mathcal{E}^n_{\mathcal{G}}[u] = \frac{1}{2} \frac{\alpha^{2n\delta}}{N^n} \sum_{i|n} \sum_{b_r \neq b_s \in \Gamma} \big(u(\psi_{i|n}(b_r)) - u(\psi_{i|n}(b_s))\big)^2, \tag{12}$$

on the domain

$$D[\mathcal{E}_{\mathcal{G}}] = \left\{ u \in C(\mathcal{G}) | \sup_{n \geq 0} \mathcal{E}^n_{\mathcal{G}}[u|_{V^n}] < +\infty \right\}.$$

Here for every $n \geq 0$ the set V^n is obtained by iteration as

$$V^n = \bigcup_{i|n} \psi_{i|n}(\Gamma). \tag{13}$$

The fractal $\mathcal{G}$ is the closure in $\mathbb{R}^2$ of the set $V^\infty = \bigcup_{n=0}^{+\infty} V^n$. We note that the functional (10) is non trivial, because it is finite on the domain $D[F] = \{u : u \in H^1(\Omega), u|_{\mathcal{G}} \in D[\mathcal{E}]\}$ which is dense in $L^2(\Omega)$ (see, *e.g.*, [13]). The functional F defines a densely defined self-adjoint operator $A = -\Delta_{\mathcal{G}}$ in the Hilbert space $L^2(\mathcal{G}, \mu_{\mathcal{G}})$, which takes the role of intrinsic Laplace operator in $\mathcal{G}$ with Neumann condition on Γ. The measure $\mu_{\mathcal{G}}$ is the (normalized) d_H-dimensional Hausdorff measure on $\mathcal{G}$, see Hutchinson [12].

The case of *Dirichlet* conditions, on both $\partial\Omega$ and Γ, is covered by the next result. The functional $F_n[u]$ of the previous theorem is now replaced by the functional

$$F_n[u] = F_\varepsilon^n[u] = \begin{cases} \int_\Omega a_{\varepsilon_n}^n(x,y)|\nabla u|^2 dxdy & \text{if } u \in D_0[F_{\varepsilon_n}^n] \\ +\infty & \text{if } u \in L^2(\Omega) \setminus D_0[F_{\varepsilon_n}^n] \end{cases} \tag{14}$$

where the domain $D_0[F_{\varepsilon_n}^n] \subset L^2(\Omega)$ is now the completion of $C_0^1(\Omega)$ in the norm $\|u\|_{D[F_{\varepsilon_n}^n]}$ given in (7). The limit functional is defined on $L^2(\Omega)$ by

$$F[u] = \begin{cases} \zeta^* \int_\Omega |\nabla u|^2 dxdy + \gamma^* \mathcal{E}[u|_{\mathcal{G}}] & \text{if } u \in H_0^1(\Omega), u|_{\mathcal{G}} \in D_0[\mathcal{E}_{\mathcal{G}}] \\ +\infty \quad \text{if } u \in L^2(\Omega) \setminus \{u : u \in H_0^1(\Omega), u|_{\mathcal{G}} \in D_0[\mathcal{E}_{\mathcal{G}}]\}. \end{cases} \tag{15}$$

The functional (15) is finite on the domain $D_0[F] = \{u : u \in H_0^1(\Omega), u|_{\mathcal{G}} \in D_0[\mathcal{E}_{\mathcal{G}}]\}$, where $D_0[\mathcal{E}_{\mathcal{G}}]$ is the subspace of all functions in $D[\mathcal{E}_{\mathcal{G}}]$ that vanish on Γ. Again, this functional is non trivial, because $D_0[F]$ is dense in $L^2(\Omega)$. The self-adjoint operator $A = -\Delta_{\mathcal{G}}$, defined now in the Hilbert space $L^2(\mathcal{G}, \mu_{\mathcal{G}})$ by the functional F, is the Laplace operator $-\Delta_{\mathcal{G}}$ in the fractal $\mathcal{G}$ with Dirichlet boundary condition on Γ.

The result in [30] is

Theorem 2.2 *Under the same scaling assumptions as in Theorem 2.1, the sequence of functionals F_n defined in (14) M-converges in $L^2(\Omega)$ to the functional F defined in (15) as $n \to +\infty$.*

The special scaling laws for the parameters imply in particular that the transversal thickness of the fibers tends to zero while their conductivity diverges to $+\infty$, the product of them remaining bounded, as $n \to +\infty$. In the same paper some cases where this condition is not satisfied are also studied.

We point out that in both theorems the asymptotic energy has two interacting components, the standard Dirichlet integral extended to the two dimensional domain Ω, and a lower-dimensional fractal energy term. Globally, the limit functional F defines a self-adjoint operator A in the space $L^2(\Omega)$. Formally, such operator A is given by the two dimensional Laplace operator Δ in the open set $\Omega \setminus \mathcal{G}$—with Neumann or Dirichlet boundary condition on $\partial\Omega$—together with the fractal-Laplacian $\Delta_{\mathcal{G}}$ on $\mathcal{G}$—with Neumann or Dirichlet condition on $\Gamma = \mathcal{G} \cap \partial\Omega$. The two operators are coupled by a second order transmission condition on $\mathcal{G}$. The condition states that the jump of the normal derivative of the function u from Ω across $\mathcal{G}$, taken on $\mathcal{G}$, equals the Laplacian $\Delta_{\mathcal{G}}$ acting on the trace of u on $\mathcal{G}$. In the case of Dirichlet condition on Γ, a fractal analogue of the Lions-Magenes trace space, mentioned earlier, comes into play. For a rigorous definition of the transmission problems when $\mathcal{G}$ is the von Koch curve we refer to [16] and [18].

As mentioned in the introduction, the convergence of the energy functionals implies the convergence of the spectral measures and of the spectral subspaces.

Theorem 2.3 *In the same assumptions of Theorem 2.1 and Theorem 2.2, for every $\lambda < \mu$ which are not in the point spectrum of the operator A in $L^2(\Omega)$, the projection operator $P^n((\lambda,\mu])$ of the spectral resolution P^n of the operator A^n in $L^2(\Omega)$ converges strongly in $L^2(\Omega)$ to the projection operator $P((\lambda,\mu])$ of the spectral resolution P of the operator A in $L^2(\Omega)$.*

This result follows from the convergence of the functionals, by applying Theorem 2.4.1 and its Corollary 2.7.1 from [24].

In the problems considered so far in this section the parameter ζ^* is positive. This is the case when the medium in which the fibers are embedded keeps finite positive viscosity up to the limit. Then, as seen before, the energy is only partially absorbed into the lower dimensional fractal inclusion. The *vanishing viscosity* case, when

$$\lim \zeta_n = \zeta^* = 0$$

has been considered in [29]. In this case the limit functional is composed *only* by the fractal energy term.

Such a collapse of geometry and energy on a lower dimensional fractal set is an interesting feature, both in fractal and PDEs theories. It shows, in particular, that fractal Laplacians can be obtained as the (spectral) limit of singular second order elliptic operators in divergence form.

In the vanishing viscosity case, however, there is a loss of coercivity as $n \to +\infty$. In fact, the uniform H^1 estimate that plays a basic role in the previous theorems, that is

$$c\|\nabla u\|^2_{L^2(\Omega)} \le F^n_{\varepsilon_n}[u]$$

with $c > 0$ independent of n, now fails, due to the vanishing of the coefficient ζ_n as $n \to +\infty$. The domains $D[F_n]$ of the functionals F_n and the domain $D[F]$ of the limit functional F are no more contained in the single Hilbert space $H = L^2(\Omega)$, which is the space where the convergence of the previous theorems takes place.

This difficulty has been overcome in [29], by relying on a generalization of the M-convergence of functionals to variable Hilbert spaces, developed by Kuwae-Shioya in [15]. Generally speaking, the convergence of the functionals takes now place in a larger Hilbert space, $\bigotimes_0^\infty H^n$.

In [29], for every $n \ge 0$ the following Hilbert space is considered:

$$H^n = L^2\big(\Omega, \mu^n_{\varepsilon_n}\big),$$

where the Borel measure $\mu^n_{\varepsilon_n}$ in Ω is defined by

$$\mu_n = \zeta_n \mathbf{1}_{\Omega \setminus \Sigma^n_{2\varepsilon_n}} \mathcal{L} + \mathbf{1}_{\Sigma^n_{2\varepsilon_n} \setminus \Sigma^n_{\varepsilon_n}} \mathcal{L} + \tau_n w^n_{\varepsilon_n} \mathbf{1}_{\Sigma^n_{\varepsilon_n}} \mathcal{L}.$$

Here $\mathcal{L}$ is the 2-dimensional Lebesgue measure; $0 < \zeta_n \le 1$ are the viscosity parameters and τ_n are scaling constants, depending on the fractal, that will be specified

later on. The functional F_n is now defined for each n in the spaces H^n as follows:

$$F_n[u] = F^n_{\varepsilon_n}[u] = \begin{cases} \int_\Omega a^n_{\varepsilon_n}(x,y)|\nabla u|^2 dxdy & \text{if } u \in D[F_n] \\ +\infty & \text{if } u \in L^2(\Omega,\mu_n) \setminus D[F_n] \end{cases}$$

where $a^n_{\varepsilon_n}\mathrm{Id}$ is again the coefficient matrix defined in (4) for $\varepsilon = \varepsilon_n$ and now the domain $D[F_n]$—$D[F_n] \subset L^2(\Omega,\mu_n) = H^n$—is the space of all functions $u \in L^2(\Omega,\mu_n)$ with distribution weak gradient in $L^2(\Omega,\mu_n)$. The functional F_n defines a regular, closed Dirichlet form in H^n. The generator of such a form is a self-adjoint operator $-A^n$ densely defined in the space H^n. The operator A^n is the positive-definite self-adjoint realization in the space $L^2(\Omega,\mu_n)$ of the second order elliptic operator in divergence form (1), with natural Neumann conditions on $\partial\Omega$. The spectrum of the operator $A^n + \mathrm{Id}_{H_n}$ is a point spectrum, with eigenvalues $\lambda^n_k \to +\infty$ as $k \to +\infty$.

The measures μ_n are the so-called *speed measures* of the Markov processes generated by $-A^n$. They replace, in the choice of the Hilbert space, the two dimensional Lebesgue measure of the non-vanishing viscosity case. The transmission condition at the interface of $\mathcal{G}$ and Ω is affected by this change.

We now summarize the assumptions in the present case. The coefficients $a^n_{\varepsilon_n}$ are defined as previously in (4), and they depend on the two sequences of constants ζ_n and γ_n. The constants N, α, δ and ρ—that depend on the fractal—are the same as specified before. As before, we also take

$$\varepsilon_n = \left(\frac{\rho}{N}\right)^n \omega_n, \quad (\text{with } \omega_n \to 0 \text{ as } n \to +\infty), \qquad \sigma_n = \left(\frac{\rho}{\alpha}\right)^n. \tag{16}$$

In addition, for the Sierpiński case considered in [29], we assume that

$$\tau_n = 3\left(\frac{\rho}{\alpha}\right)^n$$

(for other fractals, the numerical coefficient 3 may be replaced by another numerical constant depending on the cardinality of Γ). With this choice of the constants τ_n, it is proved in [29] that the measures μ_n weak* converge to the measure $\mu_{\mathcal{G}}$ as $n \to +\infty$, that is

$$\int_\Omega \phi d\mu_n \to \int_\Omega \phi d\mu_{\mathcal{G}}$$

as $n \to +\infty$, for every $\phi \in C(\bar{\Omega})$.

The result of [29] for the Sierpiński fractal is:

Theorem 2.4 *With the scaling constants ε_n, σ_n and τ_n specified before, let the constants $0 < \zeta_n \le 1$ and γ_n be such that $\lim \zeta_n = 0$ and $\lim \gamma_n = \gamma^* \in (0, +\infty)$ as $n \to \infty$. Then the sequence of functionals F_n in H^n M-converges (in the sense of*

Kuwae-Shioya) *to the functional*

$$F[u] = \begin{cases} \gamma^* \mathcal{E}_{\mathcal{G}}[u] & \text{if } u \in D[F] \\ +\infty & \text{if } u \in L^2(\mathcal{G}, \mu_{\mathcal{G}}) \setminus D[F] \end{cases} \tag{17}$$

with domain $D[F] = \{u \in L^2(\mathcal{G}, \mu_{\mathcal{G}}) : u \in D[\mathcal{E}_{\mathcal{G}}]\}$ *where* $\mathcal{E}_{\mathcal{G}}[u]$ *is the energy functional on the fractal* $\mathcal{G}$ *with domain* $D[\mathcal{E}_{\mathcal{G}}] \subset L^2(\mathcal{G}, \mu_{\mathcal{G}})$.

We note that, in the present context—that is, $H^n = L^2(\Omega, \mu_n)$, $H = L^2(\mathcal{G}, \mu_{\mathcal{G}})$—the M-convergence of the functionals F_n to the functional F, in the sense of Kuwae-Shioya, is defined as the usual M-convergence of functionals, provided strong and weak convergence of sequences of vectors are defined in the following way: a sequence of vectors $u_n \in H^n$ *converges strongly* to a vector $u \in H$ if there exists a sequence $\phi_m \in C(\bar{\Omega})$, such that $\|\phi_m - u\|_H \to 0$ as $m \to 0$, and

$$\lim_m \limsup_n \|\phi_m - u_n\|_{H^n} \to 0, \quad \text{as } n \to +\infty \text{ and } n \to 0.$$

The sequence $u_n \in H^n$ *converges weakly* to $u \in H$, if the inner product $(u_n, v_n)_{H^n}$ converge to the inner product $(u, v)_H$ for every v_n converging strongly to v as $n \to +\infty$.

Similarly as before, from the convergence of the functionals we get the convergence of the spectral measures, see Theorem 3.4 in [15]:

Theorem 2.5 *In the same assumptions of Theorem 2.4, for every* $\lambda < \mu$ *not in the point spectrum of the self-adjoint operator* $A = -\Delta_{\mathcal{G}}$ *in* $L^2(\mathcal{G}, \mu_{\mathcal{G}})$, *defined by* F, *the projection operator* $P^n((\lambda, \mu])$ *of the spectral resolution* P^n *of the self-adjoint operators* A^n *in* $L^2(\Omega, \mu_n)$, *defined by* F_n, *converges strongly to the projection operator* $P((\lambda, \mu])$ *of the spectral resolution* P *of the operator* A, *in the Kuwae-Shioya sense.*

In this statement, the strong convergence of the spectral projectors has to be intended according to the following general definition: a sequence of bounded operators B_n in H^n converges strongly to a bounded operator B in H if for every $u_n \in H^n$ converging strongly to $u \in H$ the sequence $B_n u_n \in H^n$ converges strongly to $Bu \in H$, with the strong convergence of vectors defined as before.

3 Elliptic Operators with Fractal Degeneracy

In this section we report on the recent papers by Capitanelli and Vivaldi, [6, 7]. The problem studied in these papers is the boundary approximation with suitable *insulating fibers* of Laplace equations in a domains bounded by four Koch curves.

The domain Ω_0 is now the square $\{(x, y) : 0 < x < 1, -1 < y < 0\}$, with vertices $A = (0, 0)$, $B = (1, 0)$, $C = (1, -1)$ and $D = (0, -1)$. On each one of the four sides

a Koch curve K_j, $j = 1, \ldots, 4$, is constructed, moving outward from the square. At the iteration n, the domain bounded by the four pre-fractal Koch curves K_j^n is denoted by Ω^n. For each n, and for every $0 < \varepsilon \leq \varepsilon_0 < 1/2$, the open set Ω^n is enlarged to become the open set

$$\Omega_\varepsilon^n = \bar{\Omega}^n \cup \Sigma_{j,\varepsilon}^n,$$

where for each j the set $\Sigma_{j,\varepsilon}^n$ is the open fibered neighborhood of K_j^n constructed by similarity from the initial reference fiber $\Sigma_{0,\varepsilon}^0$ already described in Sect. 1. However, since now the pre-fractal is on the boundary, we split the reference fiber in half, by keeping only the (open) half fiber that lays above the *x-axis*. The fibered set $\cup\Sigma_{j,\varepsilon}^n$ lies then externally to the domain Ω^n and is disjoint from $\bar{\Omega}^n$.

The conductivity coefficients of the fibered set $\Sigma_{j,\varepsilon}^n$ is defined, as in the previous section, in terms of the constants γ_n and σ_n and the functions w_ε^n. However, in the definition of w_ε^n, a substantial change is performed: in the definition of the conductivity of the reference fiber $\Sigma_{0,\varepsilon}$, the factor $|P - P^\perp|^{-1}$ is replaced by the factor $|P - P^\perp|$ (and the numerical coefficients are modified conveniently). With this change, the fibers present vanishing conductivity as ε tends to 0. The conductivity $a_\varepsilon^n(x, y)$ of the enlarged domain Ω_ε^n is then defined to be equal to $w_\varepsilon^n(x, y)$ if $(x, y) \in \Sigma_\varepsilon^n$, and equal to 1 if $(x, y) \in \bar{\Omega}^n$.

The spaces $H^1(\Omega_\varepsilon^n, w_\varepsilon^n)$ and $H_0^1(\Omega_\varepsilon^n, w_\varepsilon^n)$ are defined to be the completion of $C^1(\bar{\Omega}_\varepsilon^n)$ and $C_0^1(\Omega_\varepsilon^n)$, respectively, in the norm

$$\|u\|_{H^1(\Omega_\varepsilon^n, w_\varepsilon^n)} = \left\{ \int_{\Omega_\varepsilon^n} |u|^2 dxdy + \int_{\Omega_\varepsilon^n} |\nabla u|^2 w_\varepsilon^n dxdy \right\}^{\frac{1}{2}}.$$

By Ω^* we denote the unit disc with center at $P_0 = (1/2, 1/2)$. We then consider the following functionals:

$$F_n[u] = F_{\varepsilon_n}^n[u] = \begin{cases} \int_{\Omega_{\varepsilon_n}^n} a_{\varepsilon_n}^n(x, y) |\nabla u|^2 dxdy \\ \qquad \text{if } u \in L^2(\Omega^*) \text{ and } u|_{\Omega_\varepsilon^n} \in H_0^1(\Omega_\varepsilon^n, w_\varepsilon^n) \\ +\infty \quad \text{if } u \in L^2(\Omega^*) \text{ and } u|_{\Omega_\varepsilon^n} \notin H_0^1(\Omega_\varepsilon^n, w_\varepsilon^n). \end{cases}$$

By $\mu_{\partial\Omega}$ we denote the measure on $\partial\Omega$ such that the restriction of $\mu_{\partial\Omega}$ to each fractal component K_j of $\partial\Omega$ coincides with the Hausdorff measure μ_{K_j} of K_j, $j = 1, \ldots, 4$.

Then the following result is given in [6]:

Theorem 3.1 *Let us assume that $\gamma_n > 0$ $\gamma^* > 0$ and $\gamma_n \to \gamma^*$ as $n \to +\infty$. Let ε_n be an arbitrary sequence such that $\varepsilon_n \to 0$ as $n \to +\infty$. Then the functional F_n M-converge to the functional*

$$F[u] = \begin{cases} \int_\Omega |\nabla u|^2 dxdy + \gamma^* \int_{\partial\Omega} |u|^2 d\mu_{\partial\Omega} \quad \text{if } u \in L^2(\Omega^*) \text{ and } u|_\Omega \in H^1(\Omega) \\ +\infty \quad \text{if } u \in L^2(\Omega^*) \text{ and } u|_\Omega \notin H^1(\Omega). \end{cases}$$

We point out that the boundary value problem for the Laplace operator in Ω associated with the limit functional F implies, if $\gamma^* > 0$, a Robin type condition on $\partial\Omega$.

In [6], the case when $\gamma_n \to 0$—leading to Neumann boundary condition on $\partial\Omega$—and the case $\gamma_n \to +\infty$ with an additional assumption on the rate of convergence of $\varepsilon_n \to 0$—leading to a Dirichlet condition on $\partial\Omega$—are also studied, as well the generalization of the boundaries to the irregularly scaled Koch curves, or *Koch mixtures*, in the sense of [4] and [26].

4 Interfacial Heat Transmission

Two-dimensional second order transmission problems across a highly conductive layer of Koch type have been studied by Lancia, Vernole and co-authors in a series of recent papers, [8, 17, 19–22].

In reporting on this work in the context of this note, we confine ourselves mainly to the papers [8, 19, 20]. In [19] the authors obtain their first results on the heat transmission problem that we already mentioned in the Introduction. In particular, they show the existence and uniqueness of the strict solution for both the fractal and pre-fractal problems, moreover they study the regularity and the convergence of the solutions of the pre-fractal problems as the pre-fractal layer converges to the fractal set. In [8], the authors provide the finite element approximation for this kind of problems.

The pre-fractal transmission problems studied in [20] can be formally stated as follows:

$$(\overline{P_n}) \qquad \begin{cases} \frac{du_n(t,P)}{dt} - \Delta u_n(t,P) = f(t,P) & \text{in } [0,T]\times\Omega_n^i,\ i=1,2 \quad \text{(i)} \\ \frac{du_n}{dt} - \Delta_{K_n} u_n(t,P) = [\frac{\partial u_n(t,P)}{\partial \nu}] + f & \text{on } [0,T]\times K_n, \quad \text{(ii)} \\ u_n(t,P) = 0 & \text{on } [0,T]\times\partial\Omega, \quad \text{(iii)} \\ u_n^1(t,P) = u_n^2(t,P) & \text{on } [0,T]\times K_n, \quad \text{(iv)} \\ u_n(t,P) = 0 & \text{on } [0,T]\times\partial K_n \quad \text{(v)} \\ u_n(0,P) = 0 & \text{on } \Omega \quad \text{(vi)} \end{cases}$$

In this problem Ω is a rectangular domain, for example the open rectangle with vertices $A = (0, -\sqrt{3}/2)$, $B = (1, -\sqrt{3}/2)$, $C = (1, \sqrt{3}/2)$ and $D = (0, \sqrt{3}/2)$. The source term $f(t,P)$ is a given function in $C^\delta([0,T]; L^2(\Omega, m_n))$ with $\delta \in (0,1)$. For a fixed n, K_n is the pre-fractal Koch curve with endpoints are $(0,0)$ and $(1,0)$. The curve K_n separates Ω into two open subsets, Ω_n^1 and Ω_n^2. The restriction of u_n to Ω_n^i is denoted by u_n^i, $i = 1,2$. The piecewise-tangential Laplacian defined on the polygonal curve K_n is denoted by Δ_{K_n}. The jump of the normal derivatives across K_n is given by

$$\left[\frac{\partial u_n}{\partial \nu}\right] = \frac{\partial u_n^1}{\partial \nu_1} + \frac{\partial u_n^2}{\partial \nu_2},$$

where ν_i is the inward normal vector to the boundary of Ω_n^i.

Let us introduce the Hilbert space $L^2(\Omega, m_n)$, where

$$dm_n = dxdy + ds, \tag{18}$$

with inner product $(\cdot,\cdot)_{m_n}$ and norm $\|u\|_{2,m_n} = (\int_{\Omega_n} |u|^2 dxdy + \int_{K_n} |u|^2 ds)^{\frac{1}{2}}$ and the forms

$$E^{(n)}(u_n, u_n) = \int_{\Omega} |\nabla u_n|^2 dxdy + \int_{K_n} |\nabla_\tau \gamma_0 u_n|^2 ds, \tag{19}$$

defined on the domain

$$V(\Omega, K_n) = \left\{u_n \in H_0^1(\Omega) : \gamma_0 u_n \in H_0^1(K_n)\right\}. \tag{20}$$

In (20), $H_0^1(\Omega)$ denotes the usual Sobolev space in Ω, $H_0^1(K_n)$ the trace space on K_n and $\gamma_0 u_n$ is the trace of u_n on K_n (denoted simply by u_n below). Moreover, the second integral at the right-hand side of (19) is defined piece-wise by

$$\int_{K_n} |\nabla_\tau \gamma_0 u_n|^2 ds = \sum_{M \in F^n} \int_M |\nabla_\tau \gamma_0 u_n|^2 ds,$$

where the sum is taken over the segments M that compose K_n, ∇_τ is the tangential derivative along M. The measure ds is the one-dimensional arc length measure on K_n. This integral expresses the energy $E_{K_n}(\cdot,\cdot)$ of the curve K_n. The space $V(\Omega, K_n)$ given by (20) is a Hilbert space under the norm

$$\|u_n\|_{V(\Omega,K_n)} = \left\{E^{(n)}(u_n, u_n)\right\}^{1/2}. \tag{21}$$

Moreover, for each $n \in \mathbb{N}$, $E^{(n)}(\cdot,\cdot)$, with domain $V(\Omega, K_n)$, is a regular, strongly local Dirichlet form in $L^2(\Omega)$ and in $L^2(\Omega, m_n)$, respectively.

In [20], Problem $(\overline{P_n})$ is dealt with by semigroup methods. More precisely, for every fixed n the following abstract Cauchy problem is studied

$$(P_n) \quad \begin{cases} \frac{\mathrm{d}u_n(t)}{\mathrm{d}t} = A_n\, u_n(t) + f(t), \quad 0 \le t \le T \\ u_n(0) = 0 \end{cases} \tag{22}$$

where $A_n : \mathcal{D}(A_n) \subset L^2(\Omega, m_n) \to L^2(\Omega, m_n)$ is the generator associated to the energy form $E^{(n)}$,

$$E^{(n)}(u_n, v) = -\int_{\Omega} A_n u_n v\, dm_n, \quad u_n \in \mathcal{D}(A_n),\ v \in V(\Omega, K_n).$$

The following existence and uniqueness result is then obtained

Theorem 4.1 *Let $0 < \delta < 1$, $f \in C^\delta([0,T], L^2(\Omega, m_n))$, and let*

$$u_n(t) = T_t u_0 + \int_0^t T_n(t-s) f(s) \mathrm{d}s \quad \textit{for every } n \in \mathbb{N}, \tag{23}$$

where $T_n(t)$ is the analytic semigroup generated by A_n. Then u_n is the unique "strict" solution of (P_n). Moreover,

$$\|u_n\|_{C^1([0,T],L^2(\Omega,m_n))} + \|u_n\|_{C^0([0,T],\mathcal{D}(\mathrm{A_n}))} \le c\|f\|_{C^\delta([0,T],L^2(\Omega,m_n))}, \tag{24}$$

where c is a constant independent of n.

The solution of the abstract Cauchy problem (P_n) is the "strong" solution of Problem $(\overline{P}_n)$, as described by this result in [8]:

Theorem 4.2 *Let u_n be the solution of Problem (P_n). For every fixed $t \in [0,T]$ we have*

$$\begin{cases} \frac{\mathrm{d}u_n(t,P)}{\mathrm{d}t} - \Delta u_n(t,P) = f(t,P), & \text{for } P \in \Omega_n^i, i = 1,2, \\ \frac{\partial u_n^i}{\partial \nu_i} \in L^2(K_n), & i = 1,2, \\ \frac{\mathrm{d}u_n}{\mathrm{d}t} - \Delta_{K_n} u_n|_{K_n} = [\frac{\partial u_n}{\partial \nu}] + f, & \text{in } L^2(K_n), \\ u_n(t,P) = 0, & \text{for } P \in \partial\Omega. \end{cases} \tag{25}$$

Moreover, $\frac{\partial u_n^i}{\partial \nu_i} \in C([0,T], L^2(K_n))$, $i = 1,2$.

In [19] and [20] the following regularity result is also obtained

Theorem 4.3 *For any fixed $t \in [0,T]$, $u_n^1 \in H^{2,\alpha_1}(\Omega_n^1)$ with $\alpha_1 > \frac{2}{5}$; $u_n^2 \in H^{2,\alpha_2}(\Omega_n^2)$ with $\alpha_2 > \frac{1}{4}$, and $u_n \in C^0(\overline{\Omega})$, $u_n|_{K_n} \in H^2(K_n)$.*

The definition of the weighted Sobolev spaces $H^{2,\alpha^i}(\Omega_n^i)$ is rather delicate. If $\mathcal{D}$ is a non-convex polygonal domain in $\mathbb{R}^2$ and $\alpha > 0$, the space $H^{2,\alpha}(\mathcal{D})$ tis defined to be the space

$$H^{2,\alpha}(\mathcal{D}) = \left\{ v \in H^1(\mathcal{D}) : r^\alpha \cdot D^\beta v \in L^2(\mathcal{D}),\ \beta = (\beta_1,\beta_2) \in \mathbb{N}\times\mathbb{N} \text{ s.t. } |\beta| = 2 \right\},$$

equipped with the norm

$$\|v\|_{H^{2,\alpha}(\mathcal{D})} := \left\{ \|v\|^2_{H^1(\mathcal{D})} + \sum_{|\beta|=2} \left\| r^\alpha \cdot D^\beta v \right\|^2_{L^2(\mathcal{D})} \right\}^{1/2}.$$

The delicate point in this definition is the construction of the weight function $r : \mathcal{D} \to \mathbb{R}_+$, that we now describe. Let $\{P_j, 1 \le j \le \mathcal{N}\}$ be the set of vertices of $\mathcal{D}$. For $j = 1,\dots,\mathcal{N}$, let θ_j be the interior angle of $\mathcal{D}$ at P_j. Let $\mathcal{R}$ be the set of the indices $\{j = 1,\dots,\mathcal{N} : \frac{\pi}{\theta_j} < 1\}$ and let $\mathcal{Q} = \{P_j\}_{j\in\mathcal{R}}$ be the subset of the vertices with reentrant angles θ_j (these are the points where the solutions are singular). We set $\eta := \{\frac{1}{4} \cdot \min |P_j - P_k|;\ j,k \in \mathcal{R},\ j \ne k\}$ and arbitrarily choose $0 < \varepsilon < \eta$. For j in $\mathcal{R}$, we define $r_j(P) := |P - P_j|$ for all P in $\mathrm{B}_\varepsilon(P_j) = \{P \in \mathcal{D} : |P - P_j| < \varepsilon\}$. We then define the function $r : \mathcal{D} \to \mathbb{R}_+$ by putting $r(P) := r_j(P)$, for all $P \in \mathrm{B}_\varepsilon(P_j)$ and j in $\mathcal{R}$, and $r(P) := 1$ for all $P \in \mathcal{D} \setminus \bigcup_{j\in\mathcal{R}} \mathrm{B}_{2\varepsilon}(P_j)$.

We conclude this section with some remarks on the numerical approximation of these problems, reporting mainly on the papers [8] and [9].

The pre-fractal curve K_n induces a natural triangulation $\mathcal{T}_{n,h}$ of the domain Ω, in which the vertices of K_n belongs to the set of nodes of $\mathcal{T}_{n,h}$. Starting with this triangulation, a mesh refinement process is given, that generates a regular and conformal family of finer triangulations $\{\mathcal{T}_{n,h}\}$.

The need for such refined triangulations comes from the presence of reentrant angles in the boundaries of the domains Ω_n^1 and Ω_n^2, which were previously described. As already mentioned, the solution u_n is singular at these angles, indeed u_n is not in the Sobolev space $H^2(\Omega_n^i)$, as it is the case of a smoothly bounded domain. Instead, as seen with Theorem 4.3, $u_n^i \in H^{2,\alpha_i}(\Omega_n^i)$, $i = 1, 2$, with $\alpha_1 > \frac{2}{5}$ and $\alpha_2 > \frac{1}{4}$. In view of these singularities, in order to get optimal rate of convergence for the finite element approximations the triangulation of the domains Ω_n^i must be refined, according to the conditions introduced in this regard by Grisvard in [11].

The authors are able to implement Grisvard's conditions by satisfying at the same time an additional important property for their refinements. The refined meshes are constructed as a "nested" sequence of meshes, *i.e.*, all the nodes of $\mathcal{T}_{n,h}$ belong also to $\mathcal{T}_{n+1,h}$. This property is of course of great help when the numerical approximation is carried out at various levels of the fractal iteration. We refer to [8] and [9] for more details. We also point out that in [9] more complicated boundaries, made by suitable mixtures of Koch curves, are also considered.

With the appropriate triangulations at hand, the numerical approximation of the problem $(\overline{P_n})$ is carried out in two steps. In the first step the semi-discrete problem, obtained by discretizing with a Galerkin method only the space variable, is considered. The following a priori error estimates of the order of convergence is then obtained

Theorem 4.4 *Let $u_n(t)$ be the solution of $(\overline{P_n})$, $u_n^i(t)$ be the restriction to Ω_n^i of $u_n(t)$, for $i = 1, 2$, and $u_{n,h}(t)$ be the semi-discrete solution. Then for every $t \in [0, T]$ we have*

$$\begin{aligned}&\left\|u_n(t) - u_{n,h}(t)\right\|_{2,m_n}^2 + \int_0^t \left\|u_n(\tau) - u_{n,h}(\tau)\right\|_{V(\Omega,K_n)}^2 d\tau \\ &\quad \le ch^2\left(\int_0^t \left\|f(\tau)\right\|_{2,m_n}^2 d\tau\right)\end{aligned}$$

where c is a suitable constant independent of h.

In the second step the fully discretized problem is considered. By applying a finite difference scheme on the time variable, the so-called θ method, an error estimate between the semi-discrete solution $u_{n,h}(t_l)$ and the fully discrete solution $u_{n,h}^l$ is obtained.

From this estimate and from Theorem 4.4, they finally get

Theorem 4.5 *Assume that* $f \in C^{\delta}([0,T]; L^2(\Omega, m_n))$ *and* $\frac{\partial f}{\partial t} \in L^2([0,T] \times \Omega, dt \times dm_n)$. *Let* n *be fixed and let* $u_n(t)$ *be the solution of problem* $(\overline{P_n})$, $u^l_{n,h}$ *be the fully discretized solution, as given by the* θ*-method with* $\frac{1}{2} \leq \theta \leq 1$. *Then,*

$$\left\| u_n(t_l) - u^l_{n,h} \right\|^2_{2,m_n} \leq ch^2 \left(\int_0^T \left\| f(\tau) \right\|^2_{2,m_n} d\tau \right) + C_\theta \triangle t^2 \cdot \left(\left\| f(0) \right\|^2_{2,m_n} + \int_0^T \left\| \frac{\partial f}{\partial \tau}(\tau) \right\|^2_{2,m_n} d\tau \right).$$

A final remark about future research. In all the problems discussed in this paper an important question remains to be investigated, namely, to obtain some quantitative estimate for the asymptotic fractal limit. Such estimates should reflect the stability properties of the problem at hand in presence of the wild changes in the geometry. The very nature of the estimates—whether they can be stated in suitable function spaces or they are just of scalar energy kind—must be understood, in each one of the special cases described before.

Acknowledgements This work was supported by the NSF grant No. 1109356.

The author wishes to thank the authors of the papers presented in this note for their kindness in providing him with the preprints of their work.

The author wishes also to thank the Colleagues in Pavia who organized the Conference for their invitation and hospitality and the Editors of the Volume of the conference for the opportunity given to him of contributing this paper.

References

1. Achdou, Y., Sabot, C., Tchou, N.: A multiscale numerical method for Poisson problems in some ramified domains with fractal boundary. Multiscale Model. Simul. **5**(3), 828–860 (2006)
2. Attouch, H.: Variational Convergence for Functions and Operators. Pitman, London (1984)
3. Bagnerini, P., Buffa, A., Vacca, E.: Finite elements for a prefractal transmission problem. C. R. Math. **342**(3), 211–214 (2006)
4. Barlow, M.T., Hambly, B.M.: Transition density estimates for Brownian motion on scale irregular Sierpiński gaskets. Ann. Inst. Henri Poincaré **33**(5), 531–557 (1997)
5. Cannon, J.R., Meyer, G.H.: On a diffusion in a fractured medium. SIAM J. Appl. Math. **3**, 434–448 (1971)
6. Capitanelli, R., Vivaldi, M.A.: Insulating layers and Robin problems on Koch mixtures. J. Differ. Equ. **251**(4–5), 1332–1353 (2011)
7. Capitanelli, R., Vivaldi, M.A.: On the Laplacean transfer across a fractal mixture Asymptot. Anal. doi:10.3233/ASY-2012-1149
8. Cefalo, M., Dell'Acqua, G., Lancia, M.R.: Numerical approximation of transmission problems across Koch-type highly conductive layers. Appl. Math. Comput. **218**(9.1), 5453–5473 (2012)
9. Cefalo, M., Lancia, M.R., Liang, H.: Heat flow problems across fractal mixtures: regularity results of the solutions and numerical approximation (to appear)
10. Evans, E.: Extension operators and finite elements for fractal boundary value problems. PhD thesis, Worcester Polytechnic Institute, USA (2011)
11. Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Pitman, London (1985)
12. Hutchinson, J.E.: Fractals and self-similarity. Indiana Univ. Math. J. **30**, 713–747 (1981)

13. Jonsson, A.: Dirichlet forms and Brownian motion penetrating fractals. Potential Anal. **13**, 69–80 (2000)
14. Jonsson, A., Wallin, H.: Function Spaces on Subsets of $\mathbb{R}^n$, Part 1. Math. Reports, vol. 2. Harwood Academic, London (1984)
15. Kuwae, K., Shioya, T.: Convergence of spectral structures: a functional analytic theory and its applications to spectral geometry. Commun. Anal. Geom. **11**(4), 599–673 (2003)
16. Lancia, M.R.: A transmission problem with a fractal interface. Z. Anal. Anwend. **21**, 113–133 (2002)
17. Lancia, M.R., Vacca, E.: Numerical approximation of heat flow problems across highly conductive layers. In: Silhavy, M. (ed.) Mathematical Modeling of Bodies with Complicated Bulk and Boundary Behaviour. Quaderni di Matematica, pp. 57–77. Dept. Me. Mo. Mat., U. Roma La Sapienza, Roma (2007)
18. Lancia, M.R., Vivaldi, M.A.: Asymptotic convergence for energy forms. Adv. Math. Sci. Appl. **13**, 315–341 (2003)
19. Lancia, M.R., Vernole, P.: Convergence results for parabolic transmission problems across highly conductive layers with small capacity. Adv. Math. Sci. Appl. **16**, 411–445 (2006)
20. Lancia, M.R., Vernole, P.: Irregular heat flow problems. SIAM J. Math. Anal. **42**(4), 1539–1567 (2010)
21. Lancia, M.R., Vernole, P.: Semilinear evolution transmission problems across fractal layers. Nonlinear Anal., Theory Methods Appl. **75**, 4222–4240 (2012)
22. Lancia, M.R., Vernole, P.: Semilinear fractal problems: Approximation and regularity results. Nonlinear Anal., Theory Methods Appl. Article first published online 2 Nov 2012. doi:10.1016/j.na2012.08.020
23. Liang, H.: On the construction of certain fractal mixtures. MS Thesis, Worcester Polytechnic Institute (2010)
24. Mosco, U.: Composite media and asymptotic Dirichlet forms. J. Funct. Anal. **123**(2), 368–421 (1994)
25. Mosco, U.: Variational fractals. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (4) **XXV**, 683–712 (1997)
26. Mosco, U.: Harnack inequalities on scale irregular Sierpiński gaskets. In: Birman, et al. (eds.) Nonlinear Problems in Mathematical Physics and Related Topics II, pp. 305–328. Kluwer Academic/Plenum, New York (2002)
27. Mosco, U., Vivaldi, M.A.: An example of fractal singular homogenization. Georgian Math. J. **14**(1), 169–194 (2007)
28. Mosco, U., Vivaldi, M.A.: Fractal reinforcement of elastic membranes. Arch. Ration. Mech. Anal. **194**, 49–74 (2009)
29. Mosco, U., Vivaldi, M.A.: Vanishing viscosity for fractal sets. Discrete Contin. Dyn. Syst. **28**(3), 1207–1235 (2010)
30. Mosco, U., Vivaldi, M.A.: Thin fractal fibers. Math. Methods Appl. Sci. Article first published online 14 Jun 2012. doi:10.1002/mma.1621
31. Pham Huy, H., Sanchez Palencia, E.: Phénomènes des transmission à travers des couches minces de conductivité élevée. J. Math. Anal. Appl. **47**, 284–309 (1974)
32. Vacca, E.: Galerkin approximation for highly conductive layers. PhD Thesis, Dept. Me. Mo. Mat., U. Roma La Sapienza (2005)
33. Wasyk, R.: Numerical solution of a transmission problem with pre fractal interface. PhD Thesis, Worcester Polytechnic Institute (2007)

AFEM for Geometric PDE: The Laplace-Beltrami Operator

Andrea Bonito, J. Manuel Cascón, Pedro Morin, and Ricardo H. Nochetto

Abstract We present several applications governed by geometric PDE, and their parametric finite element discretization, which might yield singular behavior. The success of such discretization hinges on an adequate variational formulation of the Laplace-Beltrami operator, which we describe in detail for polynomial degree 1. We next present a complete a posteriori error analysis which accounts for the usual PDE error as well as the geometric error induced by interpolation of the surface. This leads to an adaptive finite element method (AFEM) and its convergence. We discuss a contraction property of AFEM and show its quasi-optimal cardinality.

1 Introduction

Besides its intrinsic interest in differential geometry [30, 31, 56], the Laplace-Beltrami operator (or surface Laplacian) has received a great deal of attention also

In memory of Enrico Magenes.

A. Bonito
Department of Mathematics, Texas A&M University, 3368 TAMU, College Station, TX 77843-3368, USA
e-mail: bonito@math.tamu.edu

J.M. Cascón
Departamento de Economía e Historia Económica, Universidad de Salamanca, Salamanca 37008, Spain
e-mail: casbar@usal.es

P. Morin
Instituto de Matemática Aplicada del Litoral (IMAL), Güemes 3450 and Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, 3000 Santa Fe, Argentina
e-mail: pmorin@santafe-conicet.gov.ar

R.H. Nochetto (✉)
Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA
e-mail: rhn@math.umd.edu

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*, Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_15,

in the applied and numerical communities. Basic geometric partial differential equations (PDE) such as the mean curvature flow and surface diffusion appear naturally in materials science modeling [54], whereas Willmore flow is a building block in the dynamics of membranes governed by bending energy [34]. This article is about applications, formulation, Galerkin approximation, and adaptivity for a PDE on a surface γ governed by the Laplace-Beltrami operator Δ_γ, such as

$$-\Delta_\gamma u = f. \tag{1}$$

One of the major goals of this paper is the design and analysis of *parametric* adaptive finite element methods (AFEM) for (1) of polynomial degree 1. Our discussion is based on [17].

The first FEM for the Laplace-Beltrami operator on parametric surfaces is due to G. Dziuk [35], who also developed an optimal *a priori* error analysis accounting for the approximation of the surface and PDE by piecewise linear polynomials. This seminal work was followed by parametric FEM for time dependent problems such as the mean curvature flow [36], capillary surfaces [2], surface diffusion [5, 7], Willmore flow [7, 15, 37, 50], fluid biomembranes [16], and fluid membranes with orientational order [9, 10]. The analysis of these methods is largely open, except for graphs [4, 24–27]. We refer to the survey by K. Deckelnick, G. Dziuk, and Ch. Elliott [26] for some of the early work, including level set and phase field approaches.

A. Demlow and G. Dziuk gave the first *a posteriori* error analysis for piecewise linear polynomials [29], and later A. Demlow extended it to higher polynomial degree [28]. This extension is important in light of applications in fluid dynamics [2] and biomembrane dynamics [15, 16]. O. Lakkis and R.H. Nochetto formulated an a posteriori error analysis for the mean curvature flow of graphs in [44].

Even though *adaptivity theory* for linear elliptic PDE on flat domains in any dimensions and the energy norm is now mature [21, 47, 48, 52], much less is known for elliptic problems on manifolds; we refer to the survey [49] for the state of the art of AFEM on flat domains. For the Laplace-Beltrami operator on graphs we mention the convergence theory of K. Mekchay, P. Morin and R.H. Nochetto [46], whereas for general parametric surfaces and polynomial degree we are only aware of [17]. We expose here results from [17] and restrict them to the particular case of polynomial degree 1 for the sake of clarity.

The purpose of this paper is threefold. We first discuss in Sect. 2 several applications of the Laplace-Beltrami operator we have recently developed. This serves as a motivation for the rest of the paper as well as illustration of the significance of adequate formulations and discretizations of rather complex problems which look seemingly untractable. We next discuss parametric FEM for (1) on piecewise C^1 surfaces which are merely globally Lipschitz. This is inspired by singularities observed in geometric flows, such as pinching [5–7, 15], point defects [9, 10], and line tension [38]. This in turn makes it unfeasible to use the signed distance function as in [28, 29, 35]. Our approach, developed in Sects. 3 and 4, allows for kinks aligned with the initial mesh, and yields optimal convergence rates even for surfaces which are not piecewise C^2. Our third goal is to present a rather complete discussion of adaptivity theory for AFEM on surfaces. The algorithm reads

AFEM: Given an initial surface-mesh pair $(\Gamma_0, \mathcal{T}_0)$, and parameters $\varepsilon_0 > 0$, $0 < \rho < 1$, and $\omega > 0$, set $k = 0$ and iterate

$$\begin{aligned}
&\left[\mathcal{T}_k^+, \Gamma_k^+\right] = \mathsf{ADAPT_SURFACE}(\mathcal{T}_k, \omega\varepsilon_k)\\
&[\mathcal{T}_{k+1}, \Gamma_{k+1}] = \mathsf{ADAPT_PDE}\left(\mathcal{T}_k^+, \varepsilon_k\right)\\
&\varepsilon_{k+1} = \rho\varepsilon_k; \quad k = k+1.
\end{aligned}$$

AFEM consists of two main modules: ADAPT_PDE is the usual adaptive cycle for flat domains driven by the a posteriori PDE error estimator, whereas ADAPT_SURFACE is a new module that accounts for and controls surface interpolation error. In Sect. 5 we discuss the a posteriori error analysis for (1) on parametric surfaces, with emphasis on C^1 parametric representations $\mathcal{X} : \Omega \to \mathbb{R}^{d+1}$ of γ and their piecewise linear interpolants $\mathcal{F}_\mathcal{T} : \Omega \to \mathbb{R}^{d+1}$, which describes the polyhedral counterpart $\Gamma = \mathcal{F}_\mathcal{T}(\Omega)$ of γ; hereafter $\Omega \subset \mathbb{R}^d$ is the parametric domain. This interpolation is governed by the *geometric error estimator*

$$\lambda_\Gamma := \left\| \nabla\left(\mathcal{X} - \mathcal{F}_\mathcal{T}\right) \right\|_{L^\infty(\Omega)}. \tag{2}$$

The module ADAPT_SURFACE guarantees that its output satisfies $\lambda_{\mathcal{T}_k^+} \leq \omega\varepsilon_k$, with ω a parameter small relative to 1. This is critical for ADAPT_PDE to contract, a fundamental property of AFEM shown in Sect. 7. We embark on the study of cardinality of AFEM in Sect. 8: we first prove that AFEM delivers the best asymptotic convergence rate possible for the given regularity of data γ, f and solution u (Theorem 8.2), and secondly we construct a greedy algorithm that realizes ADAPT_SURFACE (Proposition 8.1). The role of ω is crucial for the theory of Sects. 7 and 8. We conclude in Sect. 9 with a computational investigation showing that ω must be small indeed to achieve optimal performance of AFEM.

2 Motivation: Geometric PDE

The Laplace-Beltrami operator is ubiquitous in applications involving surfaces that evolve and/or are the domain of an underlying PDE. In order to motivate the study of this operator we mention a few applications where it appears naturally.

2.1 Biomembranes: Modeling and Simulations

Predicting the shape of a cell bounded by a lipid bilayer membrane has inspired a significant body of research in the past forty years ranging from purely mechanical descriptions to advanced mathematical analysis. We consider the Helfrich model for

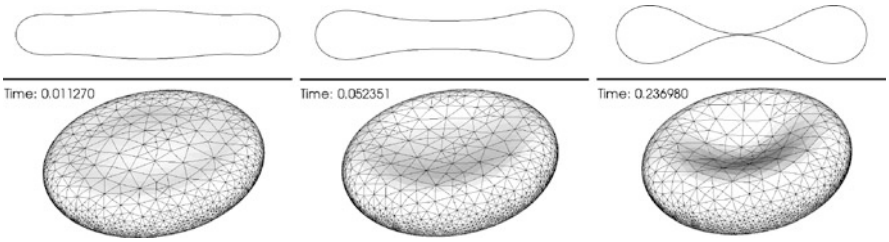


Fig. 1 Evolution of an initial axisymmetric ellipsoid of aspect ratio $5 \times 5 \times 1$. For each frame the picture on the bottom is a 3D view of the surface mesh and that on the top is a 2D cut through a symmetry plane. The equilibrium is characterized by the formation of an extreme depression of the center to the point of almost pinching (red blood cell). During the evolution the thickening of the outer circular edge occurs faster than the motion on the center, producing a depressed circular ring in between the outer edge and the center (*first frame*). This in turn is responsible for the appearance of a center bump instead of a depression. Later the evolution continues to squeeze this bump to a depression at the expense of more thickening and rounding of the outer circular edge

geometric biomembranes [41], which associates to a closed surface γ, describing the biomembrane, the *bending* (*or Willmore*) energy

$$J(\gamma) = \frac{1}{2}\int_\gamma (H - H_0)^2. \tag{3}$$

Hereafter H stands for the *mean curvature* of γ and H_0 is the *spontaneous curvature* induced by the surrounding medium.

Fluid Membranes

We start with $H_0 = 0$. The first variation (or shape derivative) of $J(\gamma)$, subject to volume and area constraints, is given in strong form by [32, 51, 56]

$$\delta_\gamma J(\gamma) = \left(\Delta_\gamma H + \frac{1}{2}H^3 - 2\kappa H\right)\boldsymbol{\nu} + (\lambda H \boldsymbol{\nu} + p\boldsymbol{\nu}), \tag{4}$$

where κ is the Gaussian curvature of γ, and λ, p are the Lagrange multipliers for the area and volume constraints, respectively. It is important to notice that $\delta_\gamma J(\gamma)$ is a vector field perpendicular to γ because $\boldsymbol{\nu}$ is the unit normal to γ. A (geometric) gradient flow consists of deforming γ in the direction opposite to the shape gradient, namely prescribing a vector velocity $\mathbf{v}$ to γ according to

$$\mathbf{v} = -\delta_\gamma J(\gamma). \tag{5}$$

This flow decreases the energy $J(\gamma)$ while keeping area and volume constant, and thus leads to equilibrium configurations such as that in Fig. 1, which mimics a *red blood cell*. The simulations in Fig. 1 were performed with the finite element

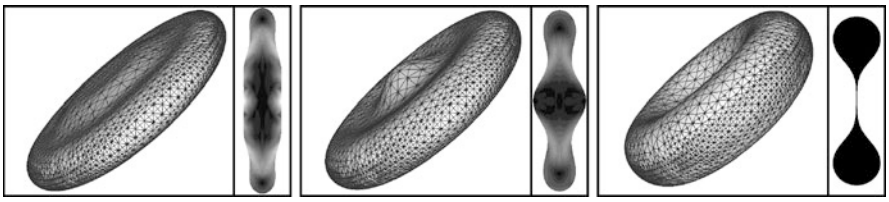


Fig. 2 Evolution of a fluid membrane with initial axisymmetric ellipsoidal shape of aspect ratio $5 \times 5 \times 1$ and final shape similar to a red blood cell. Each frame shows the membrane mesh and a symmetry cut along a big axis. The fluid flow is quite complex, creating first a bump in the middle and next moving towards the circumference and producing a depression in the center with flat pinching profile. The inertial effects are due to unrealistic physical parameters

Fig. 3 Comparison of final configuration of the geometric biomembrane of Fig. 1 and the fluid biomembrane of Fig. 2 with unrealistic (*left*) and realistic (*right*) physical parameters. For the latter the inertial effects are not significant and the purely geometric evolution is similar to the fluid driven one. The pinching on the left occurs with a much flatter and thinner neck in the center and thicker torus outside

method of A. Bonito, R.H. Nochetto, and M.S. Pauletti [15], which replaces H in (4) by the vector curvature $\mathbf{H} = H\boldsymbol{\nu}$ (see also Sect. 2.2).

We now consider the more physically realistic model that couples the membrane with a fluid. In order to do this, we assume the simplest situation in which the fluid is Newtonian, and thus is governed by the Navier-Stokes equation for incompressible fluids in the deformable domain Ω_t

$$\begin{aligned} \rho D_t \mathbf{v} - \operatorname{div}\big(-pI + \mu D(\mathbf{v})\big) = 0 \quad &\text{in } \Omega_t, \\ \operatorname{div} \mathbf{v} = 0 \quad &\text{in } \Omega_t, \end{aligned} \tag{6}$$

where $D(\mathbf{v}) = \frac{1}{2}(\nabla \mathbf{v} + \nabla \mathbf{v}^T)$ is the symmetric part of the gradient and $\Sigma = -pI + \mu D(\mathbf{v})$ is the Cauchy stress tensor. The membrane interacts with the fluid only through the boundary condition, which represents a balance of forces at the interface $\gamma = \gamma_t = \partial \Omega_t$:

$$\Sigma \boldsymbol{\nu} = k \delta_\gamma J(\gamma), \tag{7}$$

where k is the membrane bending rigidity coefficient. In [16] A. Bonito, R.H. Nochetto, and M.S. Pauletti couple the FEM of [15] with a Taylor-Hood discretization of (6) in an ALE framework involving a semi-implicit Euler method in time. Figure 2 displays the complex behavior of the fluid membrane and quite noticeable inertial effects, which lead to a more singular pinching than in Fig. 1. We give a comparison in Fig. 3.

Director Fields on Flexible Surfaces

The orientation of the bilipids is about 32° relative to the unit normal to γ for living cells. In order to describe this situation we consider the simple model introduced by S. Bartels, G. Dolzmann, R.H. Nochetto, and A. Raisch [10], which is in turn inspired on the model by M. Laradji and O.G. Mouritsen [45] for flat membranes. The starting point is to modify the energy (3) to incorporate the effect of a director field $\mathbf{n}$ so that

$$J(\gamma, \mathbf{n}) = \frac{1}{2}\int_\gamma |\operatorname{div}_\gamma \boldsymbol{\nu} - \delta \operatorname{div}_\gamma \mathbf{n}|^2 + \frac{\lambda}{2}\int_\gamma |\nabla_\gamma \mathbf{n}|^2 + \frac{1}{2\varepsilon}\int_\gamma f(\mathbf{n}\cdot\boldsymbol{\nu}), \tag{8}$$

with $|\mathbf{n}| = 1$ everywhere in γ. Here $\operatorname{div}_\gamma$, ∇_γ stand for the tangential divergence and gradient to γ, $H = -\operatorname{div}_\gamma \boldsymbol{\nu}$, and $\delta, \lambda > 0$. We thus see that $H_0 = -\delta \operatorname{div}_\gamma \mathbf{n}$ acts as a spontaneous curvature term induced by the director field $\mathbf{n}$. The function $f(x) := (x^2 - \xi_0^2)^2$ in the last term of (8) penalizes the deviation of the angle between $\mathbf{n}$ and $\boldsymbol{\nu}$ from $\arccos \xi_0$. It is worth stressing now that if this angle were constant everywhere on γ, then the projection of $\mathbf{n}$ on γ would have a constant length, which in turn would lead to the creation of defects (or singularities) of $\mathbf{n}$. This is due to the topological obstruction that there cannot be a smooth tangential vector field with nonzero constant length defined on a closed surface. Therefore the study of defects and their influence on membrane shape becomes an intriguing matter.

This is precisely what has been accomplished in [10], via an L^2-gradient flow (or relaxation dynamics) for $J(\gamma, \mathbf{n})$:

$$\mathbf{v} = -\delta_\gamma J(\gamma, \mathbf{n}), \qquad \partial_t \mathbf{n} = -\delta_n J(\gamma, \mathbf{n}), \tag{9}$$

where $\mathbf{v}$ is the velocity of γ. The expression of $\delta_\gamma J(\gamma, \mathbf{n})$, the first variation of J with respect to γ (or shape derivative) is now much more involved than (4), whereas $\delta_n J(\gamma, \mathbf{n})$ is rather simple; we refer to [10] for details. This dynamics involves again the Laplace-Beltrami operator Δ_γ.

We display in Fig. 4 the evolution of a sphere γ (first row) along with the director field $\mathbf{n}$ on a plane cutting through north and south poles. The initial director field $\mathbf{n}_0$ has a couple of defects $\pm e^{i\theta}$ of degree $+1$, which persist through the evolution and lead to the formation of cone-like singularities at the poles, one pointing inwards (north pole) and the other outwards (south pole). This configuration shows some analogies to echinocyte shapes observed in lab experiments [42]. We refer to [10] for other examples and discussion, including defects of degree ± 1.

2.2 The Laplace-Beltrami Operator and Curvature

The Laplace-Beltrami operator makes yet another fundamental appearance in the definition and calculation of *curvature*. If $\mathbf{x}$ is the identity on γ, then the following

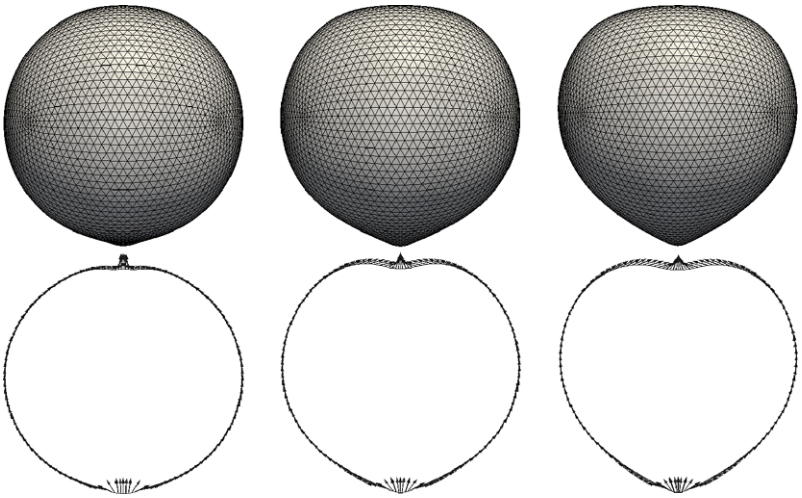


Fig. 4 Biomembrane case with inward and outward pointing defects of positive degree one: Snapshots of the surface and the director field along a (deformed) geodesic through the north and south pole after $n = 50, 500, 1400$ time steps. The surface develops inward and outward cones at the poles while the director field remains nearly unchanged during the evolution

relation for the vector curvature $\mathbf{H} = H\boldsymbol{\nu}$ is well known in differential geometry [30, 31]:

$$\mathbf{H} = -\Delta_\gamma \mathbf{x}. \tag{10}$$

This crucial formula was first used for computation by G. Dziuk [35] with piecewise linear finite elements. In the context of geometric evolution of Sect. 2.1 we advance in time from t_n to t_{n+1} via a semi-implicit Euler method $\mathbf{x}_{n+1} = \mathbf{x}_n + \tau_n \mathbf{v}_{n+1}$, which keeps the geometry explicit,

$$\int_{\gamma_n} \mathbf{H}_{n+1} \cdot \boldsymbol{\Psi} - \tau_n \int_{\gamma_n} \nabla_{\gamma_n} \mathbf{v}_{n+1} : \nabla_{\gamma_n} \boldsymbol{\Psi} = \int_{\gamma_n} \nabla_{\gamma_n} \mathbf{x}_n : \nabla_{\gamma_n} \boldsymbol{\Psi}. \tag{11}$$

This equation for $\mathbf{H}_{n+1}$ is coupled with the equation for velocity $\mathbf{v}_{n+1}$, which comes from the gradient flows (5) or (9), or the Navier-Stokes equations (6)–(7). Getting separate equations for $\mathbf{H}_{n+1}$ and $\mathbf{v}_{n+1}$ is effectively an operator splitting technique, introduced by G. Dziuk [36], which has been used in a number of papers; see e.g. [2–4, 6–9, 15, 16, 33, 37, 38, 50].

The *mean curvature flow* of a surface γ is governed by $V = -H$, with V being the scalar normal velocity of γ. On the basis of (10), this geometric PDE can be reformulated as a heat equation for the position $\mathbf{x}$ on γ, following a seminal idea of G. Dziuk [36]:

$$\partial_t \mathbf{x} = V\boldsymbol{\nu} = -\mathbf{H} = \Delta_\gamma \mathbf{x}.$$

This allows for a simple and efficient finite element discretization [36]. The analysis of the resulting FEM is still open, except for the case of graphs [24–26].

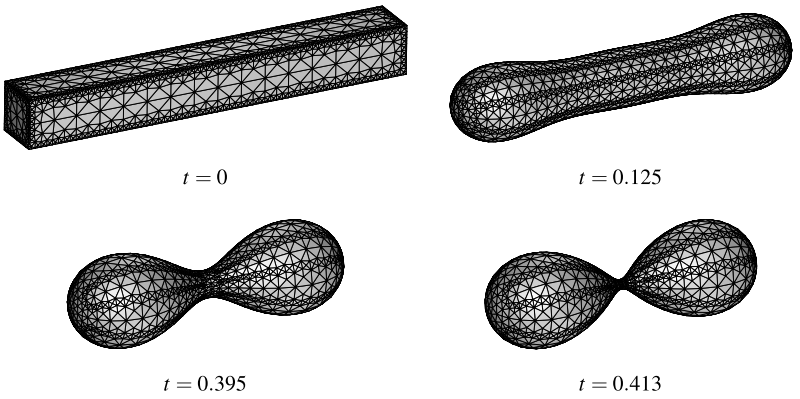


Fig. 5 Pinch-of in finite time. Evolution by surface diffusion of an $8 \times 1 \times 1$ prism at various time instants leading to a dumbbell and cusp formation

Expression (10) is also a crucial building block in the approach of E. Bänsch to Navier-Stokes equations with *free capillary surfaces* [2]. On the free surface γ, the Cauchy stress tensor Σ satisfies the Young-Laplace equation

$$\boldsymbol{\nu}\Sigma = \mathbf{H},$$

which allows for the following simple and elegant weak formulation of the boundary term

$$\int_\gamma \boldsymbol{\nu}\Sigma\mathbf{w}^T = \int_\gamma \mathbf{H}\mathbf{w}^T = -\int_\gamma \Delta_\gamma \mathbf{x}\mathbf{w}^T = \int_\gamma \nabla_\gamma \mathbf{x} : \nabla_\gamma \mathbf{w}, \tag{12}$$

where $\mathbf{w}$ is a suitable test function. This leads again to a simple and efficient FEM [2].

2.3 Surface Diffusion and Epitaxial Films

Surface diffusion is a 4th order geometric driven motion of a surface with normal velocity proportional to the surface Laplacian of mean curvature:

$$V = \Delta_\gamma H. \tag{13}$$

This PDE corresponds to the H^{-1} gradient flow of the area functional $J(\gamma) = \int_\gamma 1$, and has been studied by J. Cahn and J. Taylor [20] among others. E. Bänsch, P. Morin, and R.H. Nochetto proposed a parametric FEM upon combining (11) and (12) [5]. Other related schemes have been developed by J. Barrett, H. Garcke and R. Nürnberg [6, 7]. The analysis of this problem is still open, except for the graph case [4, 26].

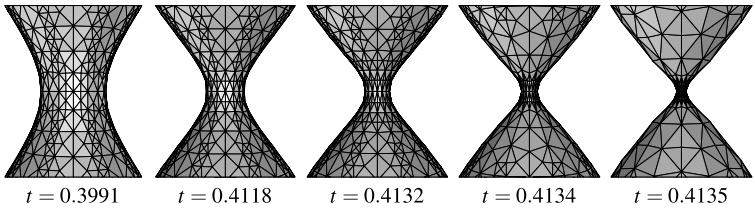


Fig. 6 Detailed view of the pinch-off produced by surface diffusion of the $8 \times 1 \times 1$ prism. Adaptivity becomes essential when approaching the pinch-off configuration

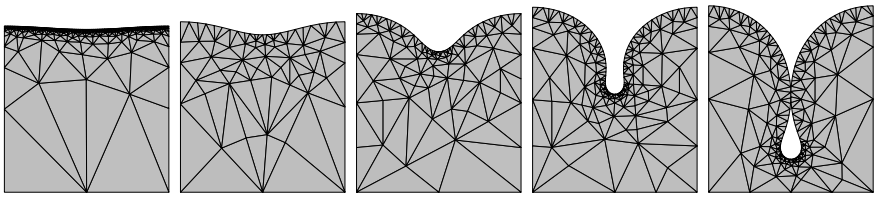


Fig. 7 Domain dynamics governed by coupling surface diffusion with the Laplace operator in the bulk. This leads to a mushroom-like free surface that gives rise to an inclusion in finite time

Surface diffusion may lead to singularity formation in finite time, depending on the initial configuration. This is depicted in Figs. 5–6 which display the evolution of an initial $8 \times 1 \times 1$ prism [5]. This simulation shows that adaptivity is essential to approximate singular situations produced by the flow.

Modeling the deformation of the free surface γ of stressed epitaxial films leads to a variant of (13). The stress accounts for the misfit between the crystalline structure of the substrate and epitaxial film, and causes a plastic deformation of γ. This morphological instability of the free surface may eventually lead to crack formation and fracture, an issue of paramount importance in Materials Science. The dynamics of γ is governed by

$$V = \Delta_\gamma (H + \varepsilon),$$

ε being the elastic energy density of the bulk enclosed by γ (see [4, 5, 33] and the references therein). Applications to material science are given in [3, 8].

Consider now a simplified situation where elasticity is replaced by the Laplace operator in the bulk Ω enclosed between the free surface γ and the substrate Σ (see Fig. 7). We let $\varepsilon = |\nabla u|^2$, where u solves the problem

$$-\Delta u = 0 \quad \text{in } \Omega, \qquad \partial_\nu u = 0 \quad \text{on } \gamma,$$

and $u = x$ on the bottom Σ and lateral boundary. This yields interesting configurations including mushroom-like formations, thereby leading to defects in materials such as inclusions [33].

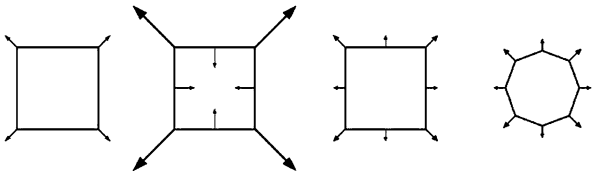


Fig. 8 Refinement procedures on a uniform partition of the unit circle using piecewise linear curves. The *arrows* on the piecewise linear curve represent the approximation of the curvature $\mathbf{H}$, all scaled down by the same multiplicative factor 0.3. We depict the starting approximation of the curvature (*first*), that after one global bisection of the surface approximation using the naive approach (*second*), and that with the GCAP method (*third*). In contrast with them GCAP algorithm, the standard algorithm does not preserve the accuracy of the geometric approximations. The last picture (*fourth*) depicts the new approximation of curvature over the surface parametrized by the vector $\mathbf{X}_*$ obtained in step (iii) of the GCAP algorithm

2.4 Geometrically Consistent Accuracy Preserving Algorithm

The chief geometric identity (10) turns out to play an important role when performing mesh modifications (refinement/coarsening/smoothing) on manifolds with incomplete information on their geometry, yet preserving position and curvature accuracy. This is typically the case when the surface γ is unknown as in the examples provided in Sects. 2.1 and 2.3: γ is known only through its approximation Γ and the approximation of its vector curvature $\mathbf{H}$, still labeled $\mathbf{H}$.

The naive approach when performing mesh modification consists of (i) apply the mesh modification to Γ; (ii) compute the corresponding curvature $\mathbf{H}$ according to a discrete version of (10) ensuring geometric consistency (here $\mathbf{X}$ is the identity on Γ). It turns out that the last step yields loss of accuracy on the approximation of the curvature regardless of polynomial degree, which is inherent to computing two derivatives numerically—an unstable process.

To circumvent this issue, A. Bonito, R.H. Nochetto, and S.M. Pauletti [14] propose a Geometrically Consistent Accuracy Preserving Algorithm (GCAP) which reverses the above process: $\mathbf{X}$ is dissociated from Γ itself in that it is no longer the identity on Γ. In essence, the GCAP algorithm proceeds as follows: (i) the mesh modifications are performed on Γ to give the new surface Γ_*; (ii) the new approximation $\mathbf{H}_*$ of vector curvature is obtained *projecting* the existing one $\mathbf{H}$ on Γ_*; (iii) the approximation $\mathbf{X}_*$ of the identity vector on Γ_* is obtained by *solving* the Laplace-Beltrami equation (10) discretely with the curvature $\mathbf{H}_*$ given in (ii). We stress that the concatenation of projection and inversion of (10) is numerically stable.

To compare the naive and GCAP algorithms, Fig. 8 depicts the effect of a global refinement on a square approximation Γ of a circle γ; here $\Gamma_* = \Gamma$. We refer to [14] for similar results for two dimensional surfaces, higher polynomial approximations, and coarsening as well as mesh smoothing.

3 Parametric Surfaces

In this section we discuss both how to represent and interpolate a parametric surface. This is instrumental for the design, analysis, and implementation of AFEM on parametric surfaces.

3.1 Representation of Parametric Surfaces

We assume that the surface γ is described as the deformation of a d dimensional polyhedral surface Γ_0 by a globally Lipschitz *homeomorphism* $P_0 : \Gamma_0 \to \gamma \subset \mathbb{R}^{d+1}$. If $\Gamma_0 = \bigcup_{i=1}^{I} \Gamma_0^i$ is made up of I (closed) faces $\Gamma_0^i, i = 1, \ldots, I$, we denote by $P_0^i : \Gamma_0^i \to \mathbb{R}^{d+1}$ the restriction of P_0 to Γ_0^i. We refer to Γ_0^i as a *macro-element* which induces the partition $\{\gamma^i\}_{i=1}^{I}$ of γ upon setting

$$\gamma^i := P_0^i\big(\Gamma_0^i\big).$$

In order to avoid technicalities, we assume that all the macro-elements are simplices, i.e. there is a (closed) reference simplex $\Omega \subset \mathbb{R}^d$, from now on called the *parametric domain*, and an affine map $\mathcal{F}_0^i : \mathbb{R}^d \to \mathbb{R}^{d+1}$ such that $\Gamma_0^i = \mathcal{F}_0^i(\Omega)$; Fig. 9 sketches the situation when $d = 2$. We thus let $\mathcal{X}^i := P_0^i \circ \mathcal{F}_0^i : \Omega \to \gamma^i$ be a local parametrization of γ which is globally bi-Lipschitz, namely there exists a universal constant $L \geq 1$ such that for all $1 \leq i \leq I$

$$L^{-1}|\hat{x} - \hat{y}| \leq \big|\mathcal{X}^i(\hat{x}) - \mathcal{X}^i(\hat{y})\big| \leq L|\hat{x} - \hat{y}|, \quad \forall \hat{x}, \hat{y} \in \Omega. \tag{14}$$

This *minimal regularity* of γ, to be soon strengthened out locally in each macro-element, implies the more familiar condition, valid for a.e. $\hat{x} \in \Omega$,

$$L^{-1}|w| \leq \big|\widehat{\nabla}\mathcal{X}^i(\hat{x})w\big| \leq L|w| \quad \forall w \in \mathbb{R}^d; \tag{15}$$

hence $L \geq 1$ is the Lipschitz constant of $\mathcal{X}^i$ and so of γ^i. We further assume that $P_0(\mathbf{v}) = \mathbf{v}$ for all vertices $\mathbf{v}$ of Γ_0, so that $\mathcal{F}_0^i$ is the nodal interpolant of $\mathcal{X}^i$ into linear polynomials.

The structure of the map P_0 depends on the application. For instance, if γ^i is described on Γ_0^i via the *distance function* $\operatorname{dist}(x)$ to γ, then

$$\gamma^i \ni \tilde{x} = x - \operatorname{dist}(x)\nabla \operatorname{dist}(x) = P_0(x) \quad \forall x \in \Gamma_0^i,$$

provided $\operatorname{dist}(x)$ is sufficiently small so that the distance is uniquely defined. If, instead, γ^i is the *zero level set* $\phi(x) = 0$ of a function ϕ, then

$$\Gamma_0^i \ni x = \tilde{x} + \frac{\nabla\phi(\tilde{x})}{|\nabla\phi(\tilde{x})|}|x - \tilde{x}| = P_0^{-1}(\tilde{x}), \quad \forall \tilde{x} \in \gamma^i,$$

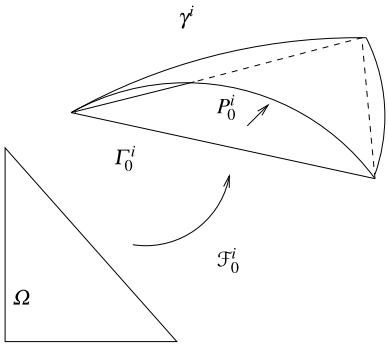


Fig. 9 Representation of each component γ^i when $d = 2$ as a parametrization from a flat triangle $\Gamma_0^i \subset \mathbb{R}^3$ as well as from the master triangle $\Omega \subset \mathbb{R}^2$. The map $\mathcal{F}_0^i : \Omega \to \Gamma_0^i$ is affine

is the inverse map of P_0. In both cases, dist and ϕ must be C^2 for P_0 to be $C^1(\Gamma_0^i)$. Yet another option is to view γ^i as a graph on Γ_0^i, in which case P_0^i is a lift in the normal direction to Γ_0^i and P_0 is $C^1(\Gamma_0^i)$ if and only if γ^i is; we refer to [46]. Notice that the inverse mapping theorem implies $(P_0^i)^{-1} \in C^1(\gamma^i)$.

The *regularity* of γ is expressed in terms of the regularity of the maps $\mathcal{X}^i$. If $s \geq 0$, $2 \leq p \leq \infty$, we say that γ is piecewise W_p^s, and write $\gamma \in W_p^s(\Gamma_0)$, whenever $\mathcal{X}^i \in [W_p^s(\Omega)]^{d+1}$, $i = 1, \ldots, I$. We denote the associated semi-norm by

$$|\gamma|_{W_p^s(\Gamma_0)} := \left(\sum_{i=1}^{I} |\mathcal{X}^i|^p_{W_p^s(\Omega)} \right)^{1/p}.$$

Note that this *non-overlapping* parametrization allows for piecewise smooth surfaces γ with possible kinks matched by the decomposition $\{\gamma^i\}_{i=1}^I$. Similarly, we say that $\gamma \in C^{1,\alpha}(\Gamma_0)$, $0 \leq \alpha \leq 1$, whenever $\mathcal{X}^i \in [C^{1,\alpha}(\Omega)]^{d+1}$, $i = 1, \ldots, I$ and define

$$|\gamma|_{C^{1,\alpha}(\Gamma_0)} := \max_{i=1,\ldots,I} |\mathcal{X}^i|_{C^{1,\alpha}(\Omega)}.$$

Finally, we note that a function $v : \gamma^i \to \mathbb{R}$ defines uniquely two functions $\hat{v} : \Omega \to \mathbb{R}$ and $\bar{v} : \Gamma_0^i \to \mathbb{R}$ via the maps $\mathcal{X}^i$ and P_0, namely

$$\hat{v}(\hat{x}) := v\big(\mathcal{X}^i(\hat{x})\big) \quad \forall \hat{x} \in \Omega \quad \text{and} \quad \bar{v}(\bar{x}) := v\big(P_0(\bar{x})\big) \quad \forall \bar{x} \in \Gamma_0^i; \tag{16}$$

we set $\tilde{x} = \mathcal{X}^i(\hat{x})$ for all $\hat{x} \in \Omega$. Conversely, a function $\hat{v} : \Omega \to \mathbb{R}$ (respectively, $\bar{v} : \Gamma_0^i \to \mathbb{R}$) defines uniquely the two functions $v : \gamma^i \to \mathbb{R}$ and $\bar{v} : \Gamma_0^i \to \mathbb{R}$ (respectively, $v : \gamma^i \to \mathbb{R}$ and $\hat{v} : \Omega \to \mathbb{R}$). We will always denote by v the two lifts $\tilde{v}$ or $\hat{v}$ of $v : \gamma^i \to \mathbb{R}$.

3.2 Interpolation of Parametric Surfaces

The initial partition of Γ_0 in macro-elements (or faces) induces a conforming triangulation $\mathcal{T}_0$ of Γ_0. We only discuss the class of conforming meshes $\mathbb{T}(\mathcal{T}_0)$ created by successive bisections of this initial mesh $\mathcal{T}_0$. However, our results remain valid for any refinement strategy satisfying Conditions 3, 4 and 6 in [12]. In particular, successive bisections, quad-refinement and red-refinement all with hanging nodes are admissible refinement strategies. For more details, we refer to [12, Sect. 6].

Given $\mathcal{T}_0$, we define a shape regular forest $\mathbb{T}(\mathcal{T}_0)$, and for each $\mathcal{T} \in \mathbb{T}(\mathcal{T}_0)$, a *piecewise affine* approximation $\Gamma = \Gamma(\mathcal{T})$ of γ, and a finite element space $\mathbb{V}(\mathcal{T})$ on Γ as follows. Note first that conforming graded bisections of each macro-element Γ_0^i induce a family of shape regular partitions $\mathcal{T}^i(\Omega)$ of the parametric domain $\Omega \subset \mathbb{R}^d$. Let $\mathbb{V}(\mathcal{T}^i(\Omega))$ be the finite element space of C^0 piecewise linear polynomials on $\mathcal{T}^i(\Omega)$, and let $\mathcal{I}_{\mathcal{T}^i} : C^0(\Omega) \to \mathbb{V}(\mathcal{T}^i(\Omega))$ be the corresponding Lagrange interpolation operator. Let $\mathcal{F}_{\mathcal{T}^i} = \mathcal{I}_{\mathcal{T}^i} \mathcal{X}^i$ be the interpolant of $\mathcal{X}^i$ in $\mathbb{V}(\mathcal{T}^i(\Omega))$, $\Gamma^i := \mathcal{F}_{\mathcal{T}^i}(\Omega)$ and

$$\mathcal{T}^i := \big\{ T = \mathcal{F}_{\mathcal{T}^i}(\widehat{T}) \;\big|\; \widehat{T} \in \mathcal{T}^i(\Omega) \big\};$$

the set Γ^i is a piecewise affine interpolation of γ^i. The global mesh $\mathcal{T}$, piecewise affine surface Γ, and parametrization $\mathcal{F}_{\mathcal{T}}$ of Γ are given by

$$\mathcal{T} := \bigcup_{i=1}^{I} \mathcal{T}^i, \qquad \Gamma := \bigcup_{i=1}^{I} \Gamma^i, \qquad \mathcal{F}_{\mathcal{T}} := \{\mathcal{F}_{\mathcal{T}^i}\}_{i=1}^{I}.$$

We need a few properties before discussing shape regularity of $\mathbb{T}(\mathcal{T}_0) = \{\mathcal{T}\}$. We define

$$\mathbb{V}(\mathcal{T}) := \Big\{ V \in C^0(\Gamma) \;\Big|\; V|_{\Gamma^i} \text{ is the lift of some } V \in \mathbb{V}\big(\mathcal{T}^i(\Omega)\big) \text{ via } \mathcal{F}_{\mathcal{T}^i}, \\ V = 0 \text{ on } \partial\Gamma, \text{ or } \int_\Gamma V = 0 \text{ if } \partial\Gamma = \emptyset \Big\},$$

and note that $\mathbb{V}(\mathcal{T})$ is not a subspace of $\mathbb{V}(\mathcal{T}_0)$, which is a lack of consistency we must account for. Since most properties discussed below are valid independently of the superscript i, we omit it from now on. Figure 10 depicts one bisection refinement for $d = 2$.

If $\widehat{T} \in \mathcal{T}(\Omega)$ and $T = \mathcal{F}_{\mathcal{T}}(\widehat{T}) \in \mathcal{T}$, we define the *geometric element indicator*

$$\lambda_\Gamma(T) := \big\| \widehat{\nabla}(\mathcal{X} - \mathcal{F}_{\mathcal{T}}) \big\|_{L^\infty(\widehat{T})}, \tag{17}$$

and the corresponding *geometric estimator*

$$\lambda_\Gamma := \max_{T \in \mathcal{T}} \lambda_\Gamma(T). \tag{18}$$

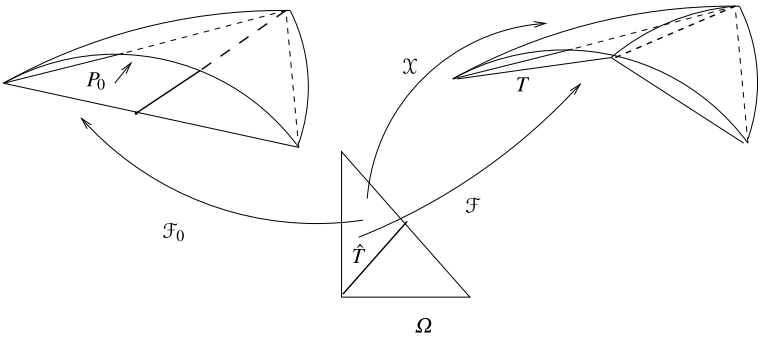


Fig. 10 Effect of one bisection of the macro-element $\mathcal{T}_0(\Omega)$ when $d = 2$ (*left*). The parametric domain Ω is split into two triangles in $\mathbb{R}^2$ via the affine map $\mathcal{F}_0^{-1}$ (*bottom*), whereas γ is interpolated by a new piecewise linear surface $\Gamma = \mathcal{F}(\Omega)$ (*right*), with $\mathcal{F} = \mathcal{I}_{\mathcal{T}}\mathcal{X}$ the piecewise linear interpolant of the parametrization $\mathcal{X}$ defined in Ω. The superscript i is omitted for simplicity

Note that two different meshes giving rise to the same surface Γ yield the same λ_Γ, which is thus of pure geometric nature; this explains the subscript Γ. Moreover, $\lambda_\Gamma(T)$ is evaluated in $\widehat{T}$, which belongs to the parametric domain Ω instead of the polyhedral surface Γ. The geometric estimator may not decrease upon each refinement, especially in the pre-asymptotic regime, but the following *quasi-monotonicity* property is valid instead: there exists a constant $\Lambda_0 \geq 1$, depending on $\mathcal{T}_0$, and dimension d, such that

$$\lambda_{\Gamma_*} \leq \Lambda_0 \lambda_\Gamma \tag{19}$$

for all conforming refinements $\mathcal{T}_*$ of $\mathcal{T}$ [17, Lemma 3.1]. This result is also valid elementwise.

We recall that $\mathbb{T}(\mathcal{T}_0)$ is the forest of all conforming refinements $\mathcal{T}$ of $\mathcal{T}_0$, denoted $\mathcal{T} \geq \mathcal{T}_0$, obtained by the aforementioned bisection procedure. We say that $\mathbb{T}(\mathcal{T}_0)$ is *shape regular* if there is a constant C_0 only depending on $\mathcal{T}_0$, such that for all $\widehat{T} \in \mathcal{T}(\Omega)$

$$C_0^{-1}|\hat{x} - \hat{y}| \leq \left|\mathcal{F}_{\mathcal{T}}(\hat{x}) - \mathcal{F}_{\mathcal{T}}(\hat{y})\right| \leq C_0|\hat{x} - \hat{y}| \quad \forall \hat{x}, \hat{y} \in \widehat{T}. \tag{20}$$

Since the forest induced by bisection on the flat parametric domain Ω is shape regular [11, 49, 53], we observe that (20) states that the deformation of $\widehat{T} \in \mathcal{T}(\Omega)$ leading to $T \in \mathcal{T}$ does not degenerate. We also point out that (20) implies the usual condition on the Jacobian $\widehat{\nabla}\mathcal{F}_{\mathcal{T}}$, valid for a.e. $\hat{x} \in \Omega$

$$C_0^{-1}|w| \leq \left|\widehat{\nabla}\mathcal{F}_{\mathcal{T}}(\hat{x})w\right| \leq C_0|w| \quad \forall w \in \mathbb{R}^d, \tag{21}$$

and that $\widehat{\nabla}\mathcal{F}_{\mathcal{T}}$ happens to be constant on $\widehat{T}$ for an affine map $\mathcal{F}_{\mathcal{T}}$ [22].

We stress that a bi-Lipschitz parametrization satisfying (14) does *not* guarantee that $\mathbb{T}(\mathcal{T}_0)$ is shape regular. This pathological situation is depicted in Fig. 11. This issue has been tackled by A. Bonito and J. Pasciak [13] assuming that the surface γ

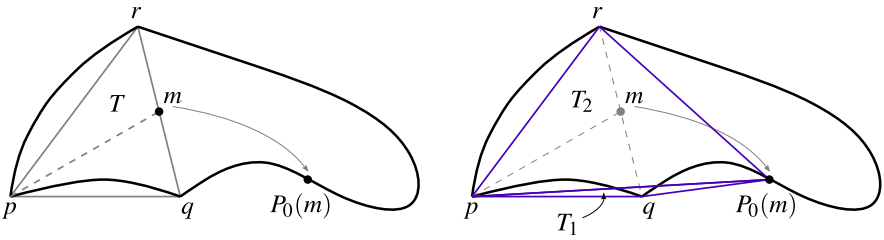


Fig. 11 Smooth surface leading to a degenerate triangle. The point $P_0(m)$ is (almost) aligned with p and q. When the triangle $T = pqr \in \mathcal{T}$ is split into pqm and rpm, the new elements of $\mathcal{T}_*$ are $T_1 = pqP_0(m)$, $T_2 = rpP_0(m)$. The triangle T_1 is degenerate and $\lambda_{\Gamma_*}(\widehat{T}_1) > (2\Lambda_0 L)^{-1}$, thus violating (22). This forces ADAPT_SURFACE to refine further, which in turn opens up T_1 leading eventually to nondegenerate descendents of T_1

is W^2_∞ and $\mathcal{T}_0$ is sufficiently fine. We now present a similar result, invoking piecewise C^1-regularity of γ, which hinges on (19): the forest $\mathbb{T}(\mathcal{T}_0)$ is shape-regular with $C_0 = 2L$ provided

$$\lambda_{\Gamma_0} \le \frac{1}{2\Lambda_0 L}, \tag{22}$$

where $L > 1$ is the constant in (14) [17, Lemma 3.2]. Figure 11 illustrates an intermediate degenerate situation in which a triangle $T \in \mathcal{T}$ is split into two triangles $T_1, T_2 \in \mathcal{T}_*$ with $\lambda_{\Gamma_*}(T_1) > (2\Lambda_0 L)^{-1}$ and (22) being violated. This thereby leads to refinement of T_1, which opens up and gives rise to nondegenerate descendents eventually satisfying (22).

4 The Laplace-Beltrami Operator

4.1 Basic Differential Geometry

In this subsection we give a matrix formulation of some basic differential geometry facts. We assume γ to be piecewise C^1, namely $\gamma^i \in C^1(\Gamma_0^i)$ for all $1 \le i \le I$, and Γ to be piecewise affine.

Our first task is to relate the gradient $\widehat{\nabla}$ in the parametric domain Ω with the tangential gradient ∇_γ on γ. To this end, let $\mathbf{T} \in \mathbb{R}^{(d+1)\times d}$ be the matrix

$$\mathbf{T} := \mathbf{T}_\gamma := [\widehat{\partial}_1 \mathcal{X}, \ldots, \widehat{\partial}_d \mathcal{X}],$$

whose ith column $\widehat{\partial}_i \mathcal{X} \in \mathbb{R}^{d+1}$ is the vector of partial derivatives of $\mathcal{X}$ with respect to the ith coordinate of Ω. Since $\mathcal{X}$ is a diffeomorphism, the set $\{\widehat{\partial}_i \mathcal{X}\}_{i=1}^d$ of tangent vectors to γ is well defined, linearly independent, and expands the tangent hyperplane to each γ^j at interior points for all $1 \le j \le I$. The *first fundamental form* of γ is the symmetric and positive definite matrix $\mathbf{g} \in \mathbb{R}^{d\times d}$ defined by

$$\mathbf{g} = (g_{\gamma,ij})_{1\le i,j\le d} := \left(\widehat{\partial}_i \mathcal{X}^T \widehat{\partial}_j \mathcal{X}\right)_{1\le i,j\le d} = \mathbf{T}^T \mathbf{T}. \tag{23}$$

Given $\hat{v}(\hat{x}) = v(\tilde{x})$, the tangent gradient $\nabla_\gamma v(\tilde{x}) = \sum_{i=1}^d \alpha_i(\hat{x}) \widehat{\partial}_i \mathcal{X}(\hat{x})$ satisfies the relation

$$\widehat{\partial}_i \hat{v}(\hat{x}) = \nabla_\gamma v(\tilde{x})\, \widehat{\partial}_i \mathcal{X}(\hat{x}) \quad \text{for } 1 \le i \le d,$$

whence

$$\widehat{\nabla} \hat{v} = \nabla_\gamma v \mathbf{T} \tag{24}$$

and $(\alpha_i)_{i=1}^d = \mathbf{g}^{-1} (\widehat{\partial}_i \hat{v})_{i=1}^d$. To get the reverse relation, we augment $\mathbf{T}$ to the matrix $\widetilde{\mathbf{T}} \in \mathbb{R}^{(d+1)\times(d+1)}$ by adding the (outer) unit normal $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{d+1}) \in \mathbb{R}^{(d+1)}$ to the tangent hyperplane $\operatorname{span}\{\widehat{\partial} \mathcal{X}_i\}_{i=1}^d$ to γ as the last column, namely

$$\widetilde{\mathbf{T}} := \left[\mathbf{T}, \boldsymbol{\nu}^T\right] = \left[\widehat{\partial}_1 \mathcal{X}, \ldots, \widehat{\partial}_d \mathcal{X}, \boldsymbol{\nu}^T\right].$$

Since $\widetilde{\mathbf{T}}$ is invertible, we let $\widetilde{\mathbf{D}} = \widetilde{\mathbf{T}}^{-1}$. We thus realize that

$$\nabla_\gamma v = \nabla_\gamma v \widetilde{\mathbf{T}} \widetilde{\mathbf{D}} = [\widehat{\nabla} \hat{v}, 0] \widetilde{\mathbf{D}} = \widehat{\nabla} \hat{v} \mathbf{D}, \tag{25}$$

where $\mathbf{D} \in \mathbb{R}^{d\times(d+1)}$ results from $\widetilde{\mathbf{D}}$ by cutting off its last row. Moreover, writing

$$\mathbf{I}_{(d+1)\times(d+1)} = \widetilde{\mathbf{T}}^{-1} \widetilde{\mathbf{T}} = \begin{bmatrix} \mathbf{D} \\ \mathbf{v} \end{bmatrix} \begin{bmatrix} \mathbf{T} & \boldsymbol{\nu}^T \end{bmatrix} = \begin{bmatrix} \mathbf{DT} & \mathbf{D}\boldsymbol{\nu}^T \\ \mathbf{vT} & \mathbf{v}\boldsymbol{\nu}^T \end{bmatrix}$$

with $\mathbf{v} \in \mathbb{R}^{d+1}$, we deduce $\mathbf{DT} = \mathbf{I}_{d\times d}$ and $\mathbf{vT} = 0$ whence $\mathbf{v}$ is parallel to $\boldsymbol{\nu}$ and $\mathbf{v} = \boldsymbol{\nu}$ because $\mathbf{v}\boldsymbol{\nu}^T = 1$. Reverting the order of multiplication, we also infer that

$$\mathbf{I}_{(d+1)\times(d+1)} = \widetilde{\mathbf{T}} \widetilde{\mathbf{T}}^{-1} = \begin{bmatrix} \mathbf{T} & \boldsymbol{\nu}^T \end{bmatrix} \begin{bmatrix} \mathbf{D} \\ \boldsymbol{\nu} \end{bmatrix} = \mathbf{TD} + \boldsymbol{\nu}^T \boldsymbol{\nu},$$

and $\mathbf{TD} = \mathbf{I}_{(d+1)\times(d+1)} - \boldsymbol{\nu}^T \boldsymbol{\nu}$. This shows that $\mathbf{TD}$ is symmetric and

$$\mathbf{TDD}^T \mathbf{T}^T = \mathbf{TDTD} = \mathbf{TD} = \mathbf{I}_{(d+1)\times(d+1)} - \boldsymbol{\nu}^T \boldsymbol{\nu}, \tag{26}$$

as well as

$$\mathbf{DD}^T \mathbf{T}^T \mathbf{T} = \mathbf{DTDT} = \mathbf{I}_{d\times d}.$$

Therefore, the first fundamental form $\mathbf{g}$ has inverse $\mathbf{g}^{-1} = \mathbf{DD}^T$. We let

$$q := \sqrt{\det \mathbf{g}} \tag{27}$$

be the elementary area of γ and point out the change of variables formula for $\omega \subset \Omega$

$$\int_\omega \hat{v} q = \int_{\mathcal{X}(\omega)} v. \tag{28}$$

The discussion above applies as well to the piecewise affine surface Γ. We denote the corresponding matrices $\mathbf{T}_\Gamma = \widehat{\nabla} \mathcal{F}_\mathcal{T}$ and $\mathbf{D}_\Gamma$ associated with $\mathcal{F}_\mathcal{T} : \Omega \to \Gamma$, and get

$$\nabla_\Gamma v = \widehat{\nabla} \hat{v} \mathbf{D}_\Gamma. \tag{29}$$

The first fundamental form $\mathbf{G}_\Gamma$ of Γ and its elementary area Q_Γ are defined by

$$\mathbf{G}_\Gamma := \mathbf{T}_\Gamma^T \mathbf{T}_\Gamma, \qquad Q_\Gamma := \sqrt{\det \mathbf{G}_\Gamma}. \tag{30}$$

It is worth noticing that, since $\mathcal{F}_\mathcal{T}$ is affine, $\mathbf{G}_\Gamma$ and Q_Γ are constant on each $\widehat{T} \in \mathcal{T}(\Omega)$ ($T \in \mathcal{T}$).

4.2 Variational Formulation and Galerkin Method

We now introduce basic Lebesgue and Sobolev spaces on the surface γ. Let

$$L^2_\#(\gamma) := \left\{ v \in L^2(\gamma) \;\middle|\; \int_\gamma v = 0 \quad \text{if } \partial\gamma = \emptyset \right\}$$

be the space of L^2 functions, with vanishing meanvalue whenever the surface γ is closed, and

$$H^1_\#(\gamma) := \big\{ v \in L^2_\#(\gamma) \;\big|\; \nabla_\gamma v|_{\gamma^i} \in \big[L^2(\gamma^i)\big]^{d+1}, \\ v|_{\gamma^i} = v|_{\gamma^j} \text{ on } \gamma^i \cap \gamma^j \; 1 \le i, j \le I, \; v = 0 \text{ on } \partial\gamma \big\},$$

where ∇_γ and traces are well defined in each component γ^i due to (25). We define the weak form of the *Laplace-Beltrami operator* $\Delta_\gamma v$ for any function $v \in H^1_\#(\gamma)$ to be

$$\langle -\Delta_\gamma v, \varphi \rangle := \sum_{i=1}^{I} \int_{\gamma^i} \nabla_\gamma v \nabla_\gamma^T \varphi \quad \forall \varphi \in H^1_\#(\gamma), \tag{31}$$

where $\langle \cdot, \cdot \rangle$ denotes the $(H^1_\#(\gamma))^* - H^1_\#(\gamma)$ duality product. In order to derive a strong form of Δ_γ, we now assume that $\mathcal{X}^i$ is C^2 and $v \in H^2(\gamma^i)$ for each $1 \le i \le d$. In view of (25), integrating by parts in Ω we obtain

$$\int_{\gamma^i} \nabla_\gamma v \nabla_\gamma^T \varphi = \int_\Omega \widehat{\nabla} \hat{v} \mathbf{D}\mathbf{D}^T \widehat{\nabla} \hat{\varphi}^T q = \int_\Omega -\frac{1}{q} \widehat{\operatorname{div}}\big(q \widehat{\nabla} \hat{v} \mathbf{g}^{-1}\big) \hat{\varphi} q + \int_{\partial\Omega} q \widehat{\nabla} \hat{v} \mathbf{g}^{-1} \widehat{\mathbf{n}}^T \hat{\varphi},$$

where $\widehat{\mathbf{n}}$ is the unit outer normal to Ω. We thus discover that inside γ^i the following expression for the Laplace-Beltrami operator holds

$$\Delta_\gamma v = \frac{1}{q} \widehat{\operatorname{div}}\big(q \widehat{\nabla} \hat{v} \mathbf{g}^{-1}\big). \tag{32}$$

The boundary term instead leads to jumps across the boundary $\partial\gamma^i$ with other pieces γ^j of γ and can be equivalently written as

$$\int_{\partial\Omega} q \widehat{\nabla} \hat{v} \mathbf{g}^{-1} \widehat{\mathbf{n}}^T \hat{\varphi} = \int_{\partial\gamma^i} \nabla_\gamma v \mathbf{n}^T \varphi, \tag{33}$$

where $\mathbf{n}$ in the unit outer normal to γ^i in the tangent plane to γ^i. Combining (32) with (33) yields

$$\int_{\gamma^i} \nabla_\gamma v \nabla_\gamma^T \varphi = \int_{\gamma^i} -\Delta_\gamma v \varphi + \int_{\partial\gamma^i} \nabla_\gamma v \mathbf{n}^T \varphi, \tag{34}$$

which is the Gauss-Green formula for C^2 surfaces.

Expression (33) is not obvious and, since it is quite important for the subsequent discussion, we prove it now. Recall that $\Omega \subset \mathbb{R}^d$ is the canonical unit simplex and notice that a change of variables in Ω dictated by a rotation leaves the left-hand side of (33) unchanged. We exploit this property to assume, for convenience, that an arbitrary $\hat{x} \in \partial\Omega$ belongs to the $(d-1)$-subsimplex $\widehat{S}$ with outer normal given by $\widehat{\mathbf{n}} = [-1, 0, \ldots, 0]$. We observe that the affine function $\widehat{\phi}(\hat{x}) = \hat{x}\widehat{\mathbf{n}}^T$ vanishes on $\widehat{S}$ and $\widehat{\nabla}\widehat{\phi} = \widehat{\mathbf{n}} = \nabla_\gamma \phi \mathbf{T}$, according to (24), whence

$$\nabla_\gamma \phi \widehat{\partial}_1 \mathcal{X} = -1, \qquad \nabla_\gamma \phi \widehat{\partial}_i \mathcal{X} = 0 \quad 2 \le i \le d;$$

moreover, $\nabla_\gamma \phi = |\nabla_\gamma \phi| \mathbf{n}$. We now introduce the matrix $\mathbf{S} \in \mathbb{R}^{(d+1)\times(d-1)}$

$$\mathbf{S} = [\widehat{\partial}_2 \mathcal{X}, \ldots, \widehat{\partial}_d \mathcal{X}], \qquad r = \sqrt{\det(\mathbf{S}^T \mathbf{S})},$$

and point out that the quantity r is the elementary area associated with the subsimplex $\widehat{S}$ at $\hat{x}$. Since the $(d-1)$-dimensional space $\operatorname{span}\{\widehat{\partial}_i \mathcal{X}\}_{i=2}^d$ is tangent to the curvilinear simplex $\widetilde{S} = \mathcal{X}(\widehat{S})$, we can decompose $\widehat{\partial}_1 \mathcal{X}$ orthogonally as follows

$$\widehat{\partial}_1 \mathcal{X} = \alpha \mathbf{n} + \mathbf{S}\mathbf{b}, \quad \alpha \in \mathbb{R}, \ \mathbf{b} \in \mathbb{R}^{d-2},$$

where $\mathbf{b}$ is the least squares solution $\mathbf{b} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T \widehat{\partial}_1 \mathcal{X}^T$ and $|\widehat{\partial}_1 \mathcal{X}|^2 = \alpha^2 + |\mathbf{S}\mathbf{b}|^2$. Hence,

$$-1 = \nabla_\gamma \phi \widehat{\partial}_1 \mathcal{X} = \alpha |\nabla_\gamma \phi| \quad \Rightarrow \quad \alpha = -|\nabla_\gamma \phi|^{-1}.$$

We compute $q^2 = \det \mathbf{g}$ using the expression for block matrices

$$\mathbf{g} = \begin{bmatrix} |\widehat{\partial}_1 \mathcal{X}|^2 & \widehat{\partial}_1 \mathcal{X}\mathbf{S} \\ \mathbf{S}^T \widehat{\partial}_1 \mathcal{X}^T & \mathbf{S}^T\mathbf{S} \end{bmatrix},$$

namely

$$\det \mathbf{g} = \det(\mathbf{S}^T\mathbf{S})\big(|\widehat{\partial}_1 \mathcal{X}|^2 - \widehat{\partial}_1 \mathcal{X}\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T \widehat{\partial}_1 \mathcal{X}^T\big),$$

to infer that

$$q^2 = r^2\alpha^2 \quad \Longrightarrow \quad |\nabla_\gamma \phi| = \frac{r}{q} \quad \Longrightarrow \quad \widehat{\mathbf{n}} = \mathbf{n}\mathbf{T}\frac{r}{q}.$$

To finally derive (33), we recall that $\mathbf{T}\mathbf{D} = \mathbf{I}_{(d+1)\times(d+1)} - \boldsymbol{\nu}^T \boldsymbol{\nu}$ and observe that

$$\int_{\widehat{S}} q \widehat{\nabla} \hat{v} \mathbf{g}^{-1} \widehat{\mathbf{n}}^T \hat{\varphi} = \int_{\widehat{S}} q \widehat{\nabla} \hat{v} \mathbf{D}\mathbf{D}^T \widehat{\mathbf{n}}^T \hat{\varphi} = \int_{\widehat{S}} r \nabla_\gamma v \mathbf{D}^T \mathbf{T}^T \mathbf{n}^T \varphi = \int_{\widetilde{S}} \nabla_\gamma v \mathbf{n}^T \varphi.$$

We now build on (31) and write the weak formulation of $-\Delta_\gamma u = f$ as follows: given $f \in L^2_\#(\gamma)$, we seek $u \in H^1_\#(\gamma)$ satisfying

$$\sum_{i=1}^{I} \int_{\gamma^i} \nabla_\gamma u \nabla_\gamma^T \varphi = \int_\gamma f\varphi, \quad \forall \varphi \in H^1_\#(\gamma). \tag{35}$$

Existence and uniqueness of a solution $u \in H^1_\#(\gamma)$ is a consequence of the Lax-Milgram theorem provided γ is Lipschitz. Combining (35) with (34) and (33) yields for each component γ^i,

$$-\Delta_{\gamma^i} u = f \quad 1 \le i \le I, \tag{36}$$

together with vanishing jump conditions at the interfaces $\gamma^i \cap \gamma^j$

$$\jmath(u)|_{\gamma^i \cap \gamma^j} = \nabla_{\gamma^i} u \mathbf{n}^i + \nabla_{\gamma^j} u \mathbf{n}^j = 0 \quad \forall 1 \le i, j \le I, \tag{37}$$

because $f \in L^2_\#(\gamma)$ cannot balance this singular term otherwise.

We next formulate an approximation to the Laplace-Beltrami operator on a piecewise affine approximation Γ of γ supported by a mesh $\mathcal{T} \in \mathbb{T}(\mathcal{T}_0)$. If $F_\Gamma \in L^2_\#(\Gamma)$ is a suitable approximation of f, then the finite element solution $U : \Gamma \to \mathbb{R}$ solves

$$U \in \mathbb{V}(\mathcal{T}): \quad \int_\Gamma \nabla_\Gamma U \nabla_\Gamma^T V = \int_\Gamma F_\Gamma V \quad \forall V \in \mathbb{V}(\mathcal{T}). \tag{38}$$

To this end we choose F_Γ to be

$$F_\Gamma := f \frac{q}{Q_\Gamma}, \tag{39}$$

because this specific choice of F_Γ satisfies the compatibility property

$$\int_\Gamma F_\Gamma = \int_\gamma f = 0, \tag{40}$$

whenever γ is closed, and allows us to handle separately the approximation of surface γ and forcing f. In particular, (38) admits a unique solution U as a consequence of the Lax-Milgram theorem. Since Γ is piecewise affine, the quantities $\widehat{\nabla} \widehat{U}, \mathbf{G}_\Gamma, Q_\Gamma$ are piecewise constant, whence

$$\Delta_\Gamma U|_T = 0 \quad \forall T \in \mathcal{T}. \tag{41}$$

We refer to [17] where we account for piecewise polynomial Γ and the fact that $\Delta_\Gamma U|_T \neq 0$. The formula (34) extends to every element $T \in \mathcal{T}$:

$$\int_T \nabla_\Gamma U \nabla_\Gamma^T V = \int_T -\Delta_\Gamma U V + \int_{\partial T} \nabla_\Gamma U \mathbf{n}_T^T V \quad \forall V \in \mathbb{V}(\mathcal{T}). \tag{42}$$

5 A Posteriori Error Analysis

In order to study the discrepancy between u and U we need to agree on comparing them in a common domain, say γ. Our goal is thus to obtain a posteriori error estimates for the energy error $\|\nabla_\gamma(u-U)\|_{L^2(\gamma)}$. This requires developing an a priori error analysis for the interpolation error committed in replacing γ by Γ in (38), which is a sort of consistency error, and its impact on the PDE error. We are concerned with these issues in this section and refer to [28, 29, 46].

5.1 Geometric Error and Estimator

We now quantify the error arising from approximating γ, the so-called *geometric error*. To this end we resort to the matrix formulation of Sect. 4.1 to relate the geometric error with the geometric estimator λ_Γ of (17).

Given $T \in \mathcal{T}$, we will deal with the regions $\widehat{T} \in \mathcal{T}(\Omega)$ and $\widetilde{T} \in \mathcal{T}(\gamma)$ given by

$$\widehat{T} := \left\{\mathcal{F}_{\mathcal{T}}^{-1}(x) | x \in T\right\} \quad \text{and} \quad \widetilde{T} := \left\{\mathcal{X}(\hat{x}) | \hat{x} \in \widehat{T}\right\}. \tag{43}$$

On mapping back and forth to $\widehat{T}$, and using (28), we easily see that

$$\int_T v = \int_{\widetilde{T}} v \frac{Q_\Gamma}{q}. \tag{44}$$

We are now able to quantify the consistency error alluded to at the beginning of this section.

Lemma 5.1 (Consistency error) *For all $v, w \in H^1(\gamma)$ there holds*

$$\int_\Gamma \nabla_\Gamma v \nabla_\Gamma^T w - \int_\gamma \nabla_\gamma v \nabla_\gamma^T w = \int_\gamma \nabla_\gamma v \mathbf{E}_\Gamma \nabla_\gamma^T w,$$

where $\mathbf{E}_\Gamma \in \mathbb{R}^{(d+1)\times(d+1)}$ stands for the following error matrix

$$\mathbf{E}_\Gamma := \frac{1}{q}\mathbf{T}\left(Q_\Gamma \mathbf{G}_\Gamma^{-1} - q\mathbf{g}^{-1}\right)\mathbf{T}^T. \tag{45}$$

Proof We first note that combining (24) with (25), we get

$$\nabla_\gamma v = \nabla_\Gamma v \mathbf{T}_\Gamma \mathbf{D} \quad \text{and} \quad \nabla_\Gamma v = \nabla_\gamma v \mathbf{T}\mathbf{D}_\Gamma, \tag{46}$$

which together with (44) gives

$$\int_\Gamma \nabla_\Gamma v \nabla_\Gamma^T w = \int_\gamma \nabla_\gamma v \mathbf{T}\mathbf{D}_\Gamma \mathbf{D}_\Gamma^T \mathbf{T}^T \nabla_\gamma^T w \frac{Q_\Gamma}{q} \quad \forall v, w \in H^1(\gamma). \tag{47}$$

Since (26) allows us to write

$$\int_\gamma \nabla_\gamma v \nabla_\gamma^T w = \int_\gamma \nabla_\gamma v \mathbf{T}\mathbf{D}\mathbf{D}^T \mathbf{T}^T \nabla_\gamma^T w \quad \forall v, w \in H^1(\gamma), \tag{48}$$

which is a counterpart of (47), the assertion follows immediately from (23) and (30). □

Our next task is to estimate $\mathbf{E}_\Gamma$ in (45), which entails dealing with $\mathbf{g}, \mathbf{G}_\Gamma$ and q, Q_Γ.

Lemma 5.2 (Properties of $\mathbf{G}_\Gamma$ and Q_Γ) *The matrices* $\mathbf{g}$ *and* $\mathbf{G}_\Gamma$ *have eigenvalues in the interval* $[L^{-2}, L^2]$ *and* $[\frac{1}{2}L^{-2}, \frac{3}{2}L^2]$, *respectively, provided the initial mesh* $\mathcal{T}_0$ *satisfies*

$$\lambda_{\Gamma_0} \leq \frac{1}{6\Lambda_0 L^3}. \tag{49}$$

Moreover, the forest $\mathbb{T}(\mathcal{T}_0)$ *is shape regular,* $L^{-d} \lesssim q, Q_\Gamma \lesssim L^d$, *and for all* $\mathcal{T} \in \mathbb{T}(\mathcal{T}_0)$

$$\|q - Q_\Gamma\|_{L^\infty(\gamma)} + \|\mathbf{g} - \mathbf{G}_\Gamma\|_{L^\infty(\gamma)} \lesssim \lambda_\Gamma. \tag{50}$$

Proof Since $L \geq 1$, (49) yields (22), which in turn gives shape regularity of the forest $\mathbb{T}(\mathcal{T}_0)$ and (21) with constant $C_0 = 2L$. Hence, using the definitions of $\mathbf{g}$ and $\mathbf{G}_\Gamma$, we deduce $\|\mathbf{g} - \mathbf{G}_\Gamma\|_{L^\infty(\gamma)} \leq 3L\lambda_\Gamma$. On the other hand, invoking (15) we see that $\boldsymbol{\xi}^T \mathbf{g} \boldsymbol{\xi} = |\nabla \mathcal{X} \boldsymbol{\xi}|^2$ for all $\boldsymbol{\xi} \in \mathbb{R}^d$, whence

$$L^{-2}|\boldsymbol{\xi}|^2 \leq \boldsymbol{\xi}^T \mathbf{g} \boldsymbol{\xi} \leq L^2 |\boldsymbol{\xi}|^2.$$

Since $\lambda_\Gamma \leq \frac{1}{6L^3}$, due to (49) and (19), then the previous estimates readily imply

$$\frac{1}{2}L^{-2}|\boldsymbol{\xi}|^2 \leq \big(L^{-2} - 3L\lambda_\Gamma\big)|\boldsymbol{\xi}|^2 \leq \boldsymbol{\xi}^T \mathbf{G}_\Gamma \boldsymbol{\xi} \leq \big(L^2 + 3L\lambda_\Gamma\big)|\boldsymbol{\xi}|^2 \leq \frac{3}{2}L^2|\boldsymbol{\xi}|^2,$$

as well as $L^{-d} \lesssim q, Q_\Gamma \lesssim L^d$ because $q^2 = \det \mathbf{g}$, $Q_\Gamma^2 = \det \mathbf{G}_\Gamma$ are products of the d eigenvalues of $\mathbf{g}, \mathbf{G}_\Gamma$. Moreover, since

$$q - Q_\Gamma = \frac{\det \mathbf{g} - \det \mathbf{G}_\Gamma}{q + Q_\Gamma},$$

it only remains to obtain an estimate for the numerator. The definition of determinant readily yields $|\det \mathbf{g} - \det \mathbf{G}| \lesssim L^{2d-1}\lambda_\Gamma$, and completes the proof. □

We stress that if $\mathcal{T}_0$ does not satisfy (49) but $\varepsilon_0 \leq (6\Lambda_0 L^3 \omega)^{-1}$, then the algorithm AFEM of Sect. 1 will first refine $\mathcal{T}_0$ to make it comply with (49) without ever solving the PDE. In this sense, (49) is not a serious restriction for AFEM, although necessary for the subsequent theory.

Corollary 5.1 (Estimate of $\mathbf{E}_\Gamma$) *If λ_{Γ_0} satisfies (49), then we have for all $\mathcal{T} \in \mathbb{T}(\mathcal{T}_0)$ and corresponding Γ*

$$\|\mathbf{E}_\Gamma\|_{L^\infty(\widehat{T})} \lesssim \lambda_\Gamma(T) \quad \forall T \in \mathcal{T},$$

where the hidden constant depends on $\mathcal{T}_0$ and the Lipschitz constant L of γ.

Proof According to (45), and $\|\mathbf{T}\|_{L^\infty(\gamma)} = \|\mathbf{T}^T\|_{L^\infty(\gamma)} \le L$, we infer that

$$\|\mathbf{E}_\Gamma\|_{L^\infty(\widehat{T})} \lesssim \big\| Q_\Gamma \mathbf{G}_\Gamma^{-1} - q\mathbf{g}^{-1} \big\|_{L^\infty(\widehat{T})}.$$

The lower bounds on the eigenvalues of $\mathbf{g}$ and $\mathbf{G}_\Gamma$ imply $\|\mathbf{g}^{-1}\|_{L^\infty(\widehat{T})}, \|\mathbf{G}_\Gamma^{-1}\|_{L^\infty(\widehat{T})} \lesssim L^2$, which together with the expression

$$Q_\Gamma \mathbf{G}_\Gamma^{-1} - q\mathbf{g}^{-1} = (Q_\Gamma - q)\mathbf{G}_\Gamma^{-1} + q\mathbf{G}_\Gamma^{-1}(\mathbf{g} - \mathbf{G}_\Gamma)\mathbf{g}^{-1}$$

and (50) gives the asserted estimate. □

We now give a constructive expression for unit normals in $\mathbb{R}^{d+1}$, thereby generalizing the usual vector product in $\mathbb{R}^3$, and next use it to derive an error estimate for $\mathbf{D}_\Gamma$.

Lemma 5.3 (Unit normal) *Let $\{\mathbf{e}_j\}_{j=1}^{d+1}$ be the canonical unit vectors of $\mathbb{R}^{d+1}$. For each $\widehat{x} \in \Omega$, and $x = \mathcal{X}(\widehat{x}) \in \gamma$, let $\mathbf{N}(\widehat{x}) = \sum_{j=1}^{d+1} A_j(\widehat{x})\mathbf{e}_j$, where A_j stands for the determinant*

$$A_j(\widehat{x}) := \det\big(\mathbf{e}_j, \widehat{\partial}_1\mathcal{X}(\widehat{x}), \dots, \widehat{\partial}_d\mathcal{X}(\widehat{x})\big).$$

We then have $|\mathbf{N}(\widehat{x})| = q(\widehat{x})$ and the unit normal vector $\boldsymbol{\nu}(x)$ to γ at x is given by $\boldsymbol{\nu}(x) = \mathbf{N}(\widehat{x})/|\mathbf{N}(\widehat{x})|$. Moreover, a similar result holds true also for Γ, upon replacing $\mathcal{X}$ by $\mathcal{F}_\mathcal{T}$, provided λ_{Γ_0} satisfies (49), i.e., $|\mathbf{N}_\Gamma(\widehat{x})| = Q_\Gamma(\widehat{x})$ and $\boldsymbol{\nu}_\Gamma(x) = \mathbf{N}_\Gamma(\widehat{x})/|\mathbf{N}_\Gamma(\widehat{x})|$.

Proof We fix $\widehat{x} \in \Omega$ and drop it from the notation. Since $\mathbf{T}$ is full rank, some A_j must be non-zero whence $\mathbf{N} \neq 0$. Moreover, the vector $\mathbf{N}$ is orthogonal to the tangent hyperplane to γ at x because

$$\mathbf{N} \cdot \widehat{\partial}_i\mathcal{X} = \sum_{j=1}^{d+1} A_j\mathbf{e}_j \cdot \widehat{\partial}_i\mathcal{X} = \det(\widehat{\partial}_i\mathcal{X}, \widehat{\partial}_1\mathcal{X}, \dots, \widehat{\partial}_i\mathcal{X}, \dots, \widehat{\partial}_d\mathcal{X}) = 0.$$

Hence, $\boldsymbol{\nu} = \mathbf{N}/|\mathbf{N}|$ is well defined. To prove that $|\mathbf{N}| = q$ recall that $\mathbf{T} = [\widehat{\partial}_1\mathcal{X}, \dots, \widehat{\partial}_d\mathcal{X}]$ to write

$$|\mathbf{N}|^2 = \sum_{j=1}^{d+1} A_j^2 = \sum_{j=1}^{d+1} A_j \det(\mathbf{e}_j, \widehat{\partial}_1\mathcal{X}, \dots, \widehat{\partial}_d\mathcal{X}) = \det(\mathbf{N}, \widehat{\partial}_1\mathcal{X}, \dots, \widehat{\partial}_d\mathcal{X})$$

$$= \left\{\det\big([\mathbf{N},\mathbf{T}]^T[\mathbf{N},\mathbf{T}]\big)\right\}^{1/2} = \left\{\det\begin{bmatrix}\mathbf{N}^T\mathbf{N} & 0\\ 0 & \mathbf{T}^T\mathbf{T}\end{bmatrix}\right\}^{1/2} = |\mathbf{N}|q.$$

This implies $|\mathbf{N}| = q$ because $|\mathbf{N}| \neq 0$. The same argument applies to Γ. □

Lemma 5.4 (Error of $\boldsymbol{\nu}$ and $\mathbf{D}$) *If* (49) *holds for the initial mesh* $\mathcal{T}_0$, *then for all* $\mathcal{T} \in \mathbb{T}(\mathcal{T}_0)$

$$\|\boldsymbol{\nu} - \boldsymbol{\nu}_\Gamma\|_{L^\infty(\gamma)} + \|\mathbf{D} - \mathbf{D}_\Gamma\|_{L^\infty(\gamma)} \lesssim \lambda_\Gamma. \tag{51}$$

Proof Lemmas 5.2 and 5.3 imply $L^{-d} \lesssim |\mathbf{N}(\widehat{x})|, |\mathbf{N}_\Gamma(\widehat{x})| \lesssim L^d$ for all $\widehat{x} \in \Omega$, whence

$$\boldsymbol{\nu} - \boldsymbol{\nu}_\Gamma = \frac{\mathbf{N}}{|\mathbf{N}|} - \frac{\mathbf{N}_\Gamma}{|\mathbf{N}_\Gamma|} = \frac{1}{|\mathbf{N}|}(\mathbf{N} - \mathbf{N}_\Gamma) + \left(\frac{1}{|\mathbf{N}|} - \frac{1}{|\mathbf{N}_\Gamma|}\right)\mathbf{N}_\Gamma$$
$$\Rightarrow \quad |\boldsymbol{\nu} - \boldsymbol{\nu}_\Gamma| \lesssim L^d|\mathbf{N} - \mathbf{N}_\Gamma|.$$

To estimate $\mathbf{N} - \mathbf{N}_\Gamma = \sum_{j=1}^{d+1}(A_j - A_{\Gamma,j})\mathbf{e}_j$, we observe that each A_j (resp. $A_{\Gamma,j}$) is a sum of factors of the form $\widehat{\partial}_i\mathcal{X}\cdot\mathbf{e}_m$ (resp. $\widehat{\partial}_i\mathcal{F}_\mathcal{T}\cdot\mathbf{e}_m$), whence

$$|A_j - A_{\Gamma,j}| \lesssim L^{d-1}\lambda_\Gamma \quad \Rightarrow \quad |\boldsymbol{\nu} - \boldsymbol{\nu}_\Gamma| \lesssim L^{2d-1}\lambda_\Gamma.$$

For the remaining estimate for $\mathbf{D} - \mathbf{D}_\Gamma$ we recall the definition $\widetilde{\mathbf{T}} = [\mathbf{T}, \boldsymbol{\nu}^T]$ to infer that

$$\|\widetilde{\mathbf{T}} - \widetilde{\mathbf{T}}_\Gamma\|_{L^\infty(\gamma)} \le \|\mathbf{T} - \mathbf{T}_\Gamma\|_{L^\infty(\gamma)} + \|\boldsymbol{\nu} - \boldsymbol{\nu}_\Gamma\|_{L^\infty(\gamma)} \lesssim \lambda_\Gamma.$$

We now show that $\widetilde{\mathbf{D}} = \widetilde{\mathbf{T}}^{-1}$ is uniformly bounded. To see this, we write $\widetilde{\mathbf{T}}\tilde{\mathbf{w}} = \mathbf{T}\mathbf{w} + w_{d+1}\boldsymbol{\nu}^T$ for $\tilde{\mathbf{w}} = (\mathbf{w}, w_{d+1}) \in \mathbb{R}^{d+1}$ and recall (15) to get

$$L^{-2}|\tilde{\mathbf{w}}|^2 \le L^{-2}|\mathbf{w}|^2 + |w_{d+1}|^2 \le |\widetilde{\mathbf{T}}\tilde{\mathbf{w}}|^2 \le L^2|\mathbf{w}|^2 + |w_{d+1}|^2 \le L^2|\tilde{\mathbf{w}}|^2,$$

as well as $\|\widetilde{\mathbf{T}}^{-1}\|_{L^\infty(\Omega)}, \|\widetilde{\mathbf{T}}_\Gamma^{-1}\|_{L^\infty(\Omega)} \lesssim L^2$. Since $\widetilde{\mathbf{D}} - \widetilde{\mathbf{D}}_\Gamma = \widetilde{\mathbf{T}}^{-1}(\widetilde{\mathbf{T}}_\Gamma - \widetilde{\mathbf{T}})\widetilde{\mathbf{T}}_\Gamma^{-1}$, the desired estimate follows immediately from the previous one for $\widetilde{\mathbf{T}} - \widetilde{\mathbf{T}}_\Gamma$. □

We finally point out the equivalence of norms on γ and Γ provided (49) is valid [17, Lemma 5.6]

$$\|v\|_{L^2(\widetilde{T})} \approx \|v\|_{L^2(T)}, \qquad |v|_{L^2(\widetilde{T})} \approx |v|_{L^2(T)} \quad \forall T \in \mathcal{T}. \tag{52}$$

5.2 Upper and Lower Bounds for the Energy Error

We now derive an error representation formula leading to lower and upper bounds for the energy error. Given $\mathcal{T} \in \mathbb{T}(\mathcal{T}_0)$, we let the usual interior and jump residual

for $V \in \mathbb{V}(\mathcal{T})$ be

$$\mathcal{R}_T(V) := F_\Gamma|_T + \Delta_\Gamma V|_T = F_\Gamma|_T \quad \forall T \in \mathcal{T},$$
$$\mathcal{J}_S(V) := \nabla_\Gamma V^+|_S \cdot \mathbf{n}_S^+ + \nabla_\Gamma V^-|_S \cdot \mathbf{n}_S^- \quad \forall S \in \mathcal{S},$$

where n_S^+ and n_S^- are outward unit normals to S with respect to T^+ and T^-, on the supporting planes containing T^+ and T^- respectively; T^+ and T^- are elements in $\mathcal{T}$ that share the side $S \in \mathcal{S}$ where $\mathcal{S}$ denotes the set of interior faces of $T \in \mathcal{T}$. We stress that, in contrast to flat domains, $\mathbf{n}_S^+ \neq \mathbf{n}_S^-$ because the vector may have different supporting hyperplanes. Similarly, $\nabla_\Gamma V^+|_S = \widehat{\nabla} V^+ \mathbf{D}_\Gamma|_{\widehat{S}}$ and $\nabla_\Gamma V^-|_S = \widehat{\nabla} V^- \mathbf{D}_\Gamma|_{\widehat{S}}$ are tangential gradients of V on T^+ and T^- restricted to S, respectively. Note that, according to (36),

$$\Delta_\Gamma V|_T = Q_\Gamma^{-1} \widehat{\operatorname{div}}\big(Q_\Gamma \widehat{\nabla} \widehat{V} \mathbf{G}_\Gamma^{-1}\big)|_{\widehat{T}} = 0 \quad \forall T \in \mathcal{T},$$

provided V and Γ are piecewise linear. We refer to [17, 29] for the case $\Delta_\Gamma V|_T \neq 0$.

Subtracting the weak formulations (35) and (38), and employing (34) to integrate by parts elementwise, we obtain for all $v \in H^1(\gamma)$:

$$\int_\gamma \nabla_\gamma (u - U) \cdot \nabla_\gamma v = I_1 + I_2 + I_3, \tag{53}$$

with

$$I_1 := \sum_{T \in \mathcal{T}} \int_T F_\Gamma (v - V) - \sum_{S \in \mathcal{S}} \int_S \mathcal{J}_S(U)(v - V),$$
$$I_2 := \int_\Gamma \nabla_\Gamma U \cdot \nabla_\Gamma v - \int_\gamma \nabla_\gamma U \cdot \nabla_\gamma v = \int_\gamma \nabla_\gamma U \mathbf{E}_\Gamma \nabla_\gamma^T v,$$
$$I_3 := \int_\gamma f v - \int_\Gamma F_\Gamma v.$$

The choice $F_\Gamma = \frac{q}{Q_\Gamma} f$ of (39) implies $I_3 = 0$ so that only I_1 and I_2 need to be estimated. Observe that I_1 is the usual residual term, whereas I_2 is the geometry consistency term studied in Sect. 5.1 which accounts for the discrepancy between γ and Γ.

We focus now on I_1. The PDE error indicator is defined as follows for any $V \in \mathbb{V}(\mathcal{T})$

$$\eta_{\mathcal{T}}(V, T)^2 := h_T^2 \|F_\Gamma\|_{L^2(T)}^2 + \frac{1}{2} \sum_{S \subset \partial T} h_T \big\|\mathcal{J}_S(V)\big\|_{L^2(S)}^2 \quad \forall T \in \mathcal{T},$$

where $h_T := |T_0|^{\frac{1}{d}}$ and T_0 is the preimage of T in the initial triangulation $\mathcal{T}_0$, i.e. $T_0 = \mathcal{F}_0 \circ \mathcal{F}_{\mathcal{T}}^{-1}(T)$. This definition of h_T guarantees the strict reduction property

$$h_{T'} \leq 2^{-b/d} h_T \tag{54}$$

for all T' obtained from T after b bisections. We also introduce the *data oscillation*

$$\operatorname{osc}_{\mathcal{T}}(f,T) := h_T \|F_\Gamma - \overline{F}_\Gamma\|_{L^2(T)} \quad \forall T \in \mathcal{T}, \tag{55}$$

where $\overline{F}_\Gamma$ stands for the meanvalue of F_Γ on $T \in \mathcal{T}$. Finally, for any subset $\tau \subset \mathcal{T}$ we set

$$\eta_{\mathcal{T}}(V,\tau)^2 := \sum_{T\in\tau} \eta_{\mathcal{T}}(V,T)^2, \quad \text{and} \quad \operatorname{osc}_{\mathcal{T}}(f,\tau)^2 := \sum_{T\in\tau} \operatorname{osc}_{\mathcal{T}}(f,T)^2,$$

and simply write $\eta_{\mathcal{T}}(V)$ and $\operatorname{osc}_{\mathcal{T}}(f)$ whenever $\tau = \mathcal{T}$.

Standard arguments [1, 55] to derive upper and lower bounds for the energy error on flat domains can be extended to this case; see [17, 29, 46]. We thus sketch the proof.

Lemma 5.5 (A posteriori upper and lower bounds) *Assume that λ_{Γ_0} satisfies (49). Let $u \in H^1(\gamma)$ be the solution of (35), $(\Gamma,\mathcal{T})$ an approximating surface-mesh pair, and $U \in \mathbb{V}(\mathcal{T})$ be the Galerkin solution of (38). Then there exist constants C_1, C_2 and Λ_1 depending only on $\mathcal{T}_0$, the Lipschitz constant of γ, and $\|f\|_{L^2(\gamma)}$, such that*

$$\big\|\nabla_\gamma(u-U)\big\|^2_{L^2(\gamma)} \le C_1\eta_{\mathcal{T}}(U)^2 + \Lambda_1\lambda_\Gamma^2, \tag{56}$$

$$C_2\eta_{\mathcal{T}}(U)^2 \le \big\|\nabla_\gamma(u-U)\big\|^2_{L^2(\gamma)} + \operatorname{osc}_{\mathcal{T}}(f)^2 + \Lambda_1\lambda_\Gamma^2. \tag{57}$$

Proof Our departing point is (53) with $v \in H^1_\#(\gamma)$ arbitrary and $V \in \mathbb{V}(\mathcal{T})$ its Scott-Zhang interpolant, built over the parametric domain Ω [19]. Using interpolation estimates and (52) yields

$$|I_1| \lesssim \eta_{\mathcal{T}}(U)\|\nabla_\gamma v\|_{L^2(\gamma)}.$$

Since $\|\nabla_\Gamma U\|_{L^2(\gamma)} \lesssim \|f\|_{L^2(\gamma)}$, invoking Corollary 5.1 gives

$$|I_2| \lesssim \lambda_\Gamma\|\nabla_\gamma v\|_{L^2(\gamma)}.$$

Since $I_3 = 0$ we obtain the upper bound (56). To prove (57) we resort to a local argument due to R. Verfürth [55]. Let $T \in \mathcal{T}$ and b_T be corresponding cubic bubble. If $v = \overline{F}_\Gamma b_T \in H^1_0(T)$, then

$$\|\nabla_\gamma v\|_{L^2(T)} \lesssim h_T^{-1}\|\overline{F}_\Gamma\|_{L^2(T)}.$$

Therefore, inserting v into (53) and taking $V = 0$ leads to

$$\|\overline{F}_\Gamma\|^2_{L^2(T)} \lesssim \int_T \overline{F}_\Gamma v \lesssim h_T^{-1}\big(\big\|\nabla_\gamma(u-U)\big\|_{L^2(T)} + \lambda_\Gamma(T)\big)\|\overline{F}_\Gamma\|_{L^2(T)}.$$

This combined with the triangle inequality gives part of (57). It remains to deal with the jump, for which we select an arbitrary side $S \in \mathcal{S}$ with adjacent elements $T^\pm$. Let

b_S be a piecewise quadratic bubble with value 1 at the midpoint of S and 0 at any other quadratic node. Let $v = \mathcal{J}_S(U) b_S \in H_0^1(\omega_S)$ where $\omega_S = T^+ \cup T^-$. Replacing v into (53) and taking $V = 0$ yields

$$\begin{aligned}\left\|\mathcal{J}_S(U)\right\|_{L^2(S)}^2 &\lesssim \int_S \mathcal{J}_S(U) v \\ &\leq \left(\left\|\nabla_\gamma (u-U)\right\|_{L^2(\omega_S)} + h_S \|F_\Gamma\|_{L^2(\omega_S)} + \lambda_\Gamma(\omega_S)\right) \|\nabla_\gamma v\|_{L^2(\omega_S)}.\end{aligned}$$

To conclude the proof we invoke the property $\|\nabla_\gamma v\|_{L^2(\omega_S)} \lesssim h_S^{-1/2} \|\mathcal{J}_S(U)\|_{L^2(S)}$ along with the previous estimate for $h_S \|F_\Gamma\|_{L^2(\omega_S)}$. □

To prove optimality of AFEM we need a localized upper bound for the distance between two discrete solutions. This bound measures $\|\nabla_\gamma (U_* - U)\|_{L^2(\gamma)}$ in terms of the PDE estimator restricted to the refined set and geometric estimator [17, Lemma 4.13].

Lemma 5.6 (Localized upper bound) *Assume that λ_{Γ_0} satisfies (49). For $(\mathcal{T}, \Gamma)$, $(\mathcal{T}_*, \Gamma_*)$ pairs of mesh-surface approximations with $\mathcal{T} \leq \mathcal{T}_*$, let $\mathcal{R} := \mathcal{R}_{\mathcal{T} \to \mathcal{T}^*} \subset \mathcal{T}$ be the set of elements refined in $\mathcal{T}$ to obtain $\mathcal{T}^*$. Let $U \in \mathbb{V}(\mathcal{T})$ and $U \in \mathbb{V}(\mathcal{T}_*)$ be the corresponding discrete solutions of (38) on Γ and Γ_*, respectively. Then the following localized upper bound is valid*

$$\left\|\nabla_\gamma (U_* - U)\right\|_{L^2(\gamma)}^2 \leq C_1 \eta_{\mathcal{T}}(U, \mathcal{R})^2 + \Lambda_1 \lambda_\Gamma(\mathcal{R})^2, \tag{58}$$

with constants C_1, Λ_1 as in Lemma 5.5.

Proof We start from the error representation formula (53) by replacing γ by Γ_* and taking as a test function $v = E_* := U - U_* \in H_\#^1(\gamma)$

$$\left\|\nabla_\gamma (U_* - U)\right\|_{L^2(\gamma)}^2 \simeq \int_{\Gamma_*} \nabla_{\Gamma_*} (U_* - U) \cdot \nabla_{\Gamma_*} E_* = I_1 + I_2 + I_3.$$

To estimate I_1, we proceed as in the flat case [21, 49, 52]. We first construct an approximation $V \in \mathbb{V}(\mathcal{T})$ of $E_* \in \mathbb{V}(\mathcal{T}_*)$. Let ω be the union of elements of $\mathcal{T}$ which are refined in $\mathcal{T}_*$, and denote by ω_i one of the connected components of its interior. Let $\mathcal{T}_i$ be the subset of $\mathcal{T}$ contained in ω_i and let $\mathbb{V}(\mathcal{T}_i)$ be the restriction of $\mathbb{V}(\mathcal{T})$ to ω_i. We now can construct the Scott-Zhang operator on the corresponding flat domains $\widehat{\omega}_i = \mathcal{F}_{\mathcal{T}}^{-1}(\omega)$ and then lift them to Γ via $\mathcal{F}_{\mathcal{T}}$. We denote these lifts by $P_i : H^1(\widehat{\omega}_i) \to \mathbb{V}(\mathcal{T}_i)$. Let $V \in \mathbb{V}(\mathcal{T})$ be the following approximation of the error $E_* \in \mathbb{V}(\mathcal{T}_*)$:

$$V := P_i E_* \quad \text{in } \omega_i, \qquad V := E_* \quad \text{elsewhere.}$$

By construction, V has conforming boundary values on $\partial\omega_i$, is continuous in Γ, i.e. $V \in \mathbb{V}(\mathcal{T})$ and is an H^1-stable approximation to E_*. Since $V = E_*$ in $\Gamma \backslash \omega$ we

obtain by standard argument

$$|I_1| \leq C_1 \eta_{\mathcal{T}}(U, \mathcal{R}) \|\nabla_\Gamma E_*\|_{L^2(\Gamma)}.$$

To estimate I_2, we first note that $I_2|_{\Gamma\setminus\omega} = 0$ because Γ and Γ_* coincide in the unrefined region $\Gamma\setminus\omega$. Adding and subtracting $\int_{\widetilde{\omega}} \nabla_\gamma U \nabla_\gamma E_*$, with $\widetilde{\omega} = \mathcal{X} \circ \mathcal{F}_{\mathcal{T}}^{-1}(\omega)$, we obtain

$$I_2 = \int_{\widetilde{\omega}} \nabla_\gamma U \mathbf{E}_\Gamma \nabla_\gamma^T E_* - \int_{\widetilde{\omega}} \nabla_\gamma U \mathbf{E}_{\Gamma_*} \nabla_\gamma^T E_*.$$

Combining Corollary 5.1 with (52) and (19), in its elementwise form, we obtain

$$|I_2| \lesssim \big(\lambda_\Gamma(\mathcal{R}) + \lambda_{\Gamma^*}(\mathcal{R})\big) \|\nabla_\Gamma E_*\|_{L^2(\gamma)} \lesssim \big(1 + \Lambda_0^2\big) \|f\|_{L^2(\gamma)} \lambda_\Gamma(\mathcal{R}).$$

We note that the choice (39) of discrete forcing terms F_{Γ_*} and F_Γ implies $I_3 = 0$. Finally, collecting the estimates above we conclude (58). □

5.3 Properties of the PDE Estimator and Data Oscillation

As indicated in (56)–(57), we have access to the energy error $\|\nabla_\gamma(u - U)\|_{L^2(\gamma)}$ only through the PDE estimator $\eta_{\mathcal{T}}(U)$, the geometric estimator λ_Γ, and data oscillation $\operatorname{osc}_{\mathcal{T}}(f)$. As is customary for flat domains, (55) guarantees that $\operatorname{osc}_{\mathcal{T}}(f)$ is dominated by $\eta_{\mathcal{T}}(U)$ locally:

$$\operatorname{osc}_{\mathcal{T}}(f, T) \leq \eta_{\mathcal{T}}(U, T) \quad \forall T \in \mathcal{T}. \tag{59}$$

The main novelty in (56)–(58) with respect to flat domains, which is also the chief challenge of the present analysis, is the presence of λ_Γ. In this respect, we show now the equivalence of $\eta_{\mathcal{T}}(U)$ and the *total error*

$$\mathcal{E}_{\mathcal{T}}(U, f) := \Big(\big\|\nabla_\gamma(u - U)\big\|_{L^2(\gamma)}^2 + \operatorname{osc}_{\mathcal{T}}(f)^2\Big)^{\frac{1}{2}} \tag{60}$$

provided λ_Γ is small relative to $\eta_{\mathcal{T}}(U)$. We refer to [21] for a similar result for flat domains.

Lemma 5.7 (Equivalence of estimator and total error) *Let C_1, C_2, Λ_1 be given in Lemma 5.5. If*

$$\lambda_\Gamma^2 \leq \frac{C_2}{2\Lambda_1} \eta_{\mathcal{T}}(U)^2, \tag{61}$$

then there exist explicit constants $C_3 \geq C_4 > 0$, depending on C_1, C_2, such that

$$C_4 \eta_{\mathcal{T}}(U) \leq \mathcal{E}_{\mathcal{T}}(U, f) \leq C_3 \eta_{\mathcal{T}}(U). \tag{62}$$

Proof Combining (56) with (61), we infer that

$$\left\|\nabla_\gamma(u-U)\right\|^2_{L^2(\gamma)} \leq \left(C_1+\frac{C_2}{2}\right)\eta_{\mathcal{T}}(U)^2. \tag{63}$$

This, together with (59), gives the upper bound in (62). We next resort to (57) and (61) to obtain

$$C_2\eta_{\mathcal{T}}(U)^2 \leq \left\|\nabla_\gamma(u-U)\right\|^2_{L^2(\gamma)} + \operatorname{osc}_{\mathcal{T}}(f)^2 + \frac{C_2}{2}\eta_{\mathcal{T}}(U)^2,$$

which implies the lower bound in (62) and concludes the proof. □

It turns out that the usual reduction property of $\eta_{\mathcal{T}}(U)$ [21, Corollary 3.4], which is instrumental to prove a contraction property of AFEM, is also polluted by the presence of λ_Γ as stated below. The following result is proved in [46, Lemma 4.2] for any polynomial degree.

Lemma 5.8 (Reduction of residual error estimator) *Let λ_{Γ_0} satisfy (49). Given a mesh-surface pair $(\mathcal{T},\Gamma)$, let $\mathcal{M}\subset\mathcal{T}$ be a subset of elements bisected at least $b\geq 1$ times in refining $\mathcal{T}$ to obtain $\mathcal{T}_*\geq\mathcal{T}$. If $\xi := 1-2^{-\frac{b}{d}}$, then there exist constants Λ_2 and Λ_3, solely depending on the shape regularity of $\mathcal{T}_0$, the Lipschitz constant L of γ, and $\|f\|_{L^2(\gamma)}$, such that for any $\delta>0$*

$$\begin{aligned}\eta_{\mathcal{T}_*}(U_*)^2 &\leq (1+\delta)\left(\eta_{\mathcal{T}}(U)^2-\xi\,\eta_{\mathcal{T}}(U,\mathcal{M})^2\right)\\ &\quad+\left(1+\delta^{-1}\right)\left(\Lambda_3\left\|\nabla_\gamma(U_*-U)\right\|^2_{L^2(\gamma)}+\Lambda_2\lambda_\Gamma^2\right).\end{aligned} \tag{64}$$

Proof Let $S\in\mathcal{S}_*$ be an interior side and $T^+,T^-\in\mathcal{T}^*$ be two elements sharing S. The component of $\nabla_{\Gamma_*}U_*$ tangential to S does not jump, because U_* is continuous across S, whence

$$\left|\jmath_S(U_*)\right| = \left|\nabla_{\Gamma_*}U_*^+ - \nabla_{\Gamma_*}U_*^-\right|,$$

where $U_*^\pm = U_*|_{T^\pm}$. Therefore

$$\begin{aligned}\left|\jmath_S(U_*)-\jmath_S(U)\right| &\leq \left|\nabla_{\Gamma_*}\left(U_*^+-U^+\right)\right| + \left|\nabla_{\Gamma_*}\left(U_*^--U^-\right)\right|\\ &\quad+\left|\nabla_{\Gamma_*}U^+-\nabla_\Gamma U^+\right|+\left|\nabla_{\Gamma_*}U^--\nabla_\Gamma U^-\right|.\end{aligned}$$

Employing an inverse estimate together with (52), the first two terms can be bounded as follows:

$$h_S\left\|\nabla_{\Gamma_*}\left(U_*^\pm-U^\pm\right)\right\|^2_{L^2(S)} \lesssim \left\|\nabla_{\Gamma_*}\left(U_*^\pm-U^\pm\right)\right\|^2_{L^2(T^\pm)} \approx \left\|\nabla_\gamma\left(U_*^\pm-U^\pm\right)\right\|^2_{L^2(\widetilde{T}^\pm)}.$$

For the next two terms we use (29), in conjunction with (51) and (19), to write

$$h_S\left\|\nabla_{\Gamma_*}U^\pm-\nabla_\Gamma U^\pm\right\|_{L^2(S)} \lesssim \left\|\widehat{\nabla}\widehat{U}^\pm(\mathbf{D}_{\Gamma_*}-\mathbf{D}_\Gamma)\right\|_{L^2(\widehat{T}^\pm)} \lesssim \lambda_\Gamma,$$

where the hidden constant depends on $\|f\|_{L^2(\gamma)}$.

We now turn our attention to the interior residual. Let $T_* \in \mathcal{T}_*$ and $T = \mathcal{F}_{\mathcal{T}} \circ \mathcal{F}_{\mathcal{T}_*}^{-1}(T_*)$, $\widehat{T} = \mathcal{F}_{\mathcal{T}_*}^{-1}(T_*)$ be the corresponding sets in Γ and Ω. Since $F_\Gamma = \frac{q}{Q_\Gamma} f$ we infer that

$$\left| \int_{T_*} |F_{\Gamma_*}|^2 - \int_T |F_\Gamma|^2 \right| = \int_{\widehat{T}} |qf|^2 \frac{|Q_\Gamma - Q_{\Gamma_*}|}{Q_\Gamma Q_{\Gamma_*}} \lesssim \lambda_\Gamma \|f\|^2_{L^2(\widetilde{T})},$$

because of (50) and the lower bounds for Q_Γ and Q_{Γ_*}, as well as (19).

Collecting the estimates above, we realize that we have derived the bound

$$\eta_{\mathcal{T}_*}(U_*)^2 \leq (1+\delta)\eta_{\mathcal{T}}(U, \mathcal{T}_*)^2 + \big(1+\delta^{-1}\big)\big(\Lambda_3 \big\|\nabla_\gamma (U_* - U)\big\|^2_{L^2(\gamma)} + \Lambda_2 \lambda_\Gamma^2\big).$$

It remains to deal with the set $\mathcal{M}$, namely to prove

$$\eta_{\mathcal{T}}(U, \mathcal{T}_*)^2 \leq \eta_{\mathcal{T}}(U)^2 - \xi \eta_{\mathcal{T}}(U, \mathcal{M})^2.$$

This is exactly the same argument as for flat domains because of the definition of meshsize h_T and (54) [21, Corollary 3.4]. This concludes the proof. □

Another difference with the theory of adaptivity for flat domains is the behavior of data oscillation under refinement. The usual situation is that $\operatorname{osc}_{\mathcal{T}}(f)$ does not increase upon refinement from $\mathcal{T}$ to $\mathcal{T}_*$ [47]. This is no longer true because $\operatorname{osc}_{\mathcal{T}}(f)$ and $\operatorname{osc}_{\mathcal{T}_*}(f)$ are defined on different domains Γ and Γ_*. Instead, we have the following substitute.

Lemma 5.9 (Quasi-monotonicity of data oscillation) *Let λ_{Γ_0} satisfy (49). Let $(\mathcal{T}, \Gamma)$, $(\mathcal{T}_*, \Gamma_*)$ be mesh-surface pairs with $\mathcal{T} \leq \mathcal{T}_*$ and discrete forcing functions defined according to (39). Then, there exists a constant $C_5 \geq 1$, depending only on $\mathcal{T}_0$ and the Lipschitz constant L of γ, such that*

$$\operatorname{osc}_{\mathcal{T}_*}(f) \leq C_5 \operatorname{osc}_{\mathcal{T}}(f). \tag{65}$$

Proof Let $T_* \in \mathcal{T}_*$ and so in Γ_*, and let $T = \mathcal{F}_\Gamma \circ \mathcal{F}_{\Gamma_*}^{-1}(T_*)$ be the corresponding set in Γ, but perhaps not in $\mathcal{T}$. Using (39) and the fact that Q_Γ is piecewise constant, we realize that

$$\begin{aligned}\int_{T_*} |F_{\Gamma_*} - \overline{F}_{\Gamma_*}|^2 &\leq \int_T \left| f\frac{q}{Q_{\Gamma_*}} - \overline{F}_\Gamma \frac{Q_\Gamma}{Q_{\Gamma_*}} \right|^2 \frac{Q_{\Gamma_*}}{Q_\Gamma} = \int_T |F_\Gamma - \overline{F}_\Gamma|^2 \frac{Q_\Gamma}{Q_{\Gamma_*}} \\ &\leq C_5^2 \int_T |F_\Gamma - \overline{F}_\Gamma|^2,\end{aligned}$$

where C_5^2 is the maximum of the ratios $Q_\Gamma / Q_{\Gamma_*}|_{T_*}$ for all $T_* \in \mathcal{T}_*$ and is bounded by L^{2d}. □

6 AFEM: Design and Properties

Since λ_Γ and $\eta_\mathcal{T}(U)$ account for quite different effects, the algorithm AFEM is designed to handle them separately via the modules ADAPT_SURFACE and ADAPT_PDE:

AFEM: Given Γ_0, $\mathcal{T}_0$, and parameters $\varepsilon_0 > 0$, $0 < \rho < 1$, and $\omega > 0$, set $k = 0$.

1. $[\mathcal{T}_k^+, \Gamma_k^+] = \mathsf{ADAPT\_SURFACE}(\mathcal{T}_k, \omega\varepsilon_k)$
2. $[\mathcal{T}_{k+1}, \Gamma_{k+1}] = \mathsf{ADAPT\_PDE}(\mathcal{T}_k^+, \varepsilon_k)$
3. $\varepsilon_{k+1} = \rho\varepsilon_k$; $k = k + 1$
4. Goto 1.

We notice the presence of the factor ω, which is employed to make the geometric error small relative to the current tolerance ε_k. This turns out to be essential for both contraction and optimality of AFEM, and is further discussed in Sects. 7–9.

6.1 Module *ADAPT_SURFACE*

Given a tolerance $\tau > 0$ and admissible subdivision $\mathcal{T}$, the call $[\mathcal{T}^+, \Gamma^+] = \mathsf{ADAPT\_SURFACE}(\mathcal{T}, \Gamma, \tau)$ improves the surface resolution until

$$\lambda_{\Gamma^+} \leq \tau \tag{66}$$

where λ_Γ is the *geometric estimator* introduced in (17). This module is based on a *greedy* algorithm

$[\mathcal{T}^+, \Gamma^+] = \mathsf{ADAPT\_SURFACE}(\mathcal{T}, \Gamma, \tau)$
  while $\mathcal{M} := \{T \in \mathcal{T} | \lambda_\mathcal{T}(T) > \tau\} \neq \emptyset$
    $\mathcal{T} := \mathsf{REFINE}(\mathcal{T}, \mathcal{M})$
    $\Gamma := \mathcal{F}_\mathcal{T}(\Omega)$
  end while
  return$(\mathcal{T},\Gamma)$

where $\mathsf{REFINE}(\mathcal{T}, \mathcal{M})$ refines all elements in the marked set $\mathcal{M}$ and keeps conformity; more details are given in Sect. 6.2. To derive convergence rates for AFEM, we require that ADAPT_SURFACE is *t-optimal*, i.e. there exists a constant C such that the set $\mathcal{M}^+$ of all the elements marked for refinement in a call to $\mathsf{ADAPT\_SURFACE}(\mathcal{T}, \Gamma, \tau)$ satisfies

$$\#\mathcal{M}^+ \leq C\tau^{-1/t}, \tag{67}$$

whenever γ belongs to a suitable approximation class, $\mathbb{B}_t$ with $0 < t \leq 1/d$ (see Sect. 8.1). In Sect. 8.3 we show that this assumption is satisfied provided that $\gamma \in W_p^{1+td}(\Gamma_0)$ for some $tp > 1$.

6.2 Module ADAPT_PDE

Given a tolerance $\varepsilon > 0$ and admissible subdivision $\mathcal{T}^+$, the call $[\mathcal{T}, U] = \mathsf{ADAPT\_PDE}(\mathcal{T}^+, \varepsilon)$ outputs a refinement $\mathcal{T} \geq \mathcal{T}^+$ and the associated finite element solution $U \in \mathbb{V}(\mathcal{T})$ such that

$$\eta_{\mathcal{T}}(U) \leq \varepsilon. \tag{68}$$

The module ADAPT_PDE is the standard adaptive sequence:

$[\mathcal{T}, \Gamma] = \mathsf{ADAPT\_PDE}(\mathcal{T}, \varepsilon)$
  $U = \mathsf{SOLVE}(\mathcal{T})$
  $\{\eta_{\mathcal{T}}(U,T)\}_{T\in\mathcal{T}} = \mathsf{ESTIMATE}(\mathcal{T}, U)$
  while $\eta_{\mathcal{T}}(U) > \varepsilon$
    $\mathcal{M} := \mathsf{MARK}(\mathcal{T}, \{\eta_{\mathcal{T}}(U,T)\}_{T\in\mathcal{T}})$
    $\mathcal{T} := \mathsf{REFINE}(\mathcal{T}, \mathcal{M})$
    $\Gamma := \mathcal{F}_{\mathcal{T}}(\Omega)$
    $U = \mathsf{SOLVE}(\mathcal{T})$
    $\{\eta_{\mathcal{T}}(U,T)\}_{T\in\mathcal{T}} = \mathsf{ESTIMATE}(\mathcal{T}, U)$
  end while
  return$(\mathcal{T},\Gamma)$

We describe below the modules SOLVE, ESTIMATE, MARK and REFINE separately.

Procedure SOLVE

This procedure solves the SPD linear system resulting for (38). For simplicity we assume that the linear system is solved exactly. In this context, the approximate solution of the discrete problem can be handled as in [52]. We refer to [43] for a hierarchical basis multigrid preconditioner and to [13] for standard variational and non-variational multigrid algorithms.

Procedure ESTIMATE

Given the Galerkin solution $U \in \mathbb{V}(\mathcal{T})$ of (38) ESTIMATE computes the PDE error indicators $\{\eta_{\mathcal{T}}(U,T)\}_{T\in\mathcal{T}}$. We emphasize that this procedure does not compute the oscillation terms, which are only needed to carry out the analysis.

The equivalence stated in Lemma 5.7 is critical to deduce that the ADAPT_PDE strategy based on the reduction of the error indicators $\eta_{\mathcal{T}}(U)$ is successful in reducing the total error $\|\nabla_\gamma(u-U)\|_{L^2(\gamma)} + \mathrm{osc}_{\mathcal{T}}(f)$. To see this we impose the constraint on the parameter ω

$$\omega \leq \omega_1 := \sqrt{\frac{C_2}{2\Lambda_0^2\Lambda_1}}, \tag{69}$$

and observe that the input $\mathcal{T}^+$ to ADAPT_PDE as well as all inner iterates satisfy, in view of (19),

$$\lambda_\Gamma^2 \leq \Lambda_0^2 \lambda_{\Gamma^+}^2 \leq \frac{C_2}{2\Lambda_1} \varepsilon_k^2.$$

Since $\eta_\mathcal{T}(U) > \varepsilon_k$, we deduce the validity of (61) whence that of (62) within ADAPT_PDE.

Procedure MARK

We rely on an optimal *Dörfler's* marking strategy for the selection of elements. Given the set of indicators $\{\eta_\mathcal{T}(U,T)\}_{T\in\mathcal{T}}$ and a marking parameter $\theta \in (0,1]$, MARK outputs a subset of marked elements $\mathcal{M} \subset \mathcal{T}$ such that

$$\eta_\mathcal{T}(U,\mathcal{M}) \geq \theta \eta_\mathcal{T}(U). \tag{70}$$

In contrast to [46], MARK only employs the error indicators and does not use the oscillation nor surface indicators. We will see that quasi-optimality of AFEM requires that $\mathcal{M}$ be *minimal* and θ sufficiently small.

Procedure REFINE

Given a triangulation $\mathcal{T}$ and a subset $\mathcal{M}$ of marked elements, the call $\mathcal{T}_* = \mathsf{REFINE}(\mathcal{T},\mathcal{M})$ bisects all elements in $\mathcal{M}$ at least $b \geq 1$ times while maintaining mesh conformity, to obtain a new mesh $\mathcal{T}_*$. The new surface Γ_* is obtained by piecewise linear interpolation of the parametrization $\mathcal{X}$ via $\mathcal{F}_{\mathcal{T}_*} = \mathcal{I}_{\mathcal{T}_*}\mathcal{X}$, namely, $\Gamma_* = \mathcal{F}_{\mathcal{T}_*}(\Omega)$.

To ensure conformity of $\mathcal{T}_*$ some additional elements of $\mathcal{T}\setminus\mathcal{M}$ need to be refined. The complexity of the overall refinement algorithm is controlled in a cumulative way, as was proved by P. Binev, W. Dahmen, and R. DeVore for $d = 2$ [11] and R. Stevenson [53] for $d > 2$; see also the survey [49]. The precise statement of this result is in the following lemma.

Lemma 6.10 (Complexity of REFINE) *Assume that $\mathcal{T}_0$ is suitably labeled* (*condition* (*b*) *of Sect.* 4 *in* [53]). *Let $\{\mathcal{T}_k\}_{k\geq 0}$ be any sequence of meshes produced by successive calls $\mathcal{T}_{k+1} = \mathsf{REFINE}(\mathcal{T}_k,\mathcal{M}_k)$. Then, there exists a constant C_6 solely depending on $\mathcal{T}_0$ and the refinement depth b such that*

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq C_6 \sum_{j=0}^{k-1} \#\mathcal{M}_j, \quad \forall k \geq 1. \tag{71}$$

It is worth noticing that the user parameter $b \geq 1$ only entails a minimal refinement, which does not force an interior node property [47, 48] or an extra refinement to improve the surface approximation [46].

Remark 6.1 (Alternative subdivision strategies) For simplicity we only discuss the refinement strategy based on simplex bisection. However, all the results obtained can be extended to any strategy satisfying Conditions 3, 4 and 6 in [12], such as quadrilaterals with hanging nodes.

7 Conditional Contraction Property

The procedure ADAPT_PDE is known to yield a contraction property in the "flat" case. In the present context, however, the surface approximation is responsible for lack of consistency in that the sequence of finite element spaces is no longer nested. This in turn leads to failure of a key orthogonality property between discrete solutions, the Pythagoras property. We have, instead, a perturbation result referred to as *quasi-orthogonality* below. Its proof follows the steps of that for graphs [46, Lemma 4.4]. In this section, we use the notation

$$e^j := \|\nabla_\gamma (u - U^j)\|_{L^2(\gamma)}, \qquad E^j := \|\nabla_\gamma (U^{j+1} - U^j)\|_{L^2(\gamma)},$$
$$\eta^j := \eta_{\mathcal{T}^j}(U^j), \quad \eta^j(\mathcal{M}^j) := \eta_{\mathcal{T}^j}(U^j, \mathcal{M}^j), \quad \lambda^j := \lambda_{\Gamma^j},$$

where $\mathcal{T}^j$ are meshes obtained after each inner iteration of ADAPT_PDE, starting with $\mathcal{T}^0 = \mathcal{T}^+$, and Γ^j, U^j are the corresponding discrete surfaces and Galerkin solutions.

Lemma 7.11 (Quasi-orthogonality) *Let $\Lambda_2 > 0$ be the constant of Lemma 5.8, which solely depends on the Lipschitz constant L of γ and $\|f\|_{L^2(\gamma)}$. Then, for $i = j, j+1$ with $j \geq 0$, we have*

$$(e^j)^2 - \frac{3}{2}(E^j)^2 - \Lambda_2 (\lambda^i)^2 \leq (e^{j+1})^2 \leq (e^j)^2 - \frac{1}{2}(E^j)^2 + \Lambda_2 (\lambda^i)^2. \tag{72}$$

Proof Since the symmetry of the Dirichlet form implies

$$(e^j)^2 = (e^{j+1})^2 + (E^j)^2 + 2\int_\gamma \nabla_\gamma (u - U^{j+1}) \nabla_\gamma^T (U^{j+1} - U^j),$$

we just have to examine the last term. Combining (35), (38), and (39) with Lemma 5.1 yields

$$\left| \int_\gamma \nabla_\gamma (u - U^{j+1}) \nabla_\gamma^T (U^{j+1} - U^j) \right| \lesssim \|f\|_{L^2(\gamma)} \lambda^j E^j,$$

which gives (72) after applying Young's inequality. □

Remark 7.2 (Validity of (72)) Relation (72) is also true for any pair of triangulations $(\mathcal{T}, \mathcal{T}_*)$, with $\mathcal{T}_* \geq \mathcal{T}$, discrete solution $U_* \in \mathbb{V}(\mathcal{T}_*)$ on the finer space, and any discrete function $V \in \mathbb{V}(\mathcal{T})$.

Theorem 7.1 (Conditional Contraction Property) *Let $\theta \in (0,1]$ be the marking parameter of* MARK *and let $\{\mathcal{T}^j, \Gamma^j, U^j\}_{j\geq 0}^{J}$ be a sequence of meshes, piecewise affine surfaces and discrete solutions generated by the procedure* ADAPT_PDE$(\mathcal{T}^0, \varepsilon)$ *within AFEM with tolerance ε, i.e. $\lambda^0 \leq \omega\varepsilon$. Assume that the AFEM parameter ω satisfies*

$$\omega \leq \omega_2 := \frac{\xi\theta^2}{\Lambda_0\sqrt{32\Lambda_2(2\Lambda_3+1)}}, \tag{73}$$

where $\xi = 1 - 2^{-b/d}$ is defined in Lemma 5.8. There exist constants $0 < \alpha < 1$ and $\beta > 0$ such that

$$\big(e^{j+1}\big)^2 + \beta\big(\eta^{j+1}\big)^2 \leq \alpha^2\big(\big(e^j\big)^2 + \beta\big(\eta^j\big)^2\big) \quad \forall 0 \leq j < J. \tag{74}$$

Moreover, the number of inner iterates J of ADAPT_PDE *is uniformly bounded.*

Proof (1) Let $\beta > 0$ be a scaling parameter to be found later. We combine (72) and (64) to write

$$\begin{aligned}
&\big(e^{j+1}\big)^2 + \beta\big(\eta^{j+1}\big)^2 \\
&\quad\leq \big(e^j\big)^2 + \Big(-\frac{1}{2} + \beta\big(1+\delta^{-1}\big)\Lambda_3\Big)\big(E^j\big)^2 \\
&\qquad + \Lambda_2\big(1+\beta\big(1+\delta^{-1}\big)\big)\big(\lambda^j\big)^2 + \beta(1+\delta)\big(\big(\eta^j\big)^2 - \xi\eta^j\big(\mathcal{M}^j\big)^2\big).
\end{aligned}$$

Here $\mathcal{M}^j$ is the set of elements in $\mathcal{T}^j$ marked for refinement at the jth subiteration. To remove the factor of E^j we now choose β dependent on δ, to be

$$\beta\big(1+\delta^{-1}\big)\Lambda_3 = \frac{1}{2} \quad\Rightarrow\quad \beta(1+\delta) = \frac{\delta}{2\Lambda_3}, \tag{75}$$

and thereby obtain

$$\begin{aligned}
\big(e^{j+1}\big)^2 + \beta\big(\eta^{j+1}\big)^2 \leq &\big(e^j\big)^2 + \Lambda_2\big(1+\beta\big(1+\delta^{-1}\big)\big)\big(\lambda^j\big)^2 \\
&+ \beta(1+\delta)\big(\big(\eta^j\big)^2 - \xi\eta^j\big(\mathcal{M}^j\big)^2\big).
\end{aligned}$$

(2) Invoking Dörfler marking (70), we deduce

$$\big(\eta^j\big)^2 - \xi\eta^j\big(\mathcal{M}^j\big)^2 \leq \big(1-\xi\theta^2\big)\big(\eta^j\big)^2.$$

Since the initial mesh $\mathcal{T}^0$ comes from ADAPT_SURFACE we know that $\lambda^0 \leq \omega\varepsilon \leq \omega\eta^j$ for all inner iterations $1 \leq j \leq J$ of ADAPT_PDE. Using (19) yields $\lambda^j \leq$

$\Lambda_0\omega\eta^j$, whence

$$\begin{aligned}\big(e^{j+1}\big)^2+\beta\big(\eta^{j+1}\big)^2 \le \big(e^j\big)^2-\beta(1+\delta)\frac{\xi\theta^2}{2}\big(\eta^j\big)^2\\ +\beta\bigg((1+\delta)\bigg(1-\frac{\xi\theta^2}{2}\bigg)+\Lambda_2\bigg(1+\frac{1}{2\Lambda_3}\bigg)\frac{\Lambda_0^2\omega^2}{\beta}\bigg)\big(\eta^j\big)^2.\end{aligned}$$

Applying the simpler upper bound (63), which is valid for the inner iterates of ADAPT_PDE, and replacing β according to (75), we obtain

$$\big(e^{j+1}\big)^2+\beta\big(\eta^{j+1}\big)^2 \le \alpha_1(\delta)\big(e^j\big)^2+\alpha_2(\delta)\beta\big(\eta^j\big)^2$$

with

$$\begin{aligned}\alpha_1(\delta)^2 &:= 1-\delta\frac{\xi\theta^2}{4\Lambda_3C_3},\\ \alpha_2(\delta)^2 &:= (1+\delta)\bigg(1-\frac{\xi\theta^2}{2}\bigg)+\Lambda_2\bigg(1+\frac{1}{2\Lambda_3}\bigg)\frac{\Lambda_0^2\omega^2}{\beta}.\end{aligned}$$

It remains to prove that δ can be chosen so that $\alpha_2(\delta)^2<1$. We then fix the parameter δ so that

$$(1+\delta)\bigg(1-\frac{\xi\theta^2}{2}\bigg)=1-\frac{\xi\theta^2}{4} \quad\Rightarrow\quad \delta=\frac{\xi\theta^2}{4-2\xi\theta^2}.$$

Now, according to (75), we obtain $\beta=\frac{\xi\theta^2}{2\Lambda_3(4-\xi\theta^2)}\ge\frac{\xi\theta^2}{8\Lambda_3}$ and since $\omega\le\omega_2$ we infer that

$$\Lambda_2\bigg(1+\frac{1}{2\Lambda_3}\bigg)\frac{\Lambda_0^2\omega^2}{\beta}\le\frac{4\Lambda_2(2\Lambda_3+1)}{\xi\theta^2}\Lambda_0^2\omega^2\le\frac{\xi\theta^2}{8}.$$

Hence $\alpha_2^2\le 1-\frac{\xi\theta^2}{8}<1$, and the choice $\alpha:=\max\{\alpha_1,\alpha_2\}<1$ yields the desired estimate (74).

(3) The contraction property (74) guarantees that ADAPT_PDE stops in a finite number of iterations J. To show that J is independent of the iteration counter k of AFEM, take $k\ge1$ and note that before the call ADAPT_PDE$(\mathcal{T}_k^+,\varepsilon_k)$ we have

$$\eta_k=\eta_{\mathcal{T}_k}(U_k)\le\varepsilon_{k-1}=\frac{\varepsilon_k}{\rho},\qquad \lambda_k=\lambda_{\Gamma_k}\le\Lambda_0\lambda_{\Gamma_{k-1}^+}\le\frac{\Lambda_0\omega}{\rho}\varepsilon_k.$$

We next combine (64), with $\delta=1$, and (72) to get

$$\eta_{\mathcal{T}_k^+}\big(U_k^+\big)^2\lesssim\eta_k^2+\lambda_k^2+\big\|\nabla_\gamma\big(U_k^+-U_k\big)\big\|_{L^2(\gamma)}^2\lesssim\eta_k^2+\lambda_k^2+\big\|\nabla_\gamma(u-U_k)\big\|_{L^2(\gamma)}^2,$$

where the hidden constants depend on Λ_2, Λ_3. The bounds on η_k,λ_k, together with (56), yield

$$\big(\eta^0\big)^2=\eta_{\mathcal{T}_k^+}\big(U_k^+\big)^2\lesssim\eta_k^2+\lambda_k^2\lesssim\varepsilon_k^2.$$

Since the stopping condition of ADAPT_PDE is $\eta^J \leq \varepsilon_k$, (74) implies that J is bounded independently of k, as asserted. □

That J is uniformly bounded dictates the complexity of ADAPT_PDE because the most expensive module SOLVE is run just J times. However, this property is not required for the study of cardinality of Sect. 8.

8 Optimal Cardinality

In this section we study the cardinality of AFEM, which is dictated by the regularity of u, f and γ. We first discuss in Sect. 8.1 the best approximation error achievable with piecewise linear polynomials for both surface and PDE solution. We show next in Sect. 8.2 that AFEM delivers the best convergence rate provided the procedure ADAPT_SURFACE is t-optimal, namely it satisfies (67). We conclude in Sect. 8.3 with a greedy algorithm for ADAPT_SURFACE that is t-optimal.

8.1 Approximation Classes

We define classes of functions and surfaces in terms of decay rate of the approximation error as a function of the number of degrees of freedom N. Let $\mathbb{T}_N \subset \mathbb{T} := \mathbb{T}(\mathcal{T}_0)$ be the set of all possible conforming triangulations, generated on γ with at most N elements more than $\mathcal{T}_0$ by successive bisection of $\mathcal{T}_0$:

$$\mathbb{T}_N := \{\mathcal{T} \in \mathbb{T} \mid \#\mathcal{T} - \#\mathcal{T}_0 \leq N\}.$$

Given $v \in H^1_\#(\gamma)$, $f \in L^2(\gamma)$, the notion of *total error*

$$\mathcal{E}_\mathcal{T}(V, f)^2 = \left\|\nabla_\gamma(v - V)\right\|^2_{L^2(\gamma)} + \operatorname{osc}_\mathcal{T}(f)^2$$

is defined in (60). In view of Lemma 5.7 and the fact that AFEM is driven by $\eta_\mathcal{T}(U)$ and λ_Γ, we assess the quality of the best approximation (v, f) with N degrees of freedom in terms of

$$\sigma(N; v, f, \gamma) := \inf_{\mathcal{T}\in\mathbb{T}_N} \inf_{V\in\mathbb{V}(\mathcal{T})} \mathcal{E}_\mathcal{T}(V, f).$$

This is consistent with the approach taken for flat domains in [21, 49]. For $s > 0$, we define the nonlinear (algebraic) approximation class $\mathbb{A}_s(\gamma)$ to be

$$\mathbb{A}_s(\gamma) := \left\{(v, f) \;\middle|\; |v, f|_{\mathbb{A}_s} := \sup_{N\geq 1}\left(N^s \sigma(N; v, f, \gamma)\right) < \infty\right\}.$$

We emphasize that the approximability of the surface γ only appears implicitly by measuring the errors on γ. In fact, the definition of data oscillation (55), and in

particular the specific choice of F_Γ, implies that $\operatorname{osc}_{\mathcal{T}}^2(f)$ entails the approximation of f by piecewise constants on γ but does not include the approximation of γ by Γ; this is quite different for higher order approximations of γ, as is shown in [17]. On the other hand, the generic range of s is dictated by polynomial degree, namely $0 < s \leq 1/d$.

An alternative and useful definition to $(u, f) \in \mathbb{A}_s(\gamma)$ is as follows: given $\varepsilon > 0$, there exists a mesh $\mathcal{T}_\varepsilon \in \mathbb{T}(\mathcal{T}_0)$ with $\mathcal{T}_\varepsilon \geq \mathcal{T}_0$ and a discrete function $V_\varepsilon \in \mathbb{V}(\mathcal{T}_\varepsilon)$ so that

$$\left\|\nabla_\gamma (u - V_\varepsilon)\right\|_{L^2(\gamma)}^2 + \operatorname{osc}_{\mathcal{T}_\varepsilon}(f)^2 \leq \varepsilon^2, \qquad \#\mathcal{T}_\varepsilon - \#\mathcal{T}_0 \leq |u, f|_{\mathbb{A}_s}^{\frac{1}{s}} \varepsilon^{-\frac{1}{s}}; \tag{76}$$

$\Gamma_\varepsilon = \mathcal{F}_{\mathcal{T}_\varepsilon}(\Omega)$ might not be a good approximating of γ. In fact, approximations of (u, f) and γ are handled separately. The characterization of $\mathbb{A}_s(\gamma)$ in terms of Besov regularity is an open issue.

Similarly, for surfaces and $t > 0$, we define the approximation class as follows:

$$\mathbb{B}_t := \left\{ \gamma \in W_\infty^1 \;\middle|\; |\gamma|_{\mathbb{B}_t} := \sup_{N \geq 1} N^t \inf_{\mathcal{T} \in \mathbb{T}_N} \lambda_\Gamma < \infty \right\}.$$

This means that surfaces in $\mathbb{B}_t$ are parametrized by Lipschitz maps $\mathcal{X} : \Omega \to \mathbb{R}^{d+1}$ which can in turn be approximated with rate N^{-t} in W_∞^1 over Ω with N degrees of freedom. In Sect. 8.3, we give a constructive greedy algorithm that realizes this rate provided γ belong to a suitable Sobolev space in the nonlinear scale of W_∞^1. The generic range of exponents t for linear elements, or equivalently polyhedral surfaces Γ, is $0 < t \leq 1/d$.

8.2 Convergence Rates

We now prove that AFEM achieves the asymptotic decay rate $\min\{s, t\}$, dictated by the classes $\mathbb{A}_s(\gamma)$ and $\mathbb{B}_t$, but without ever using either s or t in its formulation. We establish the link between the performance of AFEM and the best possible error by adapting a clever idea of R. Stevenson [52] for the Laplace operator, further extended by J.M. Cascón et al. [21] to general elliptic PDE, in flat domains; we refer to the survey [49] for a thorough discussion. The insight is that

any marking strategy that reduces the total error relative to its current value must contain a substantial portion of the error estimator, and so it can be related to Dörfler Marking. (77)

Exploiting next the minimality of Dörfler marking enables us to compare meshes generated by AFEM with the best meshes within $\mathbb{T}$. The approach of [21, 49, 52] does not apply directly in the present context because of the consistency error due to surface interpolation. We account for this discrepancy below upon making the

parameter ω of ADAPT_SURFACE sufficiently small. Let

$$\omega_3 := \frac{C_4}{\Lambda_0\sqrt{3\Lambda_1 + 2\Lambda_2}}, \qquad \omega_4 := \frac{C_4}{2\Lambda_0}\sqrt{\left(1 - \frac{\theta^2}{\theta_*^2}\right)\frac{1}{\Lambda_2}} \tag{78}$$

be two thresholds for ω to be used next and θ_* be a threshold for the Dörfler parameter θ

$$\theta_* := \frac{C_4}{\sqrt{2 + 3C_1}}; \tag{79}$$

since $C_4 = \sqrt{C_2/2}$ and $C_2 \le C_1$, we see that $\theta_* < 1$.

Lemma 8.12 (Dörfler marking) *Let λ_{Γ_0} satisfy (49), and the parameters θ and ω satisfy*

$$0 < \theta < \theta_*, \quad 0 < \omega \le \min\{\omega_1, \omega_3\}, \tag{80}$$

where θ_*, ω_3 are defined in (78), (79), and ω_1 in (69). Let $\mu = \frac{1}{2}(1 - \frac{\theta^2}{\theta_*^2})$ and $(\Gamma, \mathcal{T}, U)$ be the approximate surface, mesh and discrete solution produced by an inner iterate of ADAPT_PDE. *If $(\Gamma_*, \mathcal{T}_*, U_*)$ is a surface-mesh-solution triple with $\mathcal{T}_* \ge \mathcal{T}$, such that the total error satisfies*

$$\mathcal{E}_{\mathcal{T}_*}(U_*, f) \le \mu \mathcal{E}_{\mathcal{T}}(U, f), \tag{81}$$

then the refined set $\mathcal{R} := \mathcal{R}_{\mathcal{T}\to\mathcal{T}_*}$ satisfies Dörfler property with parameter θ, namely

$$\eta_{\mathcal{T}}(U, \mathcal{R}) \ge \theta \eta_{\mathcal{T}}(U). \tag{82}$$

Proof Using the notation $e(U) = \|\nabla_\gamma(u - U)\|_{L^2(\gamma)}$, we proceed as in [21, Lemma 5.9]. Since $\omega \le \omega_1$, we combine the lower bound of (62) with (81) to write

$$\begin{aligned}(1-\mu)C_4^2\eta_{\mathcal{T}}(U)^2 &\le (1-\mu)\big(e(U)^2 + \operatorname{osc}_{\mathcal{T}}(f)^2\big)\\ &\le e(U)^2 - e(U_*)^2 + \operatorname{osc}_{\mathcal{T}}(f)^2 - \operatorname{osc}_{\mathcal{T}_*}(f)^2.\end{aligned}$$

We now estimate separately error and oscillation terms. According to (72) and (58), we obtain

$$\begin{aligned}e(U)^2 - e(U_*)^2 &\le \frac{3}{2}\big\|\nabla_\gamma(U_* - U)\big\|^2_{L^2(\gamma)} + \Lambda_2\lambda_{\mathcal{T}}^2\\ &\le \frac{3}{2}C_1\eta_{\mathcal{T}}(U, \mathcal{R})^2 + \left(\frac{3}{2}\Lambda_1 + \Lambda_2\right)\lambda_{\Gamma}^2.\end{aligned}$$

On the other hand, the data oscillation terms verify

$$\operatorname{osc}_{\mathcal{T}}(f)^2 - \operatorname{osc}_{\mathcal{T}_*}(f)^2 \le \operatorname{osc}_{\mathcal{T}}(f, \mathcal{R})^2 \le \eta_{\mathcal{T}}(U, \mathcal{R})^2$$

because they coincide over $\mathcal{T}\backslash\mathcal{R}$ and the estimator dominates the oscillation locally (see (59)). Since $(\Gamma, \mathcal{T})$ is produced within ADAPT_PDE, we have $\eta_{\mathcal{T}}(U) > \varepsilon$ and $\lambda_{\Gamma^+} \leq \omega\varepsilon$, whence

$$\lambda_\Gamma \leq \Lambda_0 \lambda_{\Gamma^+} \leq \Lambda_0 \omega \varepsilon \leq \Lambda_0 \omega \eta_{\mathcal{T}}(U).$$

Collecting these three estimates, and using that $\omega \leq \omega_3$, we infer that

$$\left(1 + \frac{3}{2}C_1\right)\eta_{\mathcal{T}}(U,\mathcal{R})^2 \geq \left((1-\mu)C_4^2 - \Lambda_0^2\omega^2\left(\frac{3}{2}\Lambda_1 + \Lambda_2\right)\right)\eta_{\mathcal{T}}(U)^2$$
$$\geq (1-2\mu)\frac{C_4^2}{2}\eta_{\mathcal{T}}(U)^2.$$

Finally, the asserted estimate (82) is a consequence of the definition of θ_*, μ and $\theta < \theta_*$. □

Lemma 8.13 (Cardinality of $\mathcal{M}$) *Let λ_{Γ_0} satisfy (49) and the procedure* MARK *select a set $\mathcal{M}$ with minimal cardinality. Let the parameters θ and ω satisfy*

$$0 < \theta < \theta_*, \quad 0 < \omega \leq \min\{\omega_1, \omega_4\} \tag{83}$$

with $\theta_*, \omega_4, \omega_1$ given in (79), (78), and (69), respectively. Let u be the solution of (35), and let $(\Gamma, \mathcal{T}, U)$ be produced within ADAPT_PDE. *If $(u, f) \in \mathbb{A}_s(\gamma)$, then*

$$\#\mathcal{M} \lesssim |u, f|_s^{\frac{1}{s}} \mathcal{E}_{\mathcal{T}}(U, f)^{-\frac{1}{s}}.$$

Proof We set

$$\delta^2 = \hat{\mu}\mathcal{E}_{\mathcal{T}}(U,f)^2 = \hat{\mu}\big(e(U)^2 + \operatorname{osc}_{\mathcal{T}}(f)^2\big),$$

for $0 < \hat{\mu} < \mu = \frac{1}{2}(1 - \frac{\theta^2}{\theta_*^2}) < 1$ sufficiently small to be determined later. Since $(u, f) \in \mathbb{A}_s(\gamma)$, there exists a pair $(\Gamma_\delta, \mathcal{T}_\delta)$ with $\mathcal{T}_\delta \geq \mathcal{T}_0$ and a $V_\delta \in \mathbb{V}(\mathcal{T}_\delta)$ such that

$$\#\mathcal{T}_\delta - \#\mathcal{T}_0 \lesssim |u, f|_s^{\frac{1}{s}} \delta^{-\frac{1}{s}}, \qquad e(V_\delta)^2 + \operatorname{osc}_{\mathcal{T}_\delta}(f)^2 \leq \delta^2. \tag{84}$$

Let $\mathcal{T}_* = \mathcal{T} \oplus \mathcal{T}_\delta$ be the overlay of $\mathcal{T}$ and $\mathcal{T}_\delta$, which satisfies [21, Lemma 3.7], [49]

$$\#\mathcal{T}_* \leq \#\mathcal{T} + \#\mathcal{T}_\delta - \#\mathcal{T}_0. \tag{85}$$

Let $U_* \in \mathbb{V}(\mathcal{T}_*)$ be the corresponding Galerkin solution. We observe that $\mathcal{T}_* \geq \mathcal{T}_\delta, \mathcal{T}$, and invoke the upper bound of (72) in conjunction with (65) to write

$$e(U_*)^2 + \operatorname{osc}_{\mathcal{T}_*}(f)^2 \leq e(V_\delta)^2 + \Lambda_2\lambda_\Gamma^2 + C_5^2 \operatorname{osc}_{\mathcal{T}_\delta}(f)^2.$$

We recall that $\lambda_\Gamma \leq \Lambda_0\lambda_{\Gamma^+} < \Lambda_0\omega\varepsilon$, in view of (19) and (66), and $\eta_{\mathcal{T}}(U) > \varepsilon$ because of (68). Combining this with (62), we arrive at

$$\lambda_\Gamma^2 \leq \frac{\Lambda_0^2\omega^2}{C_4^2}\big(e(U)^2 + \operatorname{osc}_{\mathcal{T}}(f)^2\big) = \frac{\Lambda_0^2\omega^2}{\hat{\mu}C_4^2}\delta^2.$$

Using the fact that $C_5 \geq 1$ and $\omega \leq \omega_4$, we choose $\hat{\mu} = \frac{\mu}{2C_5^2}$ to end up with

$$e(U_*)^2 + \operatorname{osc}_{\mathcal{T}_*}(f)^2 \leq \left(C_5^2 + \frac{\Lambda_0^2 \Lambda_2 \omega^2}{\hat{\mu} C_4^2}\right)\delta^2 = \mu\big(e(U)^2 + \operatorname{osc}_{\mathcal{T}}^2(f)\big).$$

We thus deduce from Lemma 8.12 that the subset $\mathcal{R} := \mathcal{R}_{\mathcal{T}\to\mathcal{T}_*} \subset \mathcal{T}$ satisfies Dörfler property (82). Since the set $\mathcal{M} \subset \mathcal{T}$ also satisfies this property, but with minimal cardinality, we infer that

$$\#\mathcal{M} \leq \#\mathcal{R} \leq \#\mathcal{T}_* - \#\mathcal{T} \leq \#\mathcal{T}_\delta - \#\mathcal{T}_0 \lesssim |u, f|_s^{\frac{1}{s}} \delta^{-\frac{1}{s}}.$$

The asserted estimate finally follows upon using the definition of δ. □

The quasi-optimal cardinality of AFEM is a direct consequence of Lemma 8.13 and Theorem 7.1. We prove this next.

Theorem 8.2 (Convergence rate of AFEM) *Let $\gamma \in \mathbb{B}_t$ and $(u, f) \in \mathbb{A}_s(\gamma)$ for some $0 < t, s \leq n/d$. Let $\varepsilon_0 \leq (6\omega\Lambda_0 L^3)^{-1}$ be the initial tolerance, and the parameters θ, ω satisfy*

$$0 < \theta \leq \theta_*, \qquad 0 < \omega \leq \omega_* := \min\{\omega_1, \omega_2, \omega_3, \omega_4\}, \tag{86}$$

where $\theta_*, \omega_1, \ldots, \omega_4$ are given in (79), (69), (73), and (78), respectively. Let the procedure MARK *select sets with minimal cardinality, and the procedure* ADAPT_SURFACE *be t-optimal on the surface γ. Let u be the solution of (35) and $\{\Gamma_k, \mathcal{T}_k, U_k\}_{k\geq 0}$ a sequence of approximate surfaces, meshes and discrete solution generated by the outer loop of AFEM.*

Then there exists a constant C, depending on the Lipschitz constant L of γ, $\|f\|_{L^2(\gamma)}$, the refinement depth b, the initial triangulation $\mathcal{T}_0$, and AFEM parameters (θ, ω, ρ) such that

$$e(U_k) + \operatorname{osc}_{\mathcal{T}_k}(f) + \omega^{-1}\lambda_{\Gamma_k} \leq C\big(|u, f|_{\mathbb{A}_s}^{\frac{r}{s}} + \omega^{-r}|\gamma|_{\mathbb{B}_t}^{\frac{r}{t}}\big)(\#\mathcal{T}_k - \#\mathcal{T}_0)^{-r}, \tag{87}$$

with $r = \min\{s, t\}$.

Proof We start by noting that since

$$\omega\varepsilon_0 \leq \frac{1}{6\Lambda_0 L^3}$$

the first output of ADAPT_SURFACE fulfills $\lambda_{\Gamma_0^+} \leq \frac{1}{6\Lambda_0}$ which is (49) and implies that $\mathbb{T}(\mathcal{T}_0^+)$ is shape regular.

There are two instances where elements are added, inside ADAPT_SURFACE and ADAPT_PDE. For ADAPT_SURFACE we make the assumption (67) of t-optimality:

$$\#\mathcal{M}_k^+ \lesssim \omega^{-\frac{1}{t}} |\gamma|_{\mathbb{B}_t}^{\frac{1}{t}} \varepsilon_k^{-\frac{1}{t}}.$$

For ADAPT_PDE, Lemma 8.13 yields

$$\#\mathcal{M}_k^j \lesssim |u, f|_{\mathbb{A}_s}^{\frac{1}{s}} \big(e\big(U_k^j\big) + \operatorname{osc}_{\mathcal{T}_k^j}(f)\big)^{-\frac{1}{s}} \quad 0 \le j < J,$$

with j denoting the inner loop iteration counter. Since the inner iterates of ADAPT_PDE satisfy Theorem 7.1 and

$$e\big(U_k^j\big) + \operatorname{osc}_{T_k^j}(f) \approx e\big(U_k^j\big) + \eta_{\mathcal{T}_k^j}\big(U_k^j\big),$$

we deduce that

$$\big(e\big(U_k^j\big) + \operatorname{osc}_{\mathcal{T}_k^j}(f)\big)^{-\frac{1}{s}} \lesssim \alpha^{\frac{J-j-1}{s}} \big(e\big(U_k^{J-1}\big) + \eta_{\mathcal{T}_k^{J-1}}(f)\big)^{-\frac{1}{s}} \le \alpha^{\frac{J-j-1}{s}} \varepsilon_k^{-\frac{1}{s}}.$$

This implies

$$\sum_{j=0}^{J-1} \#\mathcal{M}_k^j \lesssim |u, f|_{\mathbb{A}_s}^{\frac{1}{s}} \varepsilon_k^{-\frac{1}{s}} \sum_{j=0}^{J-1} \alpha^{\frac{J-j-1}{s}} \lesssim |u, f|_{\mathbb{A}_s}^{\frac{1}{s}} \varepsilon_k^{-\frac{1}{s}}.$$

To do a full counting argument, we resort to the crucial estimate (71), which combined with the estimates above and the relation $\varepsilon_{k+1} = \rho\varepsilon_k$ of step 3 of AFEM gives

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \le C_6 \sum_{i=0}^{k-1} \left(\#\mathcal{M}_i^+ + \sum_{j=0}^{J-1} \#\mathcal{M}_i^j \right) \lesssim C_6 \big(\omega^{-\frac{1}{t}} |\gamma|_{\mathbb{B}_t}^{\frac{1}{t}} + |u, f|_{\mathbb{A}_s}^{\frac{1}{s}}\big) \sum_{i=0}^{k-1} \varepsilon_i^{-\frac{1}{r}},$$

where $r = \min\{s, t\}$. Since $\rho < 1$, we obtain $\sum_{i=0}^{k-1} \varepsilon_i^{-\frac{1}{r}} = \varepsilon_{k-1}^{-\frac{1}{r}} \sum_{i=0}^{k-1} \rho^{\frac{i}{r}} \lesssim \varepsilon_k^{-\frac{1}{r}}$, whence

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \lesssim C_6 \big(\omega^{-\frac{1}{t}} |\gamma|_{\mathbb{B}_t}^{\frac{1}{t}} + |u, f|_{\mathbb{A}_s}^{\frac{1}{s}}\big) \varepsilon_k^{-\frac{1}{r}}.$$

Moreover, the stopping criteria (66) and (68) guarantee that

$$e(U_k) + \operatorname{osc}_{\mathcal{T}_k}(f) + \omega^{-1} \lambda_{\Gamma_k} \le C\varepsilon_k,$$

which implies the desired estimate (87). □

Besides the condition $\omega \le \omega_*$ in (86), the right-hand side of (87) suggests that ω should not be too small to optimize this bound. An optimal choice of ω for the case $s = t$, which unfortunately is not computable, appears to be

$$\omega = \min\big\{\omega_*, |u, f|_{\mathbb{A}_s} |\gamma|_{\mathbb{B}_s}^{-1}\big\}.$$

8.3 Greedy Algorithm

To conclude we show that ADAPT_SURFACE is t-optimal provided γ belongs to W_p^{1+td}, which is just above the nonlinear Sobolev scale of W_∞^1 for polynomial degree 1 in d dimensions:

$$\mathrm{sob}\big(W_\infty^1\big) = 1 - \frac{d}{\infty} = 1 < \mathrm{sob}\big(W_p^{1+td}\big) = 1 + td - \frac{d}{p} \quad \Rightarrow \quad tp > 1.$$

Proposition 8.1 (Greedy algorithm) *Let γ be piecewise of class $W_p^{1+td}(\Gamma_0)$, with $tp > 1$, $t \le 1/d$, and globally of class W_∞^1. Then module $[\mathcal{T}^+, \Gamma^+] =$ ADAPT_SURFACE$(\mathcal{T}, \Gamma, \tau)$ terminates in a finite number of steps and the set $\mathcal{M}^+$ of marked elements satisfies*

$$\#\mathcal{M}^+ \le C|\gamma|^{1/t}_{W_p^{1+td}(\Gamma_0)}\tau^{-1/t},$$

where $|\gamma|_{W_p^{1+td}(\Gamma_0)} = (\sum_{i=1}^I |\mathcal{X}^i|^p_{W_p^{1+tp}(\Omega)})^{1/p}$. Moreover, $\gamma \in \mathbb{B}_t$ and

$$|\gamma|_{\mathbb{B}_t} \lesssim |\gamma|_{W_p^{1+td}(\Gamma_0)}.$$

Proof We first observe that $W_p^{1+td} \subset W_\infty^1 \subset C^0$ so that the Lagrange interpolation operator $\mathcal{I}_\mathcal{T}$ is well defined. In addition, for an approximation pair $(\Gamma, \mathcal{T})$ local interpolation estimates give

$$\lambda_\Gamma(T) \lesssim h_T^r |\mathcal{X}|_{W_p^{1+td}(\widehat{T})}, \quad \forall T \in \mathcal{T} = \mathcal{T}(\Gamma), \tag{88}$$

for $r = \mathrm{sob}(W_p^{1+td}) - \mathrm{sob}(W_\infty^1) = td - \frac{d}{p} > 0$. This shows that ADAPT_SURFACE terminates in finite number of steps, say m.

To prove that ADAPT_SURFACE is t-optimal, namely to show (67), let $\mathcal{M}^+ = \mathcal{M}_0 \cup \cdots \cup \mathcal{M}_{m-1}$ be the set of marked elements. We organize the elements in $\mathcal{M}^+$ by size in such a way that allows for a counting argument. Let $\mathcal{P}_j$ be the set of elements T of $\mathcal{M}^+$ with size

$$2^{-(j+1)} \le |T| < 2^{-j} \quad \Rightarrow \quad 2^{-(j+1)/d} \le h_T < 2^{-j/d}.$$

We recall that $|T|$ is the measure of $\overline{T}$, the preimage of T in the initial triangulation $\mathcal{T}_0$. We proceed in several steps.

(1) We first observe that all T's in $\mathcal{P}_j$ are *disjoint*. This is because if $T_1, T_2 \in \mathcal{P}_j$ and $\mathring{T}_1 \cap \mathring{T}_2 \neq \emptyset$, then one of them is contained in the other, say $T_1 \subset T_2$, due to the bisection procedure. Thus $|T_1| \le \frac{1}{2}|T_2|$, contradicting the definition of $\mathcal{P}_j$. This implies

$$2^{-(j+1)}\#\mathcal{P}_j \le |\Gamma_0| \quad \Rightarrow \quad \#\mathcal{P}_j \le |\Gamma_0| 2^{j+1}. \tag{89}$$

(2) In light of (88), we have for $T \in \mathcal{P}_j$

$$\tau \leq \lambda_\Gamma(T) \lesssim 2^{-(j/d)r} |\mathcal{X}|_{W_p^{1+td}(\widehat{T})}.$$

Therefore $\tau^p \# \mathcal{P}_j \lesssim 2^{-(j/d)rp} \sum_{T \in \mathcal{P}_j} |\mathcal{X}|^p_{W_p^{1+td}(\widehat{T})} \leq 2^{-(j/d)rp} |\gamma|^p_{W_p^{1+td}(\Gamma_0)}$, whence

$$\# \mathcal{P}_j \lesssim \tau^{-p} 2^{-(j/d)rp} |\gamma|^p_{W_p^{1+td}(\Gamma_0)}. \tag{90}$$

(3) The two bounds for $\# \mathcal{P}_j$ in (89) and (90) are complementary. The first is good for j small whereas the second is suitable for j large (think of $\tau \ll 1$). The crossover takes place for j_0 such that

$$2^{j_0+1} |\Gamma_0| = \tau^{-p} 2^{-j_0 rp/d} |\mathcal{X}|^p_{W_p^{1+td}(\Omega)} \quad \Rightarrow \quad 2^{j_0} \approx \tau^{-1/t} \frac{|\gamma|^{1/t}_{W_p^{1+td}(\Gamma_0)}}{|\Gamma_0|^{1/tp}}.$$

(4) We now compute

$$\# \mathcal{M}^+ = \sum_j \# \mathcal{P}_j \lesssim \sum_{j \leq j_0} 2^j |\Gamma_0| + \tau^{-p} |\gamma|^p_{W_p^{1+t}(\Gamma_0)} \sum_{j > j_0} \big(2^{-rp/d}\big)^j.$$

Since $\sum_{j \leq j_0} 2^j \approx 2^{j_0}$, $\sum_{j > j_0} (2^{-rp/d})^j \lesssim 2^{-(rp/d) j_0} = 2^{(1-tp) j_0}$ we can write

$$\begin{aligned} \# \mathcal{M}^+ &\lesssim \big(\tau^{-1/t} + \tau^{-p} \tau^{-1/t+p}\big) |\Gamma_0|^{1-1/tp} |\gamma|^{1/t}_{W_p^{1+td}(\Gamma_0)} \\ &\approx \tau^{-1/t} |\Gamma_0|^{1-1/tp} |\gamma|^{1/t}_{W_p^{1+td}(\Gamma_0)}. \end{aligned}$$

(5) Upon termination, $\lambda_{\Gamma^+} \leq \tau$ and $\# \mathcal{M}^+ \lesssim \tau^{-1/t} |\gamma|^{1/t}_{W_p^{1+td}(\Gamma_0)}$, which is valid regardless of the input $\mathcal{T}$ of ADAPT_SURFACE. If we take $\mathcal{T} = \mathcal{T}_0$ and invoke Lemma 6.10 we deduce that

$$\# \mathcal{T}^+ - \# \mathcal{T}_0 \lesssim \tau^{-1/t} |\gamma|^{1/t}_{W_p^{1+td}(\Gamma_0)} \quad \Rightarrow \quad \lambda_{\Gamma^+} \big(\# \mathcal{T}^+ - \# \mathcal{T}_0\big)^t \lesssim |\gamma|_{W_p^{1+td}(\Gamma_0)},$$

and that $\gamma \in \mathbb{B}_t$. This concludes the proof. □

9 Asymptotics: Role of ω

In order to analyze the role of ω in the convergence rate of AFEM and its performance, we solve the problem

$$-\Delta_\gamma u = 1, \quad \text{in } \gamma, \qquad u = 0, \quad \text{on } \partial\gamma,$$

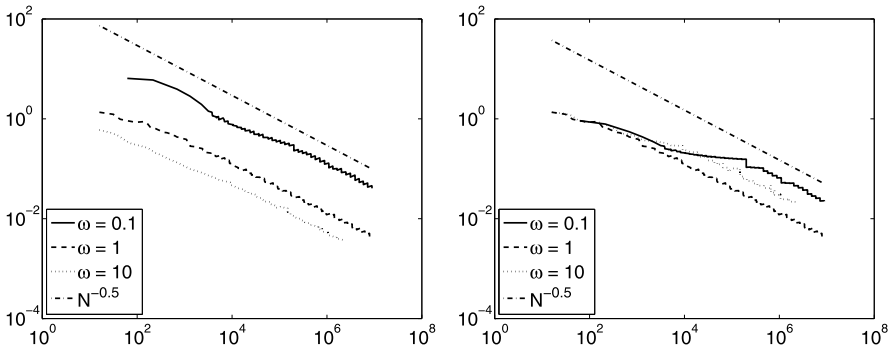


Fig. 12 $\eta_k + \lambda_k/\omega$ (*left*) and $\eta_k + \lambda_k$ (*right*) versus the number of elements in logarithmic scale for $\omega = 0.1, 1, 10$. We observe that $\eta_k + \lambda_k/\omega$ decays as $N^{-0.5}$ right from the beginning, whereas $\eta_k + \lambda_k$ shows the same decay after the meshes have some refinement, depending on the value of ω. Our theory predicts the decay of $N^{-0.5}$ for both notions of total error if ω is sufficiently small, but the best relation between the error $\eta_k + \lambda_k$ and #DOFs is obtained for $w = 1$, which is *not so small*

where γ is the graph of class $C^{1,\alpha}$ given by

$$z(x, y) = \left(0.75 - x^2 - y^2\right)_+^{1+\alpha},$$

over the flat domain $\Omega = (0, 1)^2$, and consider two cases $\alpha = 3/5$ and $\alpha = 2/5$.

It turns out that $z \in W_p^{1+2t}(\Omega)$ for $t < (\alpha + \frac{1}{p})/2$. Moreover, to enforce the gap $\mathrm{sob}(W_p^{1+2t}) - \mathrm{sob}(W_\infty^1) > 0$ with $\mathrm{sob}(W_p^{1+2t}) = 1 + 2t - \frac{2}{p}$ and $\mathrm{sob}(W_\infty^1) = 1$ we need $tp > 1$. These conditions can be achieved provided (see Sect. 8.3)

$$\alpha = 3/5\colon \quad t = 1/2,\ p > 2, \qquad \Rightarrow \quad z \in \mathbb{B}_{\frac{1}{2}}$$

$$\alpha = 2/5\colon \quad t < 2/5,\ p > 5/2, \quad \Rightarrow \quad z \in \mathbb{B}_t, \quad \forall t < 2/5.$$

On the other hand $(u, f) \in \mathbb{A}_{\frac{1}{2}}$ in both cases. This is a consequence of the fact that $\Delta_\gamma u = f$ can be written in the parameter domain Ω as $\frac{1}{q}\widehat{\mathrm{div}}(q\mathbf{g}^{-1}\widehat{\nabla}^T u) = f$, with coefficient matrix $\mathbf{A} = q\mathbf{g}^{-1} \in C^\alpha(\bar{\Omega}) \cap W_p^1(\Omega)$ and $1 < p < \frac{1}{1-\alpha}$; see (47). Extending u and qf by odd reflection and $\mathbf{A}$ by even reflection to the unit squares around Ω, u is a solution to $-\mathrm{div}(\mathbf{A}\nabla^T u) = qf$ on the ball B centered at $(1/2, 1/2)$ and radius 1, with coefficient $\mathbf{A} \in C^\alpha$ and right-hand side in L^∞. By Theorem 3.13 in [40] this implies $\nabla u \in C^\alpha(\bar{B})$, and thus

$$\mathbf{A} : D^2 u = f + \mathrm{div}\,\mathbf{A} \cdot \nabla u \in L^p(B).$$

Applying Calderón-Zygmund theory we obtain $u \in W_p^2(\Omega)$ [39, Theorem 9.11], whence $(u, f) \in \mathbb{A}_{\frac{1}{2}}$ [49, Sect. 5.4].

In the following, we use the notation $\eta_k := \eta_{\mathcal{T}_k}(U_k)$, and $\lambda_k := \lambda_{\Gamma_k}$.

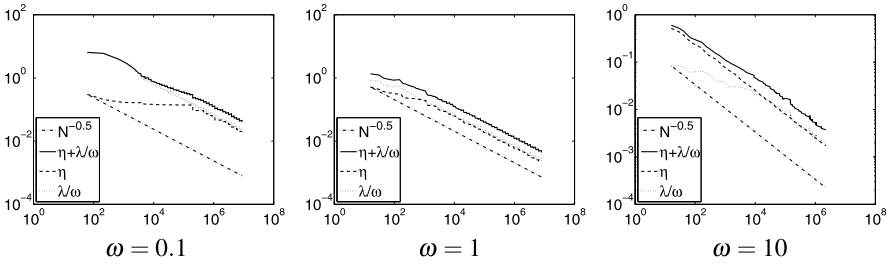


Fig. 13 η_k, λ_k/ω and $\eta_k + \lambda_k/\omega$ for $\omega = 0.1$ (*left*), $\omega = 1$ (*middle*), and $\omega = 10$ (*right*)

9.1 Case $\alpha = 3/5$

We recall that in this case $\gamma \in \mathbb{B}_{1/2}$. Since the pair $(u, f) \in \mathbb{A}_{1/2}$ we expect a decay of $\eta_k + \lambda_k/\omega$ proportional to $N_k^{-1/2}$, where $N_k = \#\mathcal{T}_k - \#\mathcal{T}_0$. In Fig. 12 we plot $\eta_k + \lambda_k/\omega$ (left) and $\eta_k + \lambda_k$ (right) versus the number of elements in logarithmic scale for $\omega = 0.1, 1, 10$, and observe that in the three cases both notions of error decay (asymptotically) as $N^{-1/2}$.

In Fig. 13 we show the behavior of the different indicators η_k, λ_k/ω and their sum, for the three values of ω considered above. We observe the following:

$\omega = 0.1$. At the beginning $\eta_k \ll \lambda_k/\omega$, thus λ_k in ADAPT_SURFACE guides the refinement initially, and η_k decreases very slowly because ADAPT_PDE exits without refining. The indicators η_k and λ_k/ω are of comparable size when the number of elements is around $2 \cdot 10^5$, when the refinement starts to occur due to both λ_k and η_k and both quantities decrease as $N^{-0.5}$.

$\omega = 1$. At the beginning $\eta_k < \lambda_k/\omega$ and the behavior is similar to the case $\omega = 0.1$. When the meshes have about 10^3 elements the curves for η_k and λ_k/ω meet and they both start to decrease at the optimal rate $N^{-0.5}$.

$\omega = 10$. At the beginning $\lambda_k/\omega < \eta_k$, and the situation is opposite to the case of ω small. The refinement is initially guided by η_k in ADAPT_PDE, and λ_k decreases very slowly because ADAPT_SURFACE exits without refining. The two curves for η_k and λ_k/ω meet when the meshes have about 10^4 elements, and they both start to decrease at the optimal rate $N^{-0.5}$.

In Fig. 14 we show three meshes after 10, 20 and 30 refinements have been performed, with 192, 1216 and 5564 elements, respectively.

9.2 Case $\alpha = 2/5$

We recall that in this case $\gamma \in \mathbb{B}_{0.4}$, whereas the pair $(u, f) \in \mathbb{A}_{1/2}$. We thus expect a decay of $\eta_k + \lambda_k/\omega$ proportional to $N^{-0.4}$. In Fig. 15 we plot $\eta_k + \lambda_k/\omega$ (left) and $\eta_k + \lambda_k$ (right) versus the number of elements in logarithmic scale for $\omega = 0.1, 1, 10$,

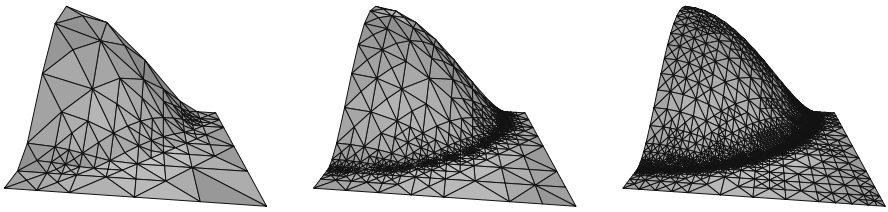


Fig. 14 Meshes after 10, 20 and 30 refinements have been performed, $C^{1.6}$-surface, with $\omega = 1$. They are composed of 192, 1216 and 5564 elements, respectively

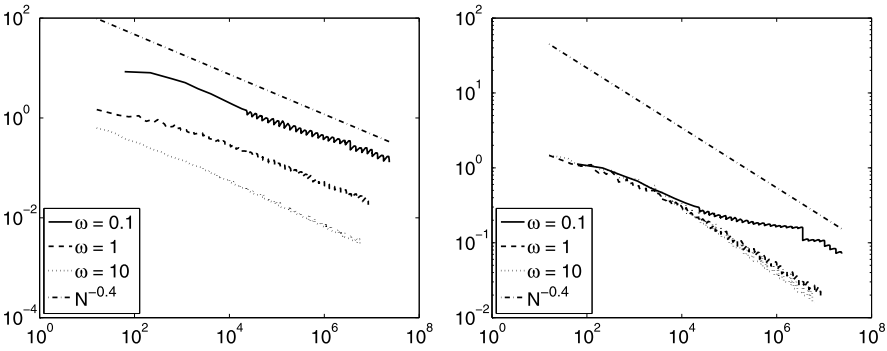


Fig. 15 $\eta_k + \lambda_k/\omega$ (*left*) and $\eta_k + \lambda_k$ (*right*) versus the number of elements in logarithmic scale for $\omega = 0.1, 1, 10$. We observe that $\eta_k + \lambda_k/\omega$ decays as $N^{-0.4}$ right from the beginning, whereas $\eta_k + \lambda_k$ shows the same decay after the meshes have some refinement, depending on the value of ω. Our theory predicts the decay of $N^{-0.4}$ for both notions of total error if ω is sufficiently small. The best relation between the error $\eta_k + \lambda_k$ seems to occur for $\omega = 1$ and $\omega = 10$

and observe that in the three cases both notions of error decay (asymptotically) as $N^{-0.4}$.

In Fig. 16 we show the behavior of η_k, λ_k/ω and their sum, for the same values of ω. We observe the following:

$\omega = 0.1$. At the beginning $\eta_k \ll \lambda_k/\omega$, thus λ_k in ADAPT_SURFACE guides the refinement initially, and η_k decreases very slowly because ADAPT_PDE exits without refining. The asymptotic regime starts when both indicators have a comparable magnitude, and both quantities decrease as $N^{-0.4}$. This instance is reached when the meshes have more than 10^6 elements, because λ_k/ω decreases more slowly than in the previous example, and takes longer to reach the initial value of η_k.

$\omega = 1$. This case is similar to the previous one, with the change of behavior occurring when the meshes have 10^4 elements.

$\omega = 10$. The situation now is opposite to the previous cases of ω small. At the beginning the refinement is initially guided by η_k in ADAPT_PDE, and λ_k decreases very slowly because ADAPT_SURFACE exits without refining. It is interesting to notice that η_k decreases as $N^{-0.5}$ in this transient initial phase. When the meshes have about 10^3 elements both indicators are of comparable size, and the overall

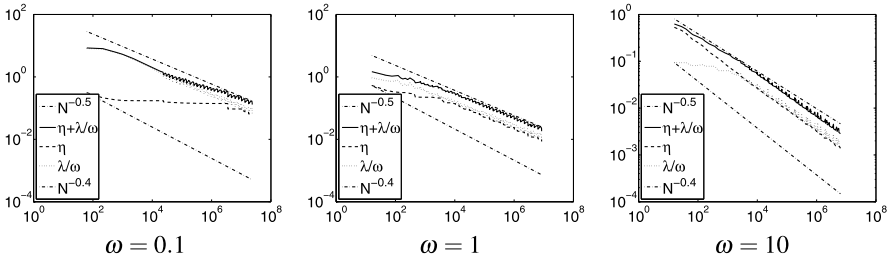


Fig. 16 η_k, λ_k/ω and $\eta_k + \lambda_k/\omega$ for $\omega = 0.1$ (*left*), $\omega = 1$ (*middle*), and $\omega = 10$ (*right*)

rate seems to be a little bit better than $N^{-0.4}$. This happens because λ_k is divided by 10 and its effect is not so visible in the picture. In the long run the decay cannot be better than $N^{-0.4}$.

It is also interesting to notice that λ_k does not decrease monotonically, mainly because the strongly curved part of γ is not aligned to the grid. This behavior is consistent with (19) and, in fact, shows that we cannot expect monotonicity of $\lambda_\Gamma(\mathcal{T})$ upon refinement, thereby justifying (19).

10 Conclusions and Comments

We finish the paper with the following remarks about this and related work.

- **Coupling PDE-Geometry:** This is a new feature in adaptivity and leads to separate handling of geometry and PDE resolution with specific relative tolerances. The current algorithm is different from that for graphs [46] studied by K. Mekchay, P. Morin, and R.H. Nochetto. The present paper studies polynomial degree 1, but the theory for parametric surfaces extends to higher polynomial degree [17].
- **Convergence rates:** We show optimal convergence rates in the energy norm

$$\left\| \nabla(u - U_k) \right\|_{L^2(\gamma)} \lesssim (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s}$$

  provided this is the rate of the best approximation of u in H^1 and that of γ in W^1_∞. This optimal result is consistent with that derived for flat domains by R. Stevenson [52] for the Poisson equation with data in H^{-1} and by J.M. Cascón et al. [21] for elliptic PDE with variable coefficients. None of them involve coarsening as the seminal paper [11] by P. Binev, W. Dahmen and R. DeVore. The present estimates extend those in [21] to the Laplace-Beltrami operator.
- **Weaker conditions on f:** We refer to A. Cohen, R. DeVore, and R.H. Nochetto [23] for convergence rates of elliptic PDE in flat domains with $f \in H^{-1}$ and A piecewise constant:

$$\operatorname{div}(A\nabla u) = f. \tag{91}$$

Paper [23] shows that approximability of u is sufficient for a complete theory. Whether this is true for the Laplace-Beltrami operator is still an open question.

- **Weaker conditions on γ:** We assume γ is W^2_p with $p > d$, which implies γ is C^1. In the flat case, this corresponds to piecewise continuous A. We refer to A. Bonito, R. DeVore, and R.H. Nochetto [18] for optimal convergence rates of AFEM for (91) with weaker regularity assumptions on A. This could be especially relevant to perform adaptivity on problems where the singularity location is not known beforehand, such as those in Sect. 2.

Acknowledgements The work of A. Bonito was partially supported by NSF Grant DMS-0914977.

The work of J.M. Cascón was partially supported by Secretaría de Estado de Investigación, Desarrollo e Innovación through grant: CGL2011-29396-C03-02 (Spain), and by Conserjería de Educación (Junta de Castilla y León), through grant: SA266A12-2.

The work of P. Morin was partially supported by CONICET through grant PIP 112-200801-02182, Universidad Nacional del Litoral through grants CAI+D 062-312, 062-309, and Agencia Nacional de Promoción Científica y Tecnológica, through grant PICT-2008-0622 (Argentina).

The work of R.H. Nochetto was partially supported by NSF Grant DMS-1109325, and the General Research Board of the University of Maryland.

References

1. Ainsworth, M., Oden, J.T.: A Posteriori Error Estimation in Finite Element Analysis. Pure and Applied Mathematics (New York). Wiley-Interscience, New York (2000), pp. xx+240
2. Bänsch, E.: Finite element discretization of the Navier-Stokes equations with a free capillary surface. Numer. Math. **88**(2), 203–235 (2001)
3. Bänsch, E., Haußer, F., Lakkis, O., Li, B., Voigt, A.: Finite element method for epitaxial growth with attachment-detachment kinetics. J. Comput. Phys. **194**(2), 409–434 (2004)
4. Bänsch, E., Morin, P., Nochetto, R.H.: Surface diffusion of graphs: variational formulation, error analysis, and simulation. SIAM J. Numer. Anal. **42**(2), 773–799 (2004)
5. Bänsch, E., Morin, P., Nochetto, R.H.: A finite element method for surface diffusion: the parametric case. J. Comput. Phys. **203**(1), 321–343 (2005)
6. Barrett, J.W., Garcke, H., Nürnberg, R.: A parametric finite element method for fourth order geometric evolution equations. J. Comput. Phys. **222**(1), 441–462 (2007)
7. Barrett, J.W., Garcke, H., Nürnberg, R.: Parametric approximation of Willmore flow and related geometric evolution equations. SIAM J. Sci. Comput. **31**(1), 225–253 (2008)
8. Barrett, J.W., Garcke, H., Nürnberg, R.: Finite-element approximation of coupled surface and grain boundary motion with applications to thermal grooving and sintering. Eur. J. Appl. Math. **21**(6), 519–556 (2010)
9. Bartels, S., Dolzmann, G., Nochetto, R.H.: A finite element scheme for the evolution of orientation order in fluid membranes. Modél. Math. Anal. Numér. **44**(1), 1–31 (2010)
10. Bartels, S., Dolzmann, G., Nochetto, R.H., Raisch, A.: Finite element methods for director fields on flexible surfaces. Interfaces Free Bound. **14**(2), 231–272 (2012). doi:10.4171/IFB/281
11. Binev, P., Dahmen, W., DeVore, R.A.: Adaptive finite element methods with convergence rates. Numer. Math. **97**(2), 219–268 (2004)
12. Bonito, A., Nochetto, R.H.: Quasi-optimal convergence rate of an adaptive discontinuous Galerkin method. SIAM J. Numer. Anal. **48**(2), 734–771 (2010)
13. Bonito, A., Pasciak, J.E.: Convergence analysis of variational and non-variational multigrid algorithm for the Laplace-Beltrami operator. Math. Comput. **81**, 1263–1288 (2012)

14. Bonito, A., Nochetto, R.H., Pauletti, M.S.: Geometrically consistent mesh modification. SIAM J. Numer. Anal. **48**(5), 1877–1899 (2010)
15. Bonito, A., Nochetto, R.H., Pauletti, M.S.: Parametric FEM for geometric biomembranes. J. Comput. Phys. **229**(9), 3171–3188 (2010)
16. Bonito, A., Nochetto, R.H., Pauletti, M.S.: Dynamics of biomembranes: effect of the bulk fluid. Math. Model. Nat. Phenom. **6**(5), 25–43 (2011)
17. Bonito, A., Cascón, J.M., Mekchay, K., Morin, P., Nochetto, R.H.: AFEM for the Laplace-Beltrami operator on parametric surfaces: convergence rates (in preparation)
18. Bonito, A., DeVore, R.A., Nochetto, R.H.: Adaptive finite element methods for elliptic problems with discontinuous coefficients (in preparation)
19. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Texts in Applied Mathematics, vol. 15, 2nd edn. Springer, New York (2002), pp. xvi+361
20. Cahn, J., Taylor, J.E.: Surface motion by surface diffusion. Acta Metall. Mater. **42**, 1045–1063 (1994)
21. Cascón, J.M., Kreuzer, C., Nochetto, R.H., Siebert, K.G.: Quasi-optimal convergence rate for an adaptive finite element method. SIAM J. Numer. Anal. **46**(5), 2524–2550 (2008)
22. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. Studies in Mathematics and its Applications, vol. 4. North-Holland, Amsterdam (1978), pp. xix+530
23. Cohen, A., DeVore, R., Nochetto, R.H.: Convergence rates for AFEM with H^{-1} data. Found. Comput. Math. **12**(5), 671–718 (2012). doi:10.1007/s10208-012-9120-1
24. Deckelnick, K., Dziuk, G.: Discrete anisotropic curvature flow of graphs. Modél. Math. Anal. Numér. **33**(6), 1203–1222 (1999)
25. Deckelnick, K., Dziuk, G.: Error estimates for a semi-implicit fully discrete finite element scheme for the mean curvature flow of graphs. Interfaces Free Bound. **2**(4), 341–359 (2000)
26. Deckelnick, K., Dziuk, G., Elliott, C.M.: Fully discrete finite element approximation for anisotropic surface diffusion of graphs. SIAM J. Numer. Anal. **43**(3), 1112–1138 (2005)
27. Deckelnick, K., Dziuk, G.: Error analysis of a finite element method for the Willmore flow of graphs. Interfaces Free Bound. **8**(1), 21–46 (2006)
28. Demlow, A.: Higher-order finite element methods and pointwise error estimates for elliptic problems on surfaces. SIAM J. Numer. Anal. **47**(2), 805–827 (2009)
29. Demlow, A., Dziuk, G.: An adaptive finite element method for the Laplace-Beltrami operator on implicitly defined surfaces. SIAM J. Numer. Anal. **45**(1), 421–442 (2007)
30. Dierkes, U., Hildebrandt, S., Küster, A., Wohlrab, O.: Minimal Surfaces. I. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 295. Springer, Berlin (1992), pp. xiv+508
31. do Carmo, M.P.: Differential Geometry of Curves and Surfaces. Prentice-Hall, Englewood Cliffs (1976), pp. viii+503
32. Doğan, G., Nochetto, R.H.: First variation of the general curvature-dependent surface energy. Modél. Math. Anal. Numér. **46**(1), 59–79 (2012)
33. Doğan, G., Morin, P., Nochetto, R.H., Verani, M.: Discrete gradient flows for shape optimization and applications. Comput. Methods Appl. Mech. Eng. **196**(37–40), 3898–3914 (2007)
34. Du, Q., Liu, C., Wang, X.: Simulating the deformation of vesicle membranes under elastic bending energy in three dimensions. J. Comput. Phys. **212**(2), 757–777 (2006)
35. Dziuk, G.: Finite elements for the Beltrami operator on arbitrary surfaces. In: Partial Differential Equations and Calculus of Variations. Lecture Notes in Math., vol. 1357, pp. 142–155 (1988)
36. Dziuk, G.: An algorithm for evolutionary surfaces. Numer. Math. **58**(6), 603–611 (1991)
37. Dziuk, G.: Computational parametric Willmore flow. Numer. Math. **111**(1), 55–80 (2008)
38. Elliott, C.M., Stinner, B.: Modeling and computation of two phase geometric biomembranes using surface finite elements. J. Comput. Phys. **229**(18), 6585–6612 (2010)
39. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 224, 2nd edn. Springer, Berlin (1983), pp. xiii+513

40. Han, Q., Lin, F.: Elliptic Partial Differential Equations. Courant Lecture Notes in Mathematics, vol. 1, 2nd edn. Courant Institute of Mathematical Sciences, New York (2011), pp. x+147
41. Helfrich, W.: Elastic properties of lipid bilayers—theory and possible experiments. Z. Nat.forsch., C J. Biosci. **28**, 693 (1973)
42. Khairy, K., Foo, J., Howard, J.: Shapes of red blood cells: comparison of 3D confocal images with the bilayer-couple model. Cell. Mol. Bioeng. **1**(2), 173–181 (2008)
43. Kornhuber, R., Yserentant, H.: Multigrid methods for discrete elliptic problems on triangular surfaces. Comput. Vis. Sci. **11**(4–6), 251–257 (2008)
44. Lakkis, O., Nochetto, R.H.: A posteriori error analysis for the mean curvature flow of graphs. SIAM J. Numer. Anal. **42**(5), 1875–1898 (2005)
45. Laradji, M., Mouritsen, O.G.: Elastic properties of surfactant monolayers at liquid-liquid interfaces: a molecular dynamics study. J. Chem. Phys. **112**(19), 8621–8630 (2000)
46. Mekchay, K., Morin, P., Nochetto, R.H.: AFEM for the Laplace-Beltrami operator on graphs: design and conditional contraction property. Math. Comput. **80**(274), 625–648 (2011)
47. Morin, P., Nochetto, R.H., Siebert, K.G.: Data oscillation and convergence of adaptive FEM. SIAM J. Numer. Anal. **38**(2), 466–488 (2000)
48. Morin, P., Nochetto, R.H., Siebert, K.G.: Convergence of adaptive finite element methods. SIAM Rev. **44**(4), 631–658 (2002)
49. Nochetto, R.H., Siebert, K.G., Veeser, A.: Theory of adaptive finite element methods: an introduction. In: Multiscale, Nonlinear and Adaptive Approximation, pp. 409–542. Springer, Berlin (2009)
50. Rusu, R.E.: An algorithm for the elastic flow of surfaces. Interfaces Free Bound. **7**(3), 229–239 (2005)
51. Sokołowski, J., Zolésio, J.-P.: Introduction to Shape Optimization. Springer Series in Computational Mathematics, vol. 16. Springer, Berlin (1992)
52. Stevenson, R.: Optimality of a standard adaptive finite element method. Found. Comput. Math. **7**(2), 245–269 (2007)
53. Stevenson, R.: The completion of locally refined simplicial partitions created by bisection. Math. Comput. **77**(261), 227–241 (2008)
54. Taylor, J.E.: Some mathematical challenges in materials science. Bull. Am. Math. Soc. (N.S.) **40**(1), 69–87 (2003)
55. Verfürth, R.: A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Technique. Wiley-Teubner, Chichester (1996)
56. Willmore, T.J.: Riemannian Geometry. Oxford Science Publications. The Clarendon Press/Oxford University Press, New York (1993), pp. xii+318

Generalized Reduced Basis Methods and n-Width Estimates for the Approximation of the Solution Manifold of Parametric PDEs

Toni Lassila, Andrea Manzoni, Alfio Quarteroni, and Gianluigi Rozza

Abstract The set of solutions of a parameter-dependent linear partial differential equation with smooth coefficients typically forms a compact manifold in a Hilbert space. In this paper we review the generalized reduced basis method as a fast computational tool for the uniform approximation of the solution manifold.

We focus on operators showing an *affine parametric dependence*, expressed as a linear combination of parameter-independent operators through some smooth, parameter-dependent scalar functions. In the case that the parameter-dependent operator has a dominant term in its affine expansion, one can prove the existence of exponentially convergent uniform approximation spaces for the entire solution manifold. These spaces can be constructed without any assumptions on the parametric regularity of the manifold—only spatial regularity of the solutions is required. The exponential convergence rate is then inherited by the generalized reduced basis method. We provide a numerical example related to parametrized elliptic equations confirming the predicted convergence rates.

T. Lassila · A. Quarteroni · G. Rozza (✉)
MATHICSE-CMCS, Modelling and Scientific Computing, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
e-mail: gianluigi.rozza@epfl.ch

T. Lassila
e-mail: toni.lassila@epfl.ch

A. Quarteroni
e-mail: alfio.quarteroni@epfl.ch

A. Manzoni · G. Rozza
SISSA Mathlab, SISSA—International School for Advanced Studies, Trieste, Italy

A. Manzoni
e-mail: amanzoni@sissa.it

G. Rozza
e-mail: gianluigi.rozza@sissa.it

A. Quarteroni
MOX, Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*, Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_16,

1 Introduction

Reduced order models (ROMs) are a crucial ingredient of many applications of increasing complexity in scientific computing related e.g. to parameter estimation, sensitivity analysis, optimal control, and design or shape optimization. In this paper we consider the reduced basis method for the numerical approximation of parameter-dependent partial differential equations (μ-PDEs). The set of solutions of such an equation depends on a finite-dimensional vector of parameters related e.g. to physical coefficients, geometrical configuration, source terms, and boundary conditions. Solving the μ-PDE for many different values of the parameters entails the exploration of the manifold of solutions, and is not affordable if each μ-PDE requires an expensive numerical approximation, such as the one built over the finite element method. Suitable structural assumptions about the parametrization enable one to decouple the computational effort into two stages. A (very expensive) preprocessing step that is performed once ("offline")—consisting in the construction of a reduced basis for the representation of the manifold of solutions, followed by very inexpensive calculations performed "online" for each new input-output evaluation required. In the reduced basis method, numerical solutions for certain parameters values are computed *offline* by a classical discretization technique. These solutions give a basis for approximating *online* the PDE solution (for a large number of new parameter values) as a linear combination of the basis elements. The rational of this approach stands on the very fast (often exponential) convergence—with respect to the number of basis—is exhibited by approximation spaces. We point out that in the *real-time* or *many-query* contexts, where the goal is to achieve a very low *marginal cost* per input-output evaluation, we can accept an increased "offline" cost—not tolerable for a single or few evaluations—in exchange for greatly decreased "online" cost for each new/additional input-output evaluation. In previous works (see e.g. [18–20]) *a priori* exponential convergence with respect to the number of basis functions is proved in the case of elliptic PDEs depending on one-dimensional parameters; several computational tests shown e.g. in [28] provide a numerical assessment of this behavior, also for larger parameter space dimensions. Several new results, such as the ones presented in [4], address an *a priori* convergence analysis in the more general case where a *greedy* algorithm is employed to build the reduced space in an automatic, adaptive way. A further improvement has been proposed in [3], where an error estimate for the greedy algorithm has been developed in terms of the Kolmogorov n-width. After recalling the basic features of a generalized version of the reduced basis method, and the main convergence results in this field, the goal of this paper is to provide both a convergence analysis for the greedy algorithm and a numerical proof of this behavior, in order to extend the *a priori* convergence results presented in [19, 20]. To do this, we rely on the introduction of a *fundamental basis*, a suitable error representation formula and an upper bound estimate for the n-width of the solution set of an elliptic parametric PDE under suitable assumptions on its parametric form.

We proceed to describe the functional setting of our problems. Let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, be a bounded domain and $X = X(\Omega)$ a Hilbert space of functions

defined on Ω with inner product $(\cdot,\cdot)_X$ and induced norm $\|\cdot\|_X = \sqrt{(\cdot,\cdot)_X}$. We consider the following problem: given a vector of parameters $\mu \in \mathcal{D}$ from a compact parameter set $\mathcal{D} \subset \mathbb{R}^P$, find $u(\mu) \in X$ s.t.

$$a\big(u(\mu), v; \mu\big) = f(v; \mu) \quad \text{for all } v \in X, \tag{1}$$

where the parameters can enter in the bilinear form $a(\cdot,\cdot;\mu)$ in several possible ways: as variable coefficients, as coefficient entering in the parametrization of the domain $\Omega \subset \mathbb{R}^d$ of the problem, in the definition of the right-hand side f that account for either forcing terms and/or boundary conditions. We denote the P-manifold of solutions

$$\mathcal{K} := \big\{u(\mu) \in X : \mu \in \mathcal{D} \subset \mathbb{R}^P\big\}$$

in the space X. In many applications $\mathcal{K}$ is a differentiable manifold. We also allow the case of a manifold that is not locally smooth at some isolated points, e.g. the parametric Helmholtz equation $\nabla \cdot (a(\cdot;\mu)\nabla u) + u = 0$, which has a smooth solution manifold except at the eigenvalues of the parametric Laplacian, $-\nabla \cdot (a(\cdot;\mu)\nabla u) = \lambda(\mu)u$. We considered this problem in [17]. A typical objective in applications is to provide a numerical approximation $\widetilde{u}(\mu)$ for $u(\mu) \in \mathcal{K}$ that is *uniform*[1] over the entire manifold $\mathcal{K}$. To fulfill this request, a necessary condition is that, for any $\varepsilon > 0$, we find a linear subspace $X_N \subset X$ of dimension N s.t.

$$\inf_{\widetilde{u} \in X_N} \big\|u(\mu) - \widetilde{u}\big\|_X < \varepsilon \quad \text{for all } \mu \in \mathcal{D},$$

where the dimension N is as small as possible. The question is: how small can we expect N to be?

To address this question, we introduce e.g. the finite element (FE) approximation of problem (1): given a vector of parameters $\mu \in \mathcal{D} \subset \mathbb{R}^P$, find $u_{h,p}(\mu) \in X_{h,p} = X_{h,p}(\Omega)$ s.t.

$$a(u_{h,p}, v_{h,p}; \mu) = f(v_{h,p}; \mu) \quad \text{for all } v_{h,p} \in X_{h,p}, \tag{2}$$

where $X_{h,p} \subset X$ is a conforming FE subspace spanned by piecewise polynomial shape functions of degree p defined on a quasi-uniform mesh of maximum element size h. Due to classical a priori error estimates such an approximation will eventually approximate well all the solutions on the manifold as the dimension $N := \dim(X_{h,p})$ increases, but only for quite large N we can expect a uniformly small error in the approximation. When $X = H^1(\Omega)$ is the standard Sobolev space, the classical a priori estimates for piecewise polynomial approximations are as follows [2]:

[1] The other option is to consider *local* or *sequential* approximations of the manifold, such as tracking a path on the manifold starting from a certain point and proceeding via a continuation method. In such cases we are usually not interested in the global behavior of the manifold.

$$\left\|u(\mu)-u_{h,p}(\mu)\right\|_1 \leq \begin{cases} C(\mu)N^{-\min\{s-1,p\}/d} & \text{with } h\text{-refinement,} \\ C(\mu)N^{-(s-1-\delta)/d} & \text{with } p\text{-refinement,} \end{cases} \tag{3}$$

where $s > 1$ denotes the number of weak derivatives of $u(\mu)$, and $\delta > 0$ is arbitrarily small. If the solution u is analytic ($s = \infty$), one obtains exponential convergence as a result of p-refinement, i.e.

$$\left\|u(\mu)-u_{h,\text{spectral}}(\mu)\right\|_1 \leq C(\mu)\exp(-\gamma N),$$

and this leads to the study of spectral methods. It should be cautioned that even if a spectral approximation can obtain in theory exponential convergence across the entire parameter range, the constants in front depend on both the dimension d and the number of parameters P of the problem. The assumption of analyticity of solutions is also often violated.

An efficient method for the approximation of the parametric manifold $\mathcal{K}$ should (i) provide exponential convergence in the dimension N of the approximation space; (ii) have the same convergence rate irrespective of the number of parameters P; and (iii) entail a computational cost that scales only moderately in N. Exploiting the *structure of the manifold* $\mathcal{K}$ is key to finding uniform approximations that satisfy (i)–(iii). Our technique for proving exponentially convergent approximation estimates for the manifold of solutions relies on a series expansion of the solution $u(\mu)$. Series expansion solutions, either by separation of variables or by power series expansion for PDEs with analytical coefficients, are classical tools for existence proofs. Analytical power series expansions, such as the decomposition method of Adomian [1], are not competitive against good numerical approximation schemes in actually providing approximate solutions to PDEs, but they do provide an interesting approach to constructing convergence estimates. The novel contribution of this work is to consider the power series expansion method forparameter-dependent PDEs by searching for solutions in a parametrically separable form

$$u(\mu) = \sum_{k=0}^{\infty} \Theta_k(\mu)\Psi_k, \tag{4}$$

where the Ψ_k do not depend on μ and the scalar functions $\Theta_k(\mu)$. The expansion (4) together with standard estimates for convergent power series then provides a construction of approximation spaces that are uniformly exponentially convergent over the entire parameter range. In order to achieve separation w.r.t. to the parameters, we must make suitable structural assumptions on the PDE. A typical assumption is that of affine dependence on the parameter, i.e. problem (1) is assumed to be of the form

$$\sum_{q=1}^{Q_a} \Theta_q^a(\mu)a_q(u,v) = \sum_{q=1}^{Q_f} \Theta_q^f(\mu)f_q(v) \quad \text{for all } v \in X, \tag{5}$$

where every $a_q : X \times X \to \mathbb{R}$ and $f_q : X \to \mathbb{R}$ are parameter-independent bilinear and linear forms respectively, whereas $\Theta_q^a : \mathcal{D} \to \mathbb{R}$ and $\Theta_q^f : \mathcal{D} \to \mathbb{R}$ are scalar coefficient functions depending only on the parameter (but not necessary in a smooth

way). We shall next describe the generalized reduced basis method (GRBM) that, given assumption (5), satisfies (i)–(iii). We then discuss some recent theoretical approximation results linking the best possible approximation space for $\mathcal{K}$ with the convergence rate obtained by the GRBM, and exhibit a model problem where, at least in the case we have $Q_a = 2$ and some additional special structure on bilinear forms a_q, we indeed observe in practice the exponential convergence predicted by theory.

2 Generalized Reduced Basis Method for Uniform Approximation of μ-PDEs

It is clear that the FE approximation $u_{h,p}(\mu)$ of (1) can be made arbitrarily accurate for all possible parameters $\mu \in \mathcal{D}$, but this usually require a considerable computational cost. In order to overcome this (sometimes unaffordable) difficulty, a possible idea is to instead consider the manifold of discrete solutions $u_{h,p}(\mu)$ given by

$$\mathcal{K}_{h,p} := \big\{u_{h,p}(\mu) \in X_{h,p} : \mu \in \mathcal{D} \subset \mathbb{R}^P\big\},$$

as a surrogate for $\mathcal{K}$, and then to look for approximations of $\mathcal{K}_{h,p}$ that converge exponentially fast (see Fig. 1 for a graphical sketch). Specifically, we consider subspaces $X_n \subset X_{h,p}$ (for $n \ll N$) that are constructed by using information coming from the snapshot solutions $u_{h,p}(\mu_i)$ computed at well-chosen points μ_i, $i = 1, \ldots, n$. More precisely, X_n is the *span* of the snapshot solutions $u_{h,p}(\mu_i)$, $i = 1, \ldots, n$. This leads to the Reduced Basis (RB) method [27, 28], which is, in brief, a Galerkin projection on an n-dimensional approximation space relying on the parametrically induced manifold $\mathcal{K}_{h,p}$. Assuming that the solutions $u_{h,p}$ satisfy the a priori convergence estimate (3) and that the approximation $u_n(\mu)$ converges exponentially to $u_{h,p}(\mu)$, we can write the total error of the reduced solution as

$$\begin{aligned}\big\|u(\mu) - u_n(\mu)\big\|_X &\le \big\|u(\mu) - u_{h,p}(\mu)\big\|_X + \big\|u_{h,p}(\mu) - u_n(\mu)\big\|_X \\ &\le C(\mu)\big[N^{-\min\{s-1,p\}/d} + \exp(-\gamma n)\big],\end{aligned}$$

and for N sufficiently large the exponential term in n dominates the error. The additional strength of this method is that in the best case we only need to solve n times the FE problem for $u_{h,p}$, and that the solution for $u_n(\mu)$ can be done with complexity depending on n but not N after some initial preprocessing steps, so that indeed N can be chosen fairly large in order to obtain highly accurate reduced order solutions.

For notational simplicity we present here the case $P = 1$ and formulate the RB method in a slightly more general form than usually given. It can be expressed in three distinct steps:

(i) **Choice of the reduced subspace**. The basic idea of every reduced basis method is to choose a finite training sample $\Xi_{\text{train}} \subset \mathcal{D}$, $|\Xi_{\text{train}}| = M$, in the

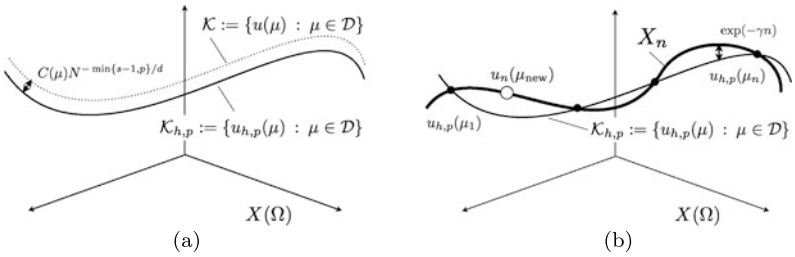


Fig. 1 (**a**) Low-dimensional manifold $\mathcal{K}_{h,p}$ on which the field variable resides and (**b**) approximation of a new solution at μ_{new} with the "snapshots" $u_{h,p}(\mu_m)$, $1 \le m \le n$

parameter space and to use the information contained in the corresponding solutions $u_{h,p}(\mu)$ for each $\mu \in \Xi_{\text{train}}$ (called snapshots) to find a representative subspace for the approximation of the manifold $\mathcal{K}_{h,p}$. The reduced subspace X_n of dimension n is found by solving [28]

$$X_n := \underset{X^* \subset X_{\text{train}}, \dim(X^*)=n}{\arg\inf} \delta\big(X^*, \mathcal{K}_{h,p}; X\big) \tag{6}$$

where $X_{\text{train}} := \text{span}\{u_{h,p}(\mu) : \mu \in \Xi_{\text{train}}\}$ is the space containing all the snapshots, and the function $X^* \mapsto \delta(X^*, \mathcal{K}_{h,p}; X) \in \mathbb{R}$ measures the distance between any subspace $X^* \subset X$ and the manifold $\mathcal{K}_{h,p}$ and is defined by

$$\delta(X_n, \mathcal{K}_{h,p}; X) := \sup_{u \in \mathcal{K}_{h,p}} \inf_{\widetilde{u} \in X_n} \|u - \widetilde{u}\|_X.$$

Since the exact distance of the subspace to the manifold is usually unknown, we must resort to computable surrogates to solve (6). Nonetheless, we mention that the exact distance is used in the so-called *strong greedy* algorithm introduced in [3] for the theoretical analysis of convergence rates of reduced basis methods. Thus we replace (6) with

$$X_n := \underset{X^* \subset X_{\text{train}}, \dim(X^*)=n}{\arg\inf} M_{\mathcal{K}_{h,p}}\big(X^*\big), \tag{7}$$

where $X^* \mapsto M_{\mathcal{K}}(X^*) \in \mathbb{R}$ is an *approximate distance* between any subspace $X^* \subset X$ and the manifold $\mathcal{K}_{h,p}$. The choice of the function $M_{\mathcal{K}}$ to be used for the approximation of $\delta(\cdot,\cdot;\cdot)$ defines which algorithm we use to choose the subspace. This is by far the most common way of constructing reduced subspaces, and we call these approaches *Lagrange ROM*s. In the GRBM we consider also the parametric sensitivities up to a suitable order, say $K-1$ (with $K \ge 1$) as part of the snapshot set

$$X_{\text{train}} := \text{span}\left\{ \frac{\partial^k u_{h,p}}{\partial \mu^k}(\mu) : \mu \in \Xi_{\text{train}},\ k = 0, \ldots, K-1 \right\},$$

giving a total number of MK snapshots. They can be computed from the *discrete sensitivity equation(s)*: find $w_k(\mu) = \frac{\partial^k u_{h,p}}{\partial \mu^k}(\mu) \in X_{h,p}$ s.t.

$$a\big(w_k(\mu), v_{h,p}; \mu\big) = \frac{\partial^k f}{\partial \mu^k}(v_{h,p}; \mu) - \sum_{\ell=1}^{k} \binom{k}{k-\ell} \frac{\partial^\ell a}{\partial \mu^\ell}\big(w_{k-l}(\mu), v_{h,p}\big)$$
$$\text{for all } v_{h,p} \in X_{h,p}, \tag{8}$$

for all $k = 1, \ldots, K$. Only the right-hand side of the system changes with k and thus any preconditioners or matrix decompositions used for the primal problem can be reused. The information contained in these snapshots is then used to build the reduced space X_n with dimension $\dim X_n = n \ll MK$ in what can also be understood as a *data compression problem*. If $K = 2$ and $M > 1$ we have a *Hermite ROM*,[2] and if $K > 1$ and $M = 1$ we have a *Taylor ROM*.[3]

Two standard choices for $M_{\mathcal{K}}$ are:

1. the *proper orthogonal decomposition* (POD), where

$$M_{\mathcal{K}}(X_n) = \frac{1}{M} \sum_{i=1}^{M} \big\| u_{h,p}(\mu_i) - \Pi_{X_n}\big(u_{h,p}(\mu_i)\big) \big\|_X^2, \tag{9}$$

and $\Pi_{X_n} : X \to X_n$ is the orthogonal projection w.r.t. the inner product of X. In this case, we choose the basis by minimizing the $\ell^2(\Xi_{\text{train}})$ error in parameter space. It turns out that the optimal bases are hierarchical and are spanned by the leading n eigenvectors of the correlation matrix

$$\mathbb{C}_{ij} = \frac{1}{M}\big(u_{h,p}(\mu_j) - \bar{u}, u_{h,p}(\mu_i) - \bar{u}\big)_X, \quad 1 \le i, j \le M,$$

where we have subtracted the mean of the snapshots

$$\bar{u} = \frac{1}{M} \sum_{i=1}^{M} u_{h,p}(\bar{\mu}).$$

[2] Ito and Ravindran [16] were perhaps the first ones to suggest using a Hermite ROM in a uniform approximation context, rather than in a pure continuation method. The Lagrange and Hermite ROMs were compared on a driven cavity problem, where the Hermite approach was somewhat superior. No stability problems were reported and the Hermite basis with only two basis functions was able to extrapolate solutions to much larger Reynolds numbers.

[3] In one of the pioneering works on RBM, Noor [24] used a Taylor ROM to build a local reduced space that was used to trace the post-buckling behavior of a nonlinear structure. The continuation idea was used also by Peterson [25] to compute Navier-Stokes solutions with increasing Reynolds number flow over a forward facing step. Again a Taylor ROM was constructed and used to extrapolate an initial guess for the Newton method at a slightly higher Reynolds number.

The eigenpairs $(\lambda_j, \boldsymbol{\psi}_j)_{j=1}^{M}$ of $\mathbb{C}$ (with the eigenvalues ordered in the decreasing order) are the solutions of

$$\mathbb{C}\boldsymbol{\psi}_j = \lambda_j \boldsymbol{\psi}_j, \quad j = 1, \ldots, |\Xi_{\text{train}}|.$$

Then, the optimal basis for the n-th dimensional space X_n generated by minimizing (9) is given by

$$\chi_0 = \bar{u}, \qquad \chi_j = \sum_{i=1}^{M} [\boldsymbol{\psi}_j]_i \big(u_{h,p}(\mu_i) - \bar{u}\big), \quad 1 \le j \le n,$$

being $[\boldsymbol{\psi}_j]_i$ the i-th component of the j-th eigenvector. Extensions of the standard POD basis to incorporate parametric sensitivities (8) were presented in [6, 12, 13] and are not discussed in detail here;[4,5]

2. the *greedy algorithm*, where

$$M_{\mathcal{K}}(X_n) = \sup_{\mu \in \Xi_{\text{train}}} \big\| u_{h,p}(\mu_i) - \Pi_{X_n}\big(u_{h,p}(\mu_i)\big) \big\|_X \tag{10}$$

i.e. minimization of the $\ell^\infty(\Xi_{\text{train}})$ error in parameter space. In practice no efficient algorithm exists to solve (6) for large-scale problems, so we approximate it by its relaxation

$$M_{\mathcal{K}}(X_n) = \sup_{\mu \in \Xi_{\text{train}}} \Delta_n\big(\widetilde{u}(\mu)\big) \tag{11}$$

where $\Delta_n(\widetilde{u}(\mu))$ is a computationally inexpensive a posteriori error estimator for the quantity $\|u_{h,p}(\mu) - \widetilde{u}(\mu)\|_X$ that should satisfy

$$C_1 \Delta_n\big(\widetilde{u}(\mu)\big) \le \big\| u(\mu) - \widetilde{u}(\mu) \big\|_X \le C_2 \Delta_n\big(\widetilde{u}(\mu)\big), \quad \forall \mu \in \mathcal{D} \tag{12}$$

for some constants $C_1 > 0$, $C_2 \ge 1$. This corresponds to the approximate minimization of the $\ell^\infty(\Xi_{\text{train}})$ error in parameter space. In this case we have a *weak greedy algorithm* as defined in [3].

[4] In the works of Hay et al. [12, 13] sensitivity information was introduced into the proper orthogonal decomposition framework. The parametric sensitivities of the POD modes were derived and computed. The test problems were related with channel flow around a cylindrical obstacle, either by using a simple parametrization as the Reynolds number, or a more involved geometric parametrization of the obstacle. The use of a Hermite ROM considerably improved the validity of the reduced solutions away from the parametric snapshots. However, in the more involved geometrical parametrization case the Hermite ROM failed completely, as it did not converge to the exact solution even when the number of POD modes was increased.

[5] Carlberg and Farhat [6] proposed an approach they call "compact POD", based on goal-oriented Petrov-Galerkin projection to minimize the approximation error subject to a chosen output criteria, and including sensitivity information with proper weighting coming from the Taylor-expansion and including "mollification" of basis functions far away from the snapshot parameter. The application was the optimization of an aeroelastic wing configuration by building local ROMs along the path to the optimal wing configuration.

Note that while conceptually the POD and the greedy algorithms can be cast in a similar framework, their practical implementations are quite different. The training set $\Xi_{\text{train}} \subset \mathcal{D}$ needs to be reasonably dense in the parameter space for $\mathcal{M}_{\mathcal{K}_{h,p}}(X^*)$ to be a good approximation of the true distance $\delta(X^*, \mathcal{K}_{h,p}; X)$ for all subspaces $X^* \subset X$. In the POD one needs to compute the FE approximations $u_{h,p}(\mu)$ for all the points in Ξ_{train}, which amounts to a considerable computational undertaking. In contrast, the weak greedy algorithm only needs to compute the exact solutions (and their parametric derivatives) at the n snapshots comprising the RB and only the computationally inexpensive a posteriori estimator $\Delta_n(u_n(\mu))$ needs to be evaluated over the entire training set. The difference in the norms used (ℓ^∞ for the greedy vs. ℓ^2 for the POD) also results in slightly different approximation behavior of the resulting bases. Typically the POD basis needed to reach a given tolerance is smaller in size but tends to be not as robust far away from the snapshots (see e.g. [28, 29]).

(ii) **(Petrov-)Galerkin projection of the equations**. In the second step we perform projection of the original problem onto the reduced trial subspace X_n using the reduced test subspace Y_n to obtain the reduced basis approximation: find $u_n(\mu) \in X_n$ s.t.

$$a\big(u_n(\mu), v_n; \mu\big) = f(v_n; \mu) \quad \text{for all } v_n \in Y_n, \tag{13}$$

where $X_n = \text{span}\{\varphi_j\}_{j=1}^n$ and $Y_n = \text{span}\{\psi_j\}_{j=1}^n$. If $Y_n = X_n$ this is a pure Galerkin method, otherwise it is a Petrov-Galerkin method. The Petrov-Galerkin approach is adopted if the underlying system is either nonsymmetric or noncoercive and can be interpreted as a form of stabilization of the ROM. Applying the assumption (5) to (13) leads to the discrete system

$$\sum_{q=1}^{Q_a} \Theta_q^a(\mu) A_q^n U^n = \sum_{q=1}^{Q_f} \Theta_q^f(\mu) F_q^n \tag{14}$$

where the matrices and vectors

$$\underset{n\times n}{\big[A_q^n\big]_{i,j}} := a_q(\varphi_j, \psi_i), \quad \underset{n\times 1}{\big[F_q^n\big]_i} := F_q(\psi_i), \quad \underset{n\times 1}{U^n} \text{ the reduced solution}, \tag{15}$$

are dense but only of dimension n and more importantly can be assembled once and then stored. The system (14) is then assembled by evaluating the coefficient functions Θ_q^a, Θ_q^f and summing together the weighted contributions from all the parts of the decomposition, and solving one small dense linear system. Assuming all the FE degrees of freedom are nodal, we can write the discrete

projectors

$$[\underset{N\times n}{\mathbb{X}_n}]_{i,j} := \varphi_j(x_i), \qquad [\underset{N\times n}{\mathbb{Y}_n}]_{i,j} := \psi_j(x_i) \tag{16}$$

where x_i are the nodal points in the full space X_N. The discrete matrices and vectors (15) can then be obtained as

$$A_q^n = \mathbb{Y}_n A_q \mathbb{X}_n^T, \qquad F_q^n = \mathbb{Y}_n F_q,$$

and the approximation of the solution $u_N(\mu)$ is obtained as $u_n(x_i) = [\mathbb{X}_n^T U_n]_i$ for $i = 1, \ldots, N$. From here on we use mainly the discrete forms of the equations.

(iii) **Certification of the ROM with error bounds**. *A posteriori* error bounds are used to both (i) certify the GRBM solution during the online stage, and (ii) construct the reduced space by means of the *weak greedy* algorithm. For the sake of simplicity we treat the case of linear, elliptic and coercive μ-PDEs—extensions to noncoercive and nonlinear problems can be found in [5, 8, 11, 21, 30]. Our error bounds rely on two basic ingredients: the dual norm of the residual and a lower bound of the stability factor (in this case, of the parameter-dependent coercivity constant). The residual $r(v;\mu) \in X'_{h,p}$ is defined as

$$r(v;\mu) \equiv f(v;\mu) - a\big(u_n(\mu), v; \mu\big), \quad \forall v \in X_{h,p} \tag{17}$$

so that exploiting (2) and the bilinearity of $a(\cdot,\cdot;\mu)$ we have the error representation for $e(\mu) = u_{h,p}(\mu) - u_n(\mu) \in X_{h,p}$ given by

$$a\big(e(\mu), v; \mu\big) = r(v;\mu), \quad \forall v \in X_{h,p}. \tag{18}$$

As a second ingredient, we need a positive lower bound $\alpha_h^{LB}(\mu)$ for the (discrete) coercivity constant $\alpha_h(\mu)$:

$$0 \le \alpha_h^{LB}(\mu) \le \alpha_h(\mu) := \inf_{w\in X_{h,p}} \frac{a(w,w;\mu)}{\|w\|^2_{X_{h,p}}}, \quad \forall \mu \in \mathcal{D}, \tag{19}$$

whose efficient evaluation as a function of μ is made possible thanks to the so-called successive constraint method (see e.g. [14, 15, 17] for a general description of this procedure). By combining (18) with 19 and using the Cauchy-Schwarz inequality, the following *a posteriori* error estimate in the energy norm holds (see [28] for a proof):

$$\left(\frac{\gamma_a(\mu)}{\alpha_h^{LB}(\mu)}\right)^{-1/2} \frac{\|r(u_n(\mu);\mu)\|_{X'_{h,p}}}{\sqrt{\alpha_h^{LB}(\mu)}} \le \big\|u_n(\mu) - u_{h,p}(\mu)\big\|_X \le \frac{\|r(u_n(\mu);\mu)\|_{X'_{h,p}}}{\sqrt{\alpha_h^{LB}(\mu)}}, \tag{20}$$

so that expression (12) is now made explicit, being

$$\Delta_n(\mu) := \frac{\|r(u_n(\mu);\mu)\|_{X'_{h,p}}}{\sqrt{\alpha_h^{LB}(\mu)}}, \qquad C_1 := \inf_{\mu\in\mathcal{D}} \left\{ \left(\frac{\gamma_a(\mu)}{\alpha_h^{LB}(\mu)} \right)^{-1/2} \right\},$$

$$C_2 := 1.$$

3 Approximation Theoretical Basis for the Generalized Reduced Basis Method

We now turn to the convergence analysis of approximations in the reduced subspaces that are obtained by (7) and the choice (11) for $M_{\mathcal{K}_{h,p}}$. We recall some recent theoretical results and provide an extension through an exponential convergence result. To do this, we rely on the introduction of a *fundamental* basis and on an intuitive error representation formula, which will be exploited in the numerical example discussed in the following section. We define the best approximation error of $\mathcal{K}_{h,p}$ obtained by the greedy algorithm (6) as

$$\sigma_n(\mathcal{K}_{h,p}; X) := \sup_{u_{h,p}\in\mathcal{K}_{h,p}} \inf_{\widetilde{u}\in X_n} \left\| u_{h,p} - \Pi_{X_n}(u_{h,p}) \right\|_X.$$

A priori convergence estimates for reduced basis approximations have been demonstrated in simple cases, such as in [19], where it was found that for a specific problem exponential convergence was achieved

$$\sigma_n(\mathcal{K}_{h,p}; X) \le C \exp\left(-n^\alpha\right), \quad \text{for some } \alpha > 0.$$

Recently much interest has been devoted to understanding why the weak greedy method (11) is able to give an approximation space X_n that exhibits exponential convergence in n. To express how well we are able to uniformly approximate a given manifold of solutions $\mathcal{K}_{h,p}$ with a finite-dimensional subspace, we recall the notion of n-width [22, 26] that is used to measure the degree in which we can uniformly approximate a subset of the space X using finite-dimensional subspaces X_n. The Kolmogorov n-width is defined as

$$d_n(\mathcal{K}_{h,p}; X) := \inf_{X_n\subset X} \sup_{u_{h,p}\in\mathcal{K}_{h,p}} \inf_{\widetilde{u}\in X_n} \|u_{h,p} - \widetilde{u}\|_X \tag{21}$$

where the first supremum is taken over all linear subspaces $X_n \subset X$ of dimension n. We also define the discrepancy between the subspace X_n and the manifold $\mathcal{K}_{h,p}$ as

$$\delta(X_n, \mathcal{K}_{h,p}; X) = \sup_{u_{h,p}\in\mathcal{K}_{h,p}} \inf_{\widetilde{u}\in X_n} \|u_{h,p} - \widetilde{u}\|_X.$$

The subspace X_n is said to be optimal if $\delta(X_n, \mathcal{K}_{h,p}; X) = d_n(\mathcal{K}_{h,p}; X)$. In general, the optimal subspace w.r.t. the Kolmogorov n-width (21) is not spanned by elements

of the set $\mathcal{K}_{h,p}$ being approximated, so that possibly $d_n(\mathcal{K}_{h,p}; X) \ll \sigma_n(\mathcal{K}_{h,p}; X)$. In the recent work [3] it was shown that

$$\sigma_n(\mathcal{K}_{h,p}; X) \leq \frac{2^{n+1}}{\sqrt{3}} d_n(\mathcal{K}_{h,p}; X),$$

and that this estimate cannot in general be improved. However, it was also shown that if the n-width converges at an exponential rate, say $d_n(\mathcal{K}_{h,p}; X) \leq C \exp(-cn^{\beta})$ for all $n > 0$ and some $\widetilde{C}, c > 0$, then

$$\sigma_n(\mathcal{K}_{h,p}; X) \leq \widetilde{C} \exp\bigl(-\alpha n^{\beta/(\beta+1)}\bigr). \tag{22}$$

A tighter estimate was proved for the case of algebraic convergence: if $d_n(\mathcal{K}_{h,p}; X) \leq CMn^{-\alpha}$ for all $n > 0$ and some $M, \alpha > 0$, then also

$$\sigma_n(\mathcal{K}_{h,p}; X) \leq CMn^{-\alpha}. \tag{23}$$

The fast (exponential) convergence of numerical approximations is often linked to spectral approximations. In this way, the reduced basis method can be understood as a spectral method, where instead of using generic global polynomial basis functions we use problem-dependent global smooth approximation basis. The analyticity of the solutions of elliptic PDEs was exploited e.g. in a recent work [7] in the special case where $\mathcal{K}_{h,p}$ is an analytic manifold. Using complex analysis techniques and a Taylor expansion approximation of the solution $u_{h,p}$ and its parametric derivatives $w_{m,k}$ for $k = 1, 2, \ldots$, the authors obtained a convergence rate for a reduced basis approximation as

$$\bigl\| u_{h,p}(\mu) - \widetilde{u}(\mu) \bigr\|_X \leq C(\mu) n^{-(1/p-1)},$$

where $0 < p < 1$ is the ℓ^p-summability exponent of a sequence related only to the diffusion coefficients of the problem. In particular, the convergence rate was independent of the spatial dimension d and the number of parameters P. In general, the reduced basis approximation of solutions of elliptic equations with regular coefficients has indeed been very successful. We would however like to convince the reader that analytic regularity of the solution manifold $\mathcal{K}_{h,p}$ is not necessary in order to successfully apply the reduced basis method.

Unfortunately, very little seems to be known about the n-width of manifolds of solutions of μ-PDEs. Very specific results concern special subspaces [10, 23]. For instance, if $Y \subset X$ is a dense, compactly embedded, and bounded subspace with inner product $(\cdot,\cdot)_Y$ then the n-width of a ball $B_Y \subset Y \subset X$ of finite radius is

$$d_n(B_Y; X) = \sqrt{\lambda_{n+1}},$$

where λ_n are the eigenvalues of the problem

$$\begin{cases} \lambda \in \mathbb{R}, \quad u \in V,\ u \neq 0, \\ (u, v)_Y = \lambda (u, v)_X \end{cases}$$

sorted in descending order (see [10], Theorem 4.5). We then obtain an algebraic decay rate for the n-width of B_Y

$$d_n(B_Y; X) = Cn^{-s/d},$$

where $s \geq 1$ similarly to (3). Provided that $\mathcal{K}_{h,p} \subset B_Y$ we obtain also an upper bound for the n-width of $\mathcal{K}_{h,p}$, since then $d_n(B_Y; X) \geq d_n(\mathcal{K}_{h,p}; X)$, consequently only algebraic convergence of the reduced basis method is predicted by (23). Such results can be misleadingly pessimistic compared to practical experiences with reduced basis methods because they do not take into account the structure of the manifold $\mathcal{K}_{h,p}$ nor the fact that approximation (4) inherits in some sense the structure of the manifold $\mathcal{K}_{h,p}$.

4 An Extended Result of Exponential Convergence

Let us now give an example of a μ-PDE where the explicit dependence of the solution manifold on the parameters can be exhibited. Let us consider the parameter-dependent problem after discretization:

$$\big(\Theta_1^a(\mu)A_1 + \Theta_2^a(\mu)A_2\big)u = \sum_{q=1}^{Q_f} \Theta_q^f(\mu)F_q. \tag{24}$$

We assume that (i) the operator A_1 is invertible; (ii) the problem satisfies a global condition for the spectral radius ρ being

$$\rho\left(\frac{\Theta_2^a(\mu)}{\Theta_1^a(\mu)}A_1^{-1}A_2\right) < 1 \quad \text{for all } \mu \in \mathcal{D}, \tag{25}$$

which we interpret as meaning that the term $\Theta_1^a(\mu)A_1$ dominates the original differential operator. Such a problem can arise for example from a discretized advection-diffusion or reaction-diffusion problem, where A_1 contains the (dominant) diffusion operator and A_2 contains all the other terms. We proceed to write explicitly the solution of this problem as

$$u = \left(I + \frac{\Theta_2^a(\mu)}{\Theta_1^a(\mu)}A_1^{-1}A_2\right)^{-1}\big(\Theta_1(\mu)A_1\big)^{-1}\left(\sum_{q=1}^{Q_f}\Theta_q^f(\mu)F_q\right),$$

which by exploiting the global spectral condition (25) leads to the series expansion for the solution

$$u = \sum_{k=0}^{\infty}\sum_{q=1}^{Q_f}\frac{(-1)^k\Theta_2^k(\mu)\Theta_q^f(\mu)}{\Theta_1^{k+1}(\mu)}\big[A_1^{-1}A_2\big]^k A_1^{-1}F_q.$$

By defining the fundamental basis vectors $\Psi_{k,q} := [A_1^{-1}A_2]^k A_1^{-1}F_q$, for $k = 0, 1, \ldots$ and $q = 1, \ldots, Q_f$, we can write the solution as a series

$$u = \sum_{k=0}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k [\Theta_2^a(\mu)]^k \Theta_q^f(\mu)}{[\Theta_1^a(\mu)]^{k+1}} \Psi_{k,q}. \tag{26}$$

Several remarks can be made about formula (26):

1. In the special case $A_2 = 0$ the parametric dependence enters only through the r.h.s. and as a consequence the series (26) truncates to a finite one

$$u = \sum_{q=1}^{Q_f} \frac{\Theta_q^f(\mu)}{\Theta_1^a(\mu)} \Psi_{0,q}, \tag{27}$$

   and so the greedy algorithm will always terminate after Q_f steps.
2. If the decay of the series coefficients in (26) is rapid, the solutions u can be well approximated by only the first few fundamental basis functions $\Psi_{k,q}$, $k = 0, 1, \ldots, K$ and $q = 1, \ldots, Q_f$. They can be computed according to an iterative procedure

$$\Psi_{0,q} = A_1^{-1}F_q, \qquad \Psi_{k+1,q} = A_1^{-1}A_2\Psi_{k,q} \quad \text{for all } q = 1, \ldots, Q_f$$

   requiring at each step one matrix multiplication and one backward substitution after obtaining once and for all the LU-decomposition of A_1.
3. In general the $\Psi_{k,q}$ are not linear combinations of solutions of (24) so that they do not constitute a reduced basis approximation. They are, however, useful for estimating the n-width of the solution set. Provided that there exist positive sequences $\{\gamma_{k,q}\}_{k=1}^{\infty}$ s.t. $\|\Psi_{k,q}\|_X \leq \gamma_{k,q}$ for each $q = 1, \ldots, Q_f$, we obtain an upper bound estimate for the n-width of the solution set $\mathcal{U}$ of (24)

$$d_m(\mathcal{U}; X) \leq \sup_{\mu \in \mathcal{D}} \sum_{k=n}^{\infty} \sum_{q=1}^{Q_f} \left| \frac{[\Theta_2^a(\mu)]^k \Theta_q^f(\mu)}{[\Theta_1^a(\mu)]^{k+1}} \right| \gamma_{k,q} \tag{28}$$

   by using the definition of the n-width, estimating upwards, and using formula (26) as

$$d_m(\mathcal{U};X) = \inf_{X_m \subset X} \sup_{\mu\in\mathcal{D}} \inf_{\widetilde{u}\in X_m} \left\| u_{h,p}(\mu) - \widetilde{u} \right\|_X \le \sup_{\mu\in\mathcal{D}} \inf_{\widetilde{u}\in X_m^\Psi} \left\| u_{h,p}(\mu) - \widetilde{u} \right\|_X$$

$$\le \sup_{\mu\in\mathcal{D}} \left\| u_{h,p}(\mu) - \sum_{k=0}^{n-1}\sum_{q=1}^{Q_f} \frac{(-1)^k[\Theta_2^a(\mu)]^k\Theta_q^f(\mu)}{[\Theta_1^a(\mu)]^{k+1}}\Psi_{k,q} \right\|_X$$

$$= \sup_{\mu\in\mathcal{D}} \left\| \sum_{k=n}^{\infty}\sum_{q=1}^{Q_f} \frac{(-1)^k[\Theta_2^a(\mu)]^k\Theta_q^f(\mu)}{[\Theta_1^a(\mu)]^{k+1}}\Psi_{k,q} \right\|_X$$

$$\le \sup_{\mu\in\mathcal{D}} \sum_{k=n}^{\infty}\sum_{q=1}^{Q_f} \left| \frac{[\Theta_2^a(\mu)]^k\Theta_q^f(\mu)}{[\Theta_1^a(\mu)]^{k+1}} \right| \|\Psi_{k,q}\|_X,$$

where $m := Q_f \cdot n$ and $X_m^\Psi := \operatorname{span}\{\Psi_{k,q} : k = 0,\dots,n-1, q = 1,\dots,Q_f\}$, i.e. the first m fundamental basis vectors. We have in fact decomposed the description of the manifold of solutions $\mathcal{U}$ into two parts: the *parametric regularity* is carried by the coefficients Θ_1, Θ_2, Θ_q^f, which can be taken just in $L^\infty(\mathcal{D})$ without affecting the n-width, and the *spatial regularity*, which is contained in the norm estimates $\gamma_{k,q}$ for the fundamental basis functions.

4. If the solution of (24) is approximated by the projection-based ROM in (14), i.e.

$$\mathbb{Y}_n\big[\Theta_1^a(\mu)A_1 + \Theta_2^a(\mu)A_2\big]\mathbb{X}_n^T U_n = \sum_{q=1}^{Q_f} \Theta_q^f(\mu)\mathbb{Y}_n F_q,$$

where the projectors were defined in (16), we obtain a similar formula for the reduced solution

$$u_n(\mu) = \sum_{k=0}^{\infty}\sum_{q=1}^{Q_f} \frac{(-1)^k[\Theta_2^a(\mu)]^k\Theta_q^f(\mu)}{[\Theta_1^a(\mu)]^{k+1}}\Psi_{k,q}^n,$$

but now with the reduced fundamental basis functions $\Psi_{k,q}^n$ defined as

$$\Psi_{0,q}^n = \big(\mathbb{Y}_n A_1 \mathbb{X}_n^T\big)^{-1}\mathbb{Y}_n F_q, \quad \Psi_{k+1,q}^n = \big(\mathbb{Y}_n A_1 \mathbb{X}_n^T\big)^{-1}\mathbb{Y}_n A_2 \mathbb{X}_n^T \Psi_{k,q}^n.$$

As a result, we obtain immediately the error representation formula

$$\varepsilon_n(\mu) := \big[u_{h,p} - \mathbb{X}_n^T u_n\big](\mu) = \sum_{k=0}^{\infty}\sum_{q=1}^{Q_f} \frac{(-1)^k[\Theta_2^a(\mu)]^k\Theta_q^f(\mu)}{[\Theta_1^a(\mu)]^{k+1}}\big(\Psi_{k,q} - \mathbb{X}_n^T\Psi_{k,q}^n\big).$$

Thus the quality of the ROM can directly be measured by observing how well it approximates the fundamental basis vectors, i.e. by looking at $\|\Psi_{k,q} - \mathbb{X}_n^T\Psi_{k,q}^n\|_X$ for all k and q.

5. Even if the global spectral condition (25) does not hold, we can try to expand the solution locally around different μ^* and obtain local approximation bases. This leads one to consider the *hp*-reduced basis method [9], where different reduced bases (analogous to *p*-refinement in the FEM) are constructed at different parts of the parameter domain (analogous to *h*-refinement in the FEM). Let $\mathcal{D}_1, \ldots, \mathcal{D}_M$ be a nonoverlapping subdivision of the original parameter domain $\mathcal{D}$ into M subdomains. The local spectral condition requires that in each subdomain $\mathcal{D}_m$

$$\exists i(m): \quad \rho\left(\frac{\Theta^a_{j(m)}(\mu)}{\Theta^a_{i(m)}(\mu)} A^{-1}_{i(m)} A_{j(m)}\right) < 1 \quad \text{for all } \mu \in \mathcal{D}_m, \text{ for } j(m) \neq i(m),$$

that is to say in each parameter subdomain $\mathcal{D}_m$ one of the terms A_q dominates, but the dominant part of the operator can change from subdomain to subdomain. If such a local spectral condition holds, our results extend straightforwardly to show the existence of local exponentially convergent approximation spaces.

With the n-width estimate (28) we can give an exponential convergence result extending that of [19]:

Proposition 4.1 *Assume that the series* (26) *converges, so that*

$$\exists \varepsilon > 0 \quad s.t. \quad \left|\frac{\Theta_2(\mu)}{\Theta_1(\mu)}\right| \leq \frac{1-\varepsilon}{\|A_1^{-1}A_2\|_X} \quad \textit{for all } \mu \in \mathcal{D}. \tag{29}$$

Then the n-width of the solution set $\mathcal{U}$ *of* (24) *converges exponentially, i.e.*

$$d_n(\mathcal{U}; X) \leq Ce^{-\alpha n} \quad \textit{for some } C, \alpha > 0. \tag{30}$$

Proof The n-width upper bound (28) gives for $m = n \cdot Q_f$

$$\begin{aligned}
d_m(\mathcal{U}; X) &\leq \sup_{\mu\in\mathcal{D}} \sum_{k=n}^{\infty} \sum_{q=1}^{Q_f} \left|\frac{[\Theta^a_2(\mu)]^k \Theta^f_q(\mu)}{[\Theta^a_1(\mu)]^{k+1}}\right| \left\| \left(A_1^{-1}A_2\right)^k A_1^{-1} F_q \right\|_X \\
&\leq Q_f \cdot \sup_{\mu,q} \left\{ \left|\frac{\Theta^f_q(\mu)}{\Theta_1(\mu)}\right| \left\| A_1^{-1} F_q \right\|_X \right\} \cdot \sum_{k=n}^{\infty} \left|\frac{[\Theta^a_2(\mu)]^k}{[\Theta^a_1(\mu)]^k}\right| \left\| A_1^{-1}A_2 \right\|_X^k \\
&= Q_f \cdot \sup_{\mu,q} \left\{ \left|\frac{\Theta^f_q(\mu)}{\Theta_1(\mu)}\right| \left\| A_1^{-1} F_q \right\|_X \right\} \cdot (1-\varepsilon)^n \sum_{k=0}^{\infty} (1-\varepsilon)^k \\
&= \frac{Q_f}{\varepsilon} \cdot \sup_{\mu,q} \left\{ \left|\frac{\Theta^f_q(\mu)}{\Theta_1(\mu)}\right| \left\| A_1^{-1} F_q \right\|_X \right\} \cdot \exp\left(\frac{\log(1-\varepsilon)}{Q_f} m\right),
\end{aligned}$$

so that the result holds with $\alpha = -\log(1-\varepsilon)/Q_f$ and $C = \frac{Q_f}{\varepsilon} \cdot \sup_{\mu,q}\{|\frac{\Theta^f_q(\mu)}{\Theta_1(\mu)}| \|A_1^{-1}F_q\|_X\}$. □

Exponential convergence of the greedy reduced basis algorithm is then predicted by [3] as in (22). It should be understood that tight n-width estimates for the rate of exponential convergence cannot be obtained by such series expansions—indeed the coefficient α will tend to 0 if we let $\varepsilon \to 0$. The factor $1/Q_f$ in the exponential is also excessively pessimistic provided that not for all q the terms converge at the same rate. In the next section we will demonstrate a problem where much faster exponential convergence of the greedy algorithm is observed, even in the parametric region when the fundamental series no longer converges rapidly.

To close this section let us briefly consider the more general case $Q_a > 2$:

$$\left(\Theta_1^a(\mu)A_1 + \sum_{r=2}^{Q_a} \Theta_r^a(\mu)A_2\right)u = \sum_{q=1}^{Q_f} \Theta_q^f(\mu)F_q. \tag{31}$$

If the global spectral condition

$$\rho\left(\sum_{r=2}^{Q_a} \frac{\Theta_r^a(\mu)}{\Theta_1^a(\mu)} A_1^{-1}A_r\right) < 1, \tag{32}$$

is satisfied, we can write the solution as

$$u = \left(I + \sum_{r=2}^{Q_a} \frac{\Theta_r^a(\mu)}{\Theta_1^a(\mu)} A_1^{-1}A_r\right)^{-1} \left(\Theta_1^a(\mu)A_1\right)^{-1} \left(\sum_{q=1}^{Q_f} \Theta_q^f(\mu)F_q\right),$$

and applying (32) leads to

$$u = \left(\sum_{k=0}^{\infty} \frac{(-1)^k}{[\Theta_1^a(\mu)]^{k+1}} \left[\sum_{r=2}^{Q_a} \Theta_r^a(\mu)A_1^{-1}A_r\right]^k A_1^{-1}\right)\left(\sum_{q=1}^{Q_f} \Theta_q^f(\mu)F_q\right)$$

and finally

$$u = \sum_{k=0}^{\infty}\sum_{q=1}^{Q_f} \frac{(-1)^k \Theta_q^f(\mu)}{[\Theta_1^a(\mu)]^{k+1}} \Psi_{k,q}(\mu), \tag{33}$$

but now the fundamental basis vectors

$$\begin{aligned} &\Psi_{0,q} = A_1^{-1}F_q, \\ &\Psi_{k+1,q}(\mu) = \left[\sum_{r=2}^{Q_a} \Theta_r^a(\mu)A_1^{-1}A_r\right]\Psi_{k,q}(\mu) \quad \text{for all } q = 1, \ldots, Q_f \end{aligned} \tag{34}$$

depend explicitly on the parameter(s) μ. Let $\rho^{(k)} = (\rho_1, \rho_2, \ldots, \rho_k)$ be a multi-index of dimension k and let $\rho^{(0)} = \emptyset$. We define a set of parameter-free basis functions $\varphi_{k,q,\rho}$ according to the recursion

$$\varphi_{0,q,\rho^{(0)}} = A_1^{-1}F_q, \qquad \varphi_{k+1,q,\rho^{(k+1)}} = A_1^{-1}A_{\rho_{k+1}}\varphi_{k,q,\rho^{(1:k)}}.$$

Using the parameter-free basis we can rewrite the recursion of the fundamental basis (34) as

$$\begin{cases} \Psi_{0,q} = \varphi_{0,q,r^{(0)}}, \\ \Psi_{1,q}(\mu) = \displaystyle\sum_{r=2}^{Q_a} \Theta_r^a(\mu) A_1^{-1} A_r \Psi_{0,q} = \sum_{r=2}^{Q_a} \Theta_r^a(\mu) \varphi_{1,q,(r)}, \\ \Psi_{2,q}(\mu) = \displaystyle\sum_{r'=2}^{Q_a} \Theta_{r'}^a(\mu) A_1^{-1} A_{r'} \Psi_{1,q} \\ \qquad\quad = \displaystyle\sum_{r'=2}^{Q_a} \sum_{r=2}^{Q_a} \Theta_{r'}^a(\mu) \Theta_r^a(\mu) A_1^{-1} A_{r'} \varphi_{1,q,(r)} \\ \qquad\quad = \displaystyle\sum_{r'=2}^{Q_a} \sum_{r=2}^{Q_a} \Theta_{r'}^a(\mu) \Theta_r^a(\mu) \varphi_{2,q,(r',r)}, \\ \Psi_{k,q}(\mu) = \displaystyle\sum_{r_1=2}^{Q_a} \cdots \sum_{r_k=2}^{Q_a} \Theta_{r_1}^a(\mu) \ldots \Theta_{r_k}^a(\mu) \varphi_{k,q,(r_k,r_{k-1},\ldots,r_1)} \end{cases}$$

and so the kth level expansion for $\Psi_{k,q}$ will contain in general $(Q_a - 1)^k$ terms, and the size of the expansion blows up exponentially. Without some strong structural assumptions the series expansion method is not suitable for deriving exponentially decaying n-width estimates in the case $Q_a \gg 1$.

5 Numerical Example of a Parameter-Dependent Diffusion Problem

In this section we shall give numerical evidence of exponential convergence of n-width upper bounds and consequently of the GRBM approximation. As a test problem we consider a diffusion problem in a disk with four circular subregions $\Omega_1, \ldots, \Omega_4$ as depicted in Fig. 2. The parametric problem can be formulated as follows: given $\mu \in \mathcal{D} \subset \mathbb{R}^8$, find $u = u(\mu)$ s.t.

$$-(1 + \mu_1 \mathbb{I}_\omega) \triangle u = \mu_4 \mathbb{I}_{\Omega \setminus \omega} + \sum_{q=1}^{4} \mu_{q+4} \mathbb{I}_{\Omega_q} \quad \text{in } \Omega,$$

$$u = 1 \quad \text{on } \Gamma_1 \cup \Gamma_4, \qquad \frac{\partial u}{\partial n} = \mu_2 \quad \text{on } \Gamma_2, \qquad \frac{\partial u}{\partial n} = \mu_3 \quad \text{on } \Gamma_3$$

where the Γ_k denote the four sides of the square, and $\omega := \bigcup_{q=1}^4 \Omega_q$ is the union of the disks. The function $\mathbb{I}_\Omega$ denotes the characteristic function of the subdomain Ω. Thus the first parameter μ_1 controls the difference between the isotropic diffusion coefficient inside the disks versus the background conductivity, while the rest of the parameters $\mu_2, \ldots, \mu_8$ enter into the boundary conditions and the source terms.

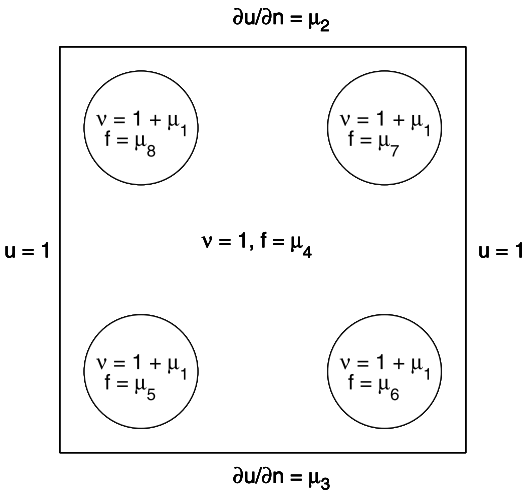


Fig. 2 Schematic description of the domain and boundary conditions of the model problem

This problem exhibits the same properties as the case discussed in the previous Sect. 4 so that the solution can be written as the combination of the fundamental basis vectors thanks to the formula (26). In this case the affine expansion (5) of the problem is given by:

$$\begin{cases} \Theta_1^a(\mu) = 1, & a_1(u,v) = \displaystyle\int_\Omega \nabla u \cdot \nabla v d\Omega, \\ \Theta_2^a(\mu) = \mu_1, & a_2(u,v) = \displaystyle\int_\omega \nabla u \cdot \nabla v d\Omega, \\ \Theta_1^f(\mu) = \mu_2, & f_1(v) = \displaystyle\int_{\Gamma_2} v d\Omega, \\ \Theta_2^f(\mu) = \mu_3, & f_2(v) = \displaystyle\int_{\Gamma_3} v d\Omega, \\ \Theta_3^f(\mu) = \mu_4, & f_3(v) = \displaystyle\int_{\Omega\setminus\omega} v d\Omega, \\ \Theta_4^f(\mu) = \mu_5, & f_4(v) = \displaystyle\int_{\Omega_1} v d\Omega, \\ \Theta_5^f(\mu) = \mu_6, & f_5(v) = \displaystyle\int_{\Omega_2} v d\Omega, \\ \Theta_6^f(\mu) = \mu_7, & f_6(v) = \displaystyle\int_{\Omega_3} v d\Omega, \\ \Theta_7^f(\mu) = \mu_8, & f_7(v) = \displaystyle\int_{\Omega_4} v d\Omega, \end{cases}$$

so that $Q_a = 2$, $Q_f = 7$, and the problem satisfies the global spectral condition (25) provided that $\mu_1 \in [-(1-\varepsilon), 1-\varepsilon]$ for some $\varepsilon > 0$.

In order to compare the n-width bounds with the observed convergence rates of the weak greedy algorithm, we considered four different cases: $\varepsilon = 0.1$, $\varepsilon = 0.5$, $\varepsilon = 0.9$, and $\varepsilon = 0.95$. Note that if $\varepsilon = 1$, the manifold of parametric solutions

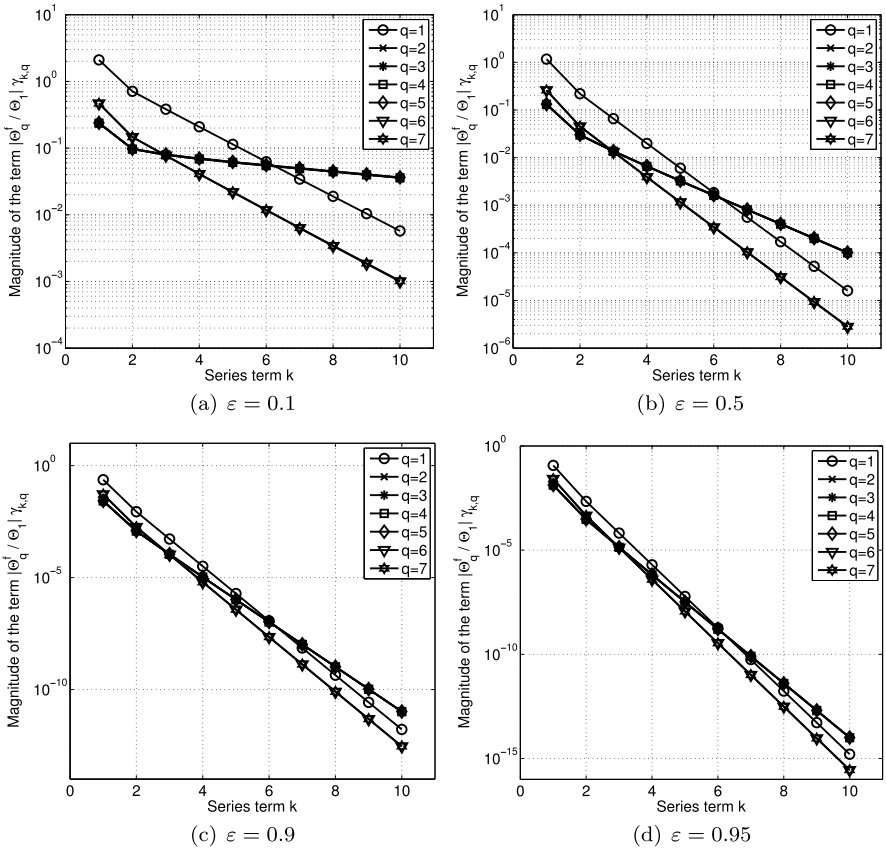


Fig. 3 Convergence of the fundamental series (26) coefficients for different values of ε in (29)

dimension is limited to a Q_f-dimensional subspace of X as indicated by (27), and so the greedy terminates after exactly 7 iterations. In Fig. 3 we have plotted the convergence of the fundamental series terms $\sup_\mu \|(\Theta_2^a(\mu)\Theta_q^f(\mu)/\Theta_1^a(\mu))\Psi_{k,q}\|_X$ that dictate the convergence rate of the n-width upper bound (30). For the value $\varepsilon = 0.1$ very weak convergence of the fundamental series is observed for some of the terms, namely $q = 2, 3, 4, 5$.

To obtain the GRBM approximation the weak greedy algorithm was driven by the residual-based a posteriori error estimator (20). In both cases the greedy was run until an absolute H^1-error below 10^{-3} was reached. This required $n = 25$ basis functions for the case $\varepsilon = 0.1$, $n = 21$ basis functions for the case $\varepsilon = 0.5$, $n = 15$ basis functions for the case $\varepsilon = 0.9$, and $n = 14$ basis functions for the case $\varepsilon = 0.95$. In Fig. 4 we have plotted the corresponding convergence rates of the greedy algorithm compared to the n-width upper bound predictions given by (30). In each case exponential convergence of the GRBM approximation is observed. The actual exponential decay rate depends on ε, where for $\varepsilon = 0.1$ the n-width estimate is much

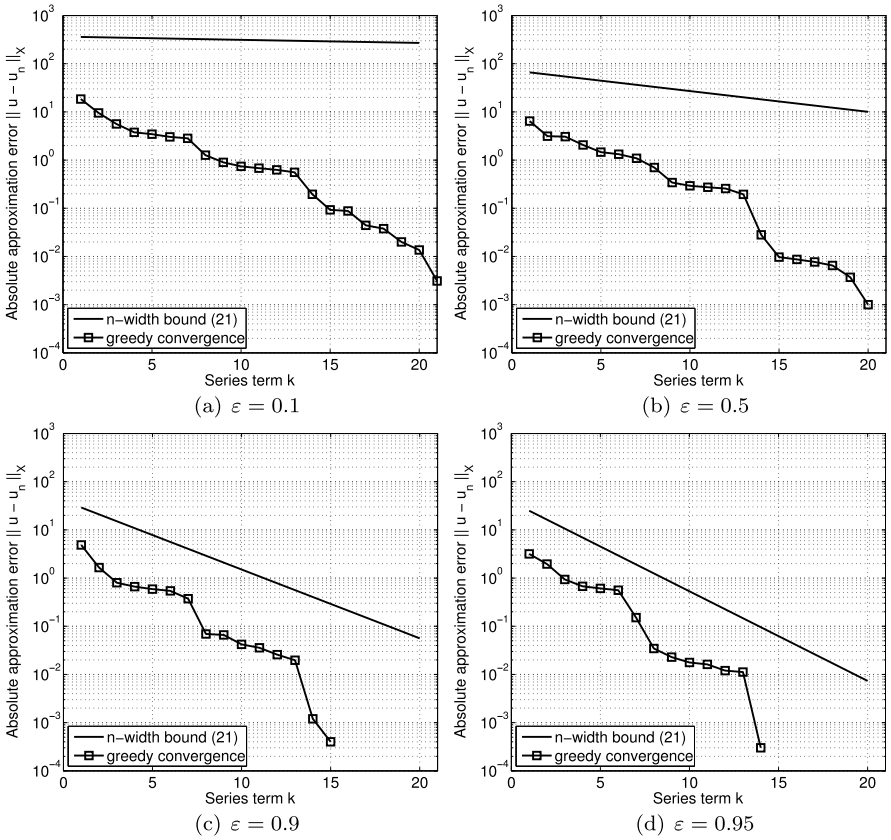


Fig. 4 Comparison of the n-width upper bound estimate (30) and the greedy convergence rate

too pessimistic when compared to the true rate of convergence. This is likely due to the weak convergence of some of the fundamental series terms (see Fig. 3(a)), and the result could be improved by considering more carefully the cutoff point for the different series terms for different q. However, as $\varepsilon \to 1$ the n-width estimate (30) becomes more and more indicative of the convergence rate observed during the greedy algorithm. According to Fig. 3(c–d) at the limit all the fundamental series coefficients converge at roughly the same rate, so that the bound (30) is expected to sharpen considerably.

6 Conclusions

We have reviewed the generalized reduced basis method for the uniform approximation of manifolds of solutions of parametric partial differential equations. These methods are typically driven by a greedy algorithm for selecting near-optimal re-

duced approximation subspaces. It has recently been shown that the convergence rate of the generalized reduced basis approximations is linked to the Kolmogorov n-width of the manifold of solutions. We have exhibited a model problem where the exact parameter-dependent solution can be expanded as a Neumann series, leading to a constructive proof that the n-width of the solution set in this case converges exponentially. Numerical experiments confirm that the reduced basis approximation also converges exponentially, and with a rate that is comparable to the one predicted by our n-width upper bound estimate. The predicted convergence rate is independent of the parametric regularity of the solution manifold and the number of parameters, but it does depend on the size of the affine expansion of the parametric problem. Future work involves finding more cases of parameter-dependent problems, where explicit solution formulas could be used to prove more general n-width estimates.

Acknowledgements This contribution is published in BUMI (Bollettino Unione Matematica Italiana) as well by a special agreement between UMI (Unione Matematica Italiana) and Springer. This work celebrates the memory of Prof. Enrico Magenes (1923–2010) after the conference held in Pavia in November 2011.

References

1. Adomian, G.: Solving Frontier Problems of Physics: The Decomposition Method. Kluwer Academic, Norwell (1994)
2. Babuška, I., Szabo, B.A., Katz, I.N.: The p-version of the finite element method. SIAM J. Numer. Anal. **18**(3), 515–545 (1981)
3. Binev, P., Cohen, A., Dahmen, W., DeVore, R., Petrova, G., Wojtaszczyk, P.: Convergence rates for greedy algorithms in reduced basis methods. SIAM J. Math. Anal. **43**, 1457–1472 (2011)
4. Buffa, A., Maday, Y., Patera, A.T., Prud'homme, C., Turinici, G.: A priori convergence of the greedy algorithm for the parametrized reduced basis method. ESAIM Math. Model. Numer. Anal. **46**(3), 595–603 (2012)
5. Canuto, C., Tonn, T., Urban, K.: A-posteriori error analysis of the reduced basis method for non-affine parameterized nonlinear PDEs. SIAM J. Numer. Anal. **47**(3), 2001–2022 (2009)
6. Carlberg, K., Farhat, C.: A low-cost, goal-oriented 'compact proper orthogonal decomposition' basis for model reduction of static systems. Int. J. Numer. Methods Eng. **86**(3), 381–402 (2011)
7. Cohen, A., DeVore, R., Schwab, C.: Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs. Anal. Appl. **9**(1), 11–47 (2011)
8. Deparis, S.: Reduced basis error bound computation of parameter-dependent Navier-Stokes equations by the natural norm approach. SIAM J. Numer. Anal. **46**(4), 2039–2067 (2008)
9. Eftang, J.L., Knezevic, D.J., Patera, A.T.: An hp certified reduced basis method for parametrized parabolic partial differential equations. Math. Comput. Model. Dyn. Syst. **17**(4), 395–422 (2011)
10. Evans, J.A., Bazilevs, Y., Babuška, I., Hughes, T.J.R.: N-widths, sup-infs, and optimality ratios for the k-version of the isogeometric finite element method. Comput. Methods Appl. Mech. Eng. **198**(21–26), 1726–1741 (2009)
11. Grepl, M.A., Maday, Y., Nguyen, N.C., Patera, A.T.: Efficient reduced-basis treatment of non-affine and nonlinear partial differential equations. ESAIM Math. Model. Numer. Anal. **41**(3), 575–605 (2007)

12. Hay, A., Borggaard, J.T., Pelletier, D.: Local improvements to reduced-order models using sensitivity analysis of the proper orthogonal decomposition. J. Fluid Mech. **629**, 41–72 (2009)
13. Hay, A., Borggaard, J., Akhtar, I., Pelletier, D.: Reduced-order models for parameter dependent geometries based on shape sensitivity analysis. J. Comput. Phys. **229**(4), 1327–1352 (2010)
14. Huynh, D.B.P., Rozza, G., Sen, S., Patera, A.T.: A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. C. R. Acad. Sci. Paris. Sér. I Math. **345**, 473–478 (2007)
15. Huynh, D.B.P., Knezevic, D., Chen, Y., Hesthaven, J., Patera, A.T.: A natural-norm successive constraint method for inf-sup lower bounds. Comput. Methods Appl. Mech. Eng. **199**(29–32), 13 (2010)
16. Ito, K., Ravindran, S.S.: A reduced order method for simulation and control of fluid flows. J. Comput. Phys. **143**(2), 403–425 (1998)
17. Lassila, T., Manzoni, A., Rozza, G.: On the approximation of stability factors for general parametrized partial differential equations with a two-level affine decomposition. ESAIM Math. Model. Numer. Anal. **46**, 1555–1576 (2012)
18. Maday, Y.: Reduced basis method for the rapid and reliable solution of partial differential equations. In: Proceedings of the International Congress of Mathematicians, vol. III, pp. 1255–1270. Eur. Math. Soc., Zürich (2006)
19. Maday, Y., Patera, A.T., Turinici, G.: Global a priori convergence theory for reduced-basis approximation of single-parameter symmetric coercive elliptic partial differential equations. C. R. Acad. Sci. Paris. Sér. I Math. **335**, 1–6 (2002)
20. Maday, Y., Patera, A.T., Turinici, G.: A priori convergence theory for reduced-basis approximations of single-parametric elliptic partial differential equations. J. Sci. Comput. **17**(1–4), 437–446 (2002)
21. Manzoni, A.: Reduced models for optimal control, shape optimization and inverse problems in haemodynamics. PhD thesis, N. 5402, École Polytechnique Fédérale de Lausanne (2012)
22. Melenk, J.M.: On n-widths for elliptic problems. J. Math. Anal. Appl. **247**, 272–289 (2000)
23. Melkman, A.A., Micchelli, C.A.: Spline spaces are optimal for L^2 n-width. Ill. J. Math. **22**(4), 541–564 (1978)
24. Noor, A.K.: Recent advances in reduction methods for nonlinear problems. Comput. Struct. **13**(1–3), 31–44 (1981)
25. Peterson, J.S.: The reduced basis method for incompressible viscous flow calculations. SIAM J. Sci. Stat. Comput. **10**, 777–786 (1989)
26. Pinkus, A.: n-Widths in Approximation Theory. Ergebnisse der Mathematik und ihrer Grenzgebiete (3). Springer, Berlin (1985)
27. Quarteroni, A., Rozza, G., Manzoni, A.: Certified reduced basis approximation for parametrized PDEs and applications. J. Math. Ind. **1**, 3 (2011). doi:10.1186/2190-5983-1-3
28. Rozza, G., Huynh, D.B.P., Patera, A.T.: Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. Arch. Comput. Methods Eng. **15**, 229–275 (2008)
29. Tonn, T., Urban, K., Volkwein, S.: Comparison of the reduced-basis and POD a posteriori error estimators for an elliptic linear-quadratic optimal control problem. Math. Comput. Model. Dyn. Syst. **17**(4), 355–369 (2011)
30. Veroy, K., Patera, A.T.: Certified real-time solution of the parametrized steady incompressible Navier-Stokes equations: rigorous reduced-basis a posteriori error bounds. Int. J. Numer. Methods Fluids **47**(8–9), 773–788 (2005)

Variational Formulation of Phase Transitions with Glass Formation

Augusto Visintin

Abstract In the framework of the theory of nonequilibrium thermodynamics, phase transitions with glass formation in binary alloys are here modeled as a multi-nonlinear system of PDEs. A weak formulation is provided for an initial- and boundary-value problem, and existence of a solution is studied. This model is then reformulated as a minimization problem, on the basis of a theory that was pioneered by Fitzpatrick [MR 1009594]. This provides a tool for the analysis of compactness and structural stability of the dependence of the solution(s) on data and operators, via De Giorgi's notion of Γ-convergence. This latter issue is here dealt with in some simpler settings.

Foreword Enrico Magenes was an outstanding mathematician, and founded an internationally renowned school. But to many persons he was much more than that, and his charismatic personality influenced the Italian and the international mathematical world. He was a determined and efficient worker; had a great ability in getting people motivated towards shared purposes, especially research; and was of example in any aspect of his life.

I first met Him in 1973 as a third-year student of mathematics at the University of Pavia, after two years of teaching of analysis by Claudio Baiocchi; these two encounters much contributed to orient me towards this branch of mathematics, and still influence my activity as a researcher and as a teacher. When the moment of choosing the thesis came, I asked some of my former teachers. I wished to write a thesis in analysis, and wondered whether I might ask Baiocchi, or Gianni Gilardi, or someone else. They told me that I had little choice: *il Capo* (the Boss, as He was often named) intended to be my advisor. I followed that suggestion, and He introduced me to boundary-value problems for P.D.E.s and to the Stefan problem.

To the memory of Enrico Magenes: an anti-fascist Resistance fighter, a charismatic leader, and more than that.

A. Visintin (✉)
Dipartimento di Matematica, Università di Trento, via Sommarive 14, Povo di Trento, 38050 Trento, Italy
e-mail: visintin@science.unitn.it

F. Brezzi et al. (eds.), *Analysis and Numerics of Partial Differential Equations*, Springer INdAM Series 4, DOI 10.1007/978-88-470-2592-9_17,

That started a collaboration that left me a great freedom of research. I could then investigate some physical aspects of phase transitions, and also address the modeling of hysteresis phenomena. It is in the spirit of those times and of that freedom that here I wish to revisit an extension of the Stefan model, with an eye for the model and one for some recently-developed analytical issues.

I would like to conclude this short souvenir mentioning that in the last years of His life we met several times in the mountains near Trento, where He used to spend a part of the Summer. In those talks I could learn about His past activity in the anti-fascist Resistance and His experience as a Dachau deportee: this revealed to me another aspect of His active and generous personality.

1 Introduction

This note is partially based on a talk that this author gave at a conference in memory of Professor Enrico Magenes, in Pavia in November 2011. That speech was devoted to recent advances in Fitzpatrick's theory on the variational representation of maximal monotone operators, and on its use to prove the structural stability of quasilinear PDEs. Those results are here reviewed, and are applied to some evolutionary problems. In this note a variational formulation is also provided for a model of phase transitions with glass formation in heterogeneous systems, that was proposed in [67] and is here reviewed, too. The goal of proving the structural stability of that problem is more demanding; here some features of that question are just discussed.

Stefan-Type Problems Phase transitions occur in many relevant processes in physics and engineering. In 1889 the physicist Josef Stefan [56–59] proposed a one-dimensional model, that accounted for heat diffusion and exchange of latent heat in the melting of the polar ice. The analytical formulation consisted in what is now called a *free boundary* (or *moving boundary*) *problem*, for a parabolic equation. This definition refers to the fact that the evolution of the surfaces that separate the phases is not known a priori: the relevant PDE actually holds in a space-time set, of which part of the boundary is free. On this unknown boundary a discontinuity condition is then prescribed.

That model was then extended in many ways, and an intense research started into two directions: phase transitions and free boundary problems. This involved a large number of physicists, engineers and mathematical analysts, giving rise to tens of monographs and tens of thousands of papers in journals. Much of those models extend the formulation introduced by Stefan, and are often labeled under the general denomination of *Stefan-type problems*.

One of the variants of the basic Stefan model concerns phase transitions in heterogeneous systems; in this case heat and mass diffusion are coupled. A first description simply consists in coupling the Fourier and Fick diffusion laws, and prescribing appropriate conditions at the phase interfaces. This formulation however

exhibits substantial physical and analytical shortcomings, that are strictly related to inconsistency with the second principle of thermodynamics. A more appropriate model stems from a neat theory that is known as *nonequilibrium* (or *irreversible*) *thermodynamics*, and is based on the second principle.

Glass Formation Here we are concerned with glass formation, namely the onset of an amorphous phase that retains (either all or at least a large part of) its latent heat of phase transition. This is an important physical phenomenon and has relevant industrial applications: many manufactured products are the outcome of a process of phase transition, and a part of them either consists in or includes a glassy phase. Polymers are also examples of amorphous materials.

A glassy phase may be formed by undercooling a liquid, because of an impressive increase (up to 18 orders of magnitude) of viscosity associated to a sufficiently deep undercooling. This requires the undercooling to be sufficiently rapid to prevent crystallization: in this case the disordered atomic configuration that is typical of the liquid phase is *frozen* into the solid state. The solid behavior of glasses is thus not due to a crystal structure, but to extremely high viscosity. Amorphous phases may persist for a long time (even millennia) in a state that is far from equilibrium. Remarkable examples of this phenomenon are provided by the windows of ancient cathedrals, which however in some cases exhibit traces of crystallization.

By what we just pointed out, glass formation is related to the process rather than just the state temperature. In order to account for this phenomenon, we represent phase transitions via a first-order dynamics, named *phase relaxation*, and model glass formation by prescribing a nonmonotone *kinetic function* (which represents the relation between transition rate and undercooling). This entails that the solid-liquid transition zone is not reduced to a surface, (in the jargon of the Stefan-milieu, this is usually labeled as the onset of a *mushy region*) so that the resulting model is not a free boundary problem.

Most of the industrial applications of phase transitions involve composite materials. Here we then deal with glass formation in (noneutectic) binary alloys. In this case the phase transition and glass formation temperatures and more generally the kinetic law of phase relaxation depend on the concentration of the two components, namely on the composition. The problem that here we consider is just a first step towards a more detailed model; for instance, this should also account for mechanical effects.

A Doubly Nonlinear Equation The model that we derive, see Problem 4.1, is an initial- and boundary-value problem for a multi-nonlinear system of the form

$$\begin{cases} \Theta \in \partial\varphi(U) \\ J = -\gamma(\Theta, \nabla\Theta) \\ D_t U + \nabla \cdot J = f(\Theta) \end{cases} \qquad \text{in } Q := \Omega \times]0, T[\ (D_t := \partial/\partial t); \tag{1}$$

here Ω is a Euclidean domain and T is a positive constant. By $\partial\varphi$ we denote the subdifferential of a convex potential φ; γ is continuous with respect to its first argument and maximal monotone with respect to the second one. Denoting by φ^* the

Fenchel conjugate function of φ, this system also reads as a single inclusion:

$$D_t \partial\varphi^*(\Theta) - \nabla \cdot \gamma(\Theta, \nabla\Theta) \ni F(\Theta) \quad \text{in } Q. \tag{2}$$

(By φ^* we denote the Fenchel convex conjugate of φ.) Under suitable restrictions, the operator $H_0^1(\Omega) \to H^{-1}(\Omega) : \Theta \mapsto -\nabla \cdot \gamma(S, \nabla\Theta)$ is maximal monotone, for any admissible S.

Apart from the nonlinear second member, Eq. (2) may be compared with *doubly nonlinear equations* of the form

$$D_t \beta(\Theta) + \alpha(\Theta) \ni 0 \quad \text{with } \alpha \text{ and } \beta \text{ maximal monotone.} \tag{3}$$

The case in which for instance β is linear is quite easier, and corresponds to a *monotone flow*:

$$D_t \Theta + \alpha(\Theta) \ni 0 \quad \text{with } \alpha \text{ maximal monotone.} \tag{4}$$

Structural Stability A basic feature of modeling is that data (e.g., initial and/or boundary conditions) and operators (e.g., $\partial\varphi$ and γ in (1)) are known only with some approximation. It is then of interest to devise topologies that provide the stability of the problem in the following sense: whenever the data and the operators converge, the corresponding solutions u_n weakly converge to a solution of the asymptotic problem (up to a subsequence); this is close to the notions of G-convergence and H-convergence.

Results have been established for the problem (4). They rest upon three main ingredients:

(i) a variational formulation of maximal monotone operators (including evolutionary ones, such as those representing diffusion or phase relaxation); this is based on a theory that was pioneered by Fitzpatrick in [30];
(ii) the definition of a suitable nonlinear notion of convergence in function spaces, see [72];
(iii) the use of De Giorgi's theory of Γ-convergence, see [20, 21].

Plan of Work This paper consists of two parts, that merge just in the final section, and may thus be read independently.

The first two sections deal with a model of phase transition with glass formation in binary alloys that was first formulated in [67]. More specifically, in Sect. 2 we review a model of phase relaxation with glass formation, and in Sect. 3 we couple it with heat and mass diffusion in binary alloys, along the lines of the theory of nonequilibrium thermodynamics. Next in Sect. 4 we formulate a nonlinear problem in the framework of Sobolev spaces; this consists in an initial- and boundary-value problem for two quasilinear PDEs, which are coupled with a nonlinear ordinary differential equation. We review a result of [67] on the existence of a weak solution of that problem, that is based on so-called *compactness by strict convexity*. Via a compactness argument that is based on an additional a priori estimate, we then

prove a novel existence theorem, that provides existence of a solution even if the phenomenological laws have no potential.

The second part concerns the variational formulation and the structural stability of first-order flows. First in Sect. 5 we state the Fitzpatrick theorem, and illustrate how De Giorgi's theory of Γ-convergence may be used to study the compactness and structural stability of a wide class of monotone PDEs, along the lines of [72]. In Sect. 6 we then apply those techniques to Eq. (4): we provide a variational formulation in term of what we name a *null-minimization problem*, and prove its structural stability. In Sect. 7 we extend the variational formulation to the flow (2), partially along the lines of [69], where the structural stability is also addressed. (The results of [69] might however be refined on the basis of the present analysis: in particular the compactness of the family of operators might be proved; this might be illustrated in a work apart.) In Sect. 7 we provide a variational formulation of doubly nonlinear flows of the form (3), and then of the above model of phase transitions with glass formation.

A large part of this paper revisits previous works, but it also includes some novel results. These comprise a new result of existence of a weak solution for the glass formation problem (Theorem 5.4), and the variational formulation of nonmonotone flows (see Sects. 7 and 8). The discussion of the variational formulation of monotone flows (see Sect. 6) also includes elements of novelty with respect to [72].

Literature Mathematical models of phase transitions have been studied in a large number of works; see e.g. the monographs of Alexiades and Solomon [1], Brokate and Sprekels [12], Elliott and Ockendon [29], Frémond [33], Gupta [36], V. [64], and the survey V. [65]. Further references may be found in the comprehensive bibliography of Tarzia [63]. Physical and engineering aspects of phase transitions, especially of solidification of metals, have been treated e.g. by Chalmers [16], Christian [18], Flemings [31], Kurz and Fisher [39], Woodruff [73].

The coherent picture of the theory of nonequilibrium thermodynamics was first formulated by Eckart [27, 28] in 1940; see e.g. the accounts of Müller and Weiss [47–49]. That work formed the basis of a comprehensive theory that was then developed by Meixner, Prigogine, Onsager, De Groot, Mazur and other physicists; this is now also called *thermodynamics of irreversible processes*. See e.g. Callen [15], De Groot [22], De Groot and Mazur [23], Kondepudi and Prigogine [38], Prigogine [53]. Some papers also applied that approach to phase transitions in heterogeneous systems, see e.g. Donnelly [25], Luckhaus and V. [42], Alexiades, Wilson and Solomon [2], Luckhaus [41], V. [64]; Chap. V and [65, 67]. Nonequilibrium thermodynamics is also at the basis of a celebrated model of phase transitions in homogeneous materials, that was proposed by Penrose and Fife in [51, 52].

Doubly-nonlinear parabolic problems were dealt with in a number of works, see e.g. DiBenedetto and Showalter [24] and Alt and Luckhaus [3]. Here we also use techniques of [19] and [55]. Further references may be found e.g. in [69].

The theorem on the variational representation of maximal monotone operators was proposed by Fitzpatrick [30] in 1988, and was then rediscovered by Martinez-Legaz and Théra [45] and (independently) by Burachik and Svaiter [13]. This started

an intense research, see e.g. [14, 34, 43, 44, 46], Ghossoub's monograph [35], and several other contributions.

The theory of Γ-convergence was pioneered by De Giorgi and Franzoni [21] in 1975, and then extensively developed by the Pisa school and others; see e.g. [5, 7, 8, 20]. A compactness result for a notion of nonlinear G-convergence of quasilinear maximal monotone operators in divergence form was also proved in [17]. This is based on a different approach from the present one, but a comparison may be of some interest. More recently in [32] H-convergence was also applied to the homogenization of nonlinear quasilinear elliptic operators; see also [4].

The present work is part of an ongoing research on the variational representation of (nonlinear) evolutionary P.D.E.s, and on the application of variational techniques to the analysis of their structural stability, see e.g. [66, 69, 71, 72]. A somehow comparable program, based on the use of the Fitzpatrick theory, has been accomplished for the homogenization of quasilinear flows, see e.g. [68] and references therein.

2 Phase Relaxation and Glass Formation

Phase Relaxation Let us first consider a homogeneous liquid-solid system, and assume that the two phases are separated by a (smooth) sharp interface $\mathcal{S}$, that moves with speed $\boldsymbol{v}$ ($\in \mathbb{R}^3$). Let us denote by $\boldsymbol{n}$ the unit normal field to $\mathcal{S}$ oriented from the liquid to the solid. Neglecting curvature effects, at and near equilibrium the interface is at the absolute temperature $\tau = \tau_E$; that is, setting $\theta := \tau - \tau_E$,

$$\theta = 0 \quad \text{on } \mathcal{S}. \tag{5}$$

At higher temperature rates one may instead assume a *kinetic law* of phase transition of the form

$$\nu \boldsymbol{v} \cdot \boldsymbol{n} = \widetilde{g}(\theta) \quad \text{on } \mathcal{S}. \tag{6}$$

Here by ν we denote a viscosity coefficient, and $\widetilde{g}$ is a prescribed continuous function $\mathbb{R} \to \mathbb{R}$ such that

$$\widetilde{g}(\theta)\theta \geq 0 \quad \forall \theta \in \mathbb{R}, \tag{7}$$

see Fig. 1. In the framework of a weak formulation of phase transition, we drop the assumption of sharp interface $\mathcal{S}$, and allow for the occurrence of a so-called *mushy region*, namely, a fine-scale solid-liquid mixture. Denoting by ρ the liquid concentration (which is proportional to the content of latent heat of phase transition), we define the *phase function* $\chi := 2\rho - 1$. Thus $-1 \leq \chi \leq 1$, and

$$\begin{aligned} &\chi = -1 \quad \text{in the solid,} \\ &\chi = 1 \quad \text{in the liquid,} \\ &-1 < \chi < 1 \quad \text{in the mushy region.} \end{aligned} \tag{8}$$

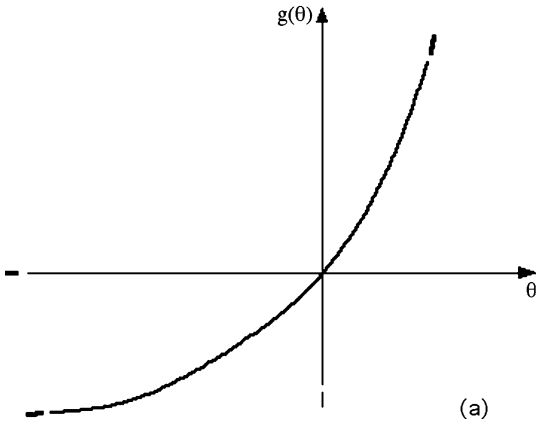


Fig. 1 Monotone kinetic function for a crystallizing material in (**a**), for the kinetic law $v\mathbf{v}\cdot\mathbf{n} = g(\theta)$

We then replace the interface dynamics (6) by a law of *phase relaxation*:

$$v D_t \chi + \partial I_{[-1,1]}(\chi) \ni \widetilde{g}(\theta) \quad \text{in } Q; \tag{9}$$

here

$$I_{[-1,1]}(\xi) := \begin{cases} 0 & \text{if } \xi \in [-1,1], \\ +\infty & \text{otherwise,} \end{cases} \tag{10}$$

and we denote by ∂ the *subdifferential* operator of convex analysis (see e.g. [26, 37, 54]).

It should be noticed that in general (6) and (9) are far from being equivalent: (6) represents phase transition by displacement of the solid-liquid front, whereas (9) accounts for phase transition by formation and growth of a mushy region; see e.g. [64]; Sect. V.1.

Glass Formation For most of substances a liquid tends to crystallize whenever $\theta < 0$, and symmetrically a solid tends to melt if $\theta > 0$. The kinetic function $\widetilde{g}$ may accordingly be assumed to be nondecreasing. If close to the interfaces and in the mushy region the temperature rate is sufficiently small, then $\widetilde{g}$ may also be linearized in a neighborhood of $\theta = 0$. This applies to systems close to thermodynamic equilibrium.

Glass formation is due to a strong increase of viscosity that impairs the mobility of particles in their migration towards the crystal sites, and thus prevents the formation of the crystal lattice. This phenomenon is thus related to the temperature dynamics, and requires the undercooling to be sufficiently fast as well as sufficiently deep. In several cases the latter requirement may be expressed in the form

$$\theta \le \theta^*, \quad \text{for a material-dependent threshold } \theta^* < 0. \tag{11}$$

Next we provide a quantitative representation of these requirements.

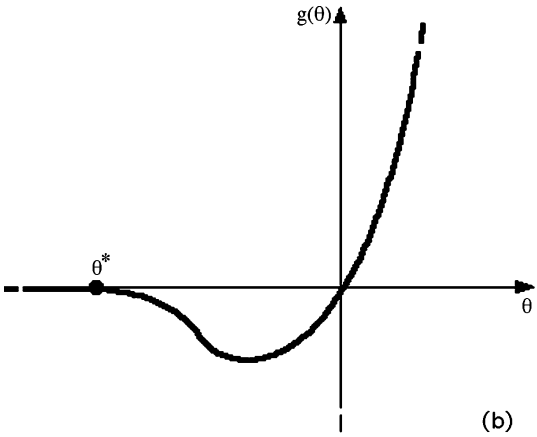


Fig. 2 Nonmonotone function for an amorphous material in (**b**), for the kinetic law $\nu\mathbf{v}\cdot\mathbf{n} = g(\theta)$

As the temperature dependence of the viscosity is the main feature of the glass behavior, in (6) and (9) we replace the constant ν by $\widehat{\nu}(\theta)$, for a prescribed function $\widehat{\nu}:\mathbb{R}\to\,]0,+\infty[$ such that

$$\widehat{\nu}(\theta)\gg 1,\quad \forall\theta<\theta^*. \tag{12}$$

Next we divide both members of (9) by $\widehat{\nu}(\theta)$; notice that, as $\widehat{\nu}(\theta)>0$,

$$\widehat{\nu}(\theta)^{-1}I_{[-1,1]}(\chi)=I_{[-1,1]}(\chi).$$

Moreover, setting $\bar{g}(\theta):=\widetilde{g}(\theta)/\widehat{\nu}(\theta)$, by (12) we have $|\bar{g}(\theta)|\ll 1$ for any $\theta<\theta^*$. It is then natural to assume that

$$\bar{g}(\theta)\theta\geq 0\quad\forall\theta\in\mathbb{R},\qquad \bar{g}(\theta)=0\quad\forall\theta<\theta^*, \tag{13}$$

see Fig. 2.

By (9) we then get the equivalent inclusion

$$D_t\chi+\partial I_{[-1,1]}(\chi)\ni\bar{g}(\theta)\quad\text{in } Q, \tag{14}$$

which is in turn equivalent to the following variational inequality:

$$\begin{cases}\chi\in[-1,1]\\ D_t\chi(\chi-v)\leq\bar{g}(\theta)(\chi-v)\quad\forall v\in[-1,1]\end{cases}\quad\text{in } Q. \tag{15}$$

(Henceforth we shall drop the tilde and the bar, and write g in place of $\widetilde{g}$ and $\bar{g}$.) Thus $D_t\chi=0$ where either

(i) $\theta=0$, or
(ii) $\theta>0$ and $\chi=1$, or
(iii) $\theta^*<\theta<0$ and $\chi=-1$, or
(iv) $\theta\leq\theta^*$.

That is, there is no phase transition at equilibrium (cases (i), (ii), (iii)) as well as in the glassy phase (case (iv)).

Dealing with heterogeneous substances this model must be amended, since the two-phase equilibrium temperature also depends on the composition.

3 Nonequilibrium Thermodynamics

In this section we review some basic elements of the theory of nonequilibrium thermodynamics, and then formulate a model of glass formation.

Eckart's Theory of Nonequilibrium Thermodynamics Next we deal with processes of coupled heat and mass diffusion with phase transition in a binary alloy, namely, a composite of two substances whose constituents are intermixed at the atomic scale.

A basic model consists in coupling the Fourier and Fick laws with appropriate conditions at the phase interface, that respectively account for heat and mass conservation. This approach has been used by material scientists and engineers, but exhibits some physical and mathematical shortcomings. Actually this model does not account for cross-effects between heat and mass diffusion. In several cases the omitted terms are not very significant quantitatively; this explains why the above approach may produce fairly acceptable numerical results. However, this model is not consistent with the second principle of thermodynamics, and of course this is quite regrettable from a theoretical viewpoint. This inconvenience also has a relevant analytical counterpart: the diffusive part of this model is represented by a system that does not have the structure of a gradient flow. As far as this author knows, in the multivariate setting no solution is known to exist even for the weak formulation.

These physical and mathematical drawbacks are overcome by a different model, that is formulated in the framework of the theory of *nonequilibrium thermodynamics*, that we now illustrate. This neat theory was first formulated by Eckart in 1940, and then exploited by Meixner, Prigogine, Onsager, De Groot, Mazur and many other physicists; see e.g. [47–49]. Here the constitutive relations are dictated by the very exigency of fulfilling the second principle. More specifically, this method provides the entropy estimate, and with that a priori estimates that contribute to make the analysis rather natural.

Next we confine ourselves to a composite of two constituents: a *binary alloy*, that is, a homogeneous mixture of two substances, that are soluble in each other in all proportions in each phase, outside a critical range of temperatures. We label this mixture as *homogeneous* since the constituents are intermixed on the atomic length-scale to form a single phase, either solid or liquid. We regard one of the two components, for instance that with the lower solid-liquid equilibrium temperature, as the *solute*—the other one as the *solvent*. We confine ourselves to a nonreacting and noneutectic binary system, although this analysis might be extended to include chemical reactions in multi-component systems.

The model that here we consider consists in two balance laws and appropriate constitutive relations:

(i) the principle of mass conservation,
(ii) the principle of energy conservation (i.e., the first principle of thermodynamics),
(iii) a constitutive relation that relates the entropy density, the temperature, the solute concentration, and the phase function (i.e., a *Gibbs-type formula*),
(iv) two constitutive relations for the energy and mass fluxes (the so-called *phenomenological laws*),
(v) a relaxation dynamics for the phase function.

The prescriptions (iv) and (v) will account for a local formulation of the second principle of thermodynamics. This will yield a parabolic doubly-nonlinear system of PDEs.

Balance Laws and Gibbs-Type Formula We shall use the following notation:

u: density of internal energy,
s: density of entropy,
τ: absolute temperature,
c: concentration of the solute (per unit volume), with $0 \le c \le 1$,
μ: difference between the *chemical potentials* of the two constituents,
λ: difference between the density of internal energy of the two phases (at constant entropy and concentration),
$\boldsymbol{j}_u$: flux of energy (per unit surface), due to flux of heat and mass,
$\boldsymbol{j}_c$: flux of the solute (per unit surface),
h: intensity of a prescribed energy source or sink, due to injection or extraction of either heat or mass.

It should be notices that λ does not coincide with the latent heat, namely the difference between the density of internal energy of the two phases at constant temperature and concentration.

Let us assume that the system under consideration occupies a domain $\Omega \subset \mathbb{R}^3$ for a time interval $]0, T[$. In the absence of chemical reactions and mechanical actions, the principles of energy and mass conservation yield

$$D_t u = -\nabla \cdot \boldsymbol{j}_u + h \quad \text{in } Q := \Omega \times]0, T[, \tag{16}$$

$$D_t c = -\nabla \cdot \boldsymbol{j}_c \quad \text{in } Q. \tag{17}$$

We shall assume that the dependence of the internal energy density u on the primal state variables s, c, χ is prescribed; that is, $u = \widehat{u}(s, c, \chi)$. By this "hat notation" we shall distinguish between the physical field, $u = u(x, t)$, and the function that represents how it depends on other variables, $u = \widehat{u}(s, c, \chi)$.

Along with a standard practice of the theory of convex analysis, we then extend $\widehat{u}$ with value $+\infty$ for $(c, \chi) \notin [0, 1] \times [-1, 1]$. We may thus assume this function to be differentiable for any $(c, \chi) \in]0, 1[\times]-1, 1[$, but of course not on the boundary

of this rectangle. The (multivalued) partial subdifferentials[1] $\partial_c \widehat{u}$ and $\partial_\chi \widehat{u}$, are then reduced to the partial derivatives $\partial \widehat{u}/\partial c$ and $\partial \widehat{u}/\partial \chi$ for any $(c, \chi) \in]0, 1[\times]-1, 1[$.

Classical thermodynamics prescribes that

$$\tau = \frac{\partial \widehat{u}}{\partial s}(s, c, \chi), \qquad \mu = \frac{\partial \widehat{u}}{\partial c}(s, c, \chi), \qquad \lambda = \frac{\partial \widehat{u}}{\partial \chi}(s, c, \chi),$$

provided that the function $\widehat{u}$ is differentiable. Thus[2]

$$\begin{aligned} &u = \widehat{u}(s, c, \chi), \\ &du = \tau ds + \mu dc + \lambda d\chi \quad \forall (s, c, \chi) \in \mathrm{Dom}(\widehat{u})^0, \end{aligned} \tag{18}$$

or more generally, without assuming the differentiability of the function $\widehat{u}$,

$$\begin{aligned} &\tau \in \partial_s \widehat{u}(s, c, \chi), \qquad \mu \in \partial_c \widehat{u}(s, c, \chi), \qquad \lambda \in \partial_\chi \widehat{u}(s, c, \chi) \\ &\forall (u, c, \chi) \in \mathrm{Dom}(\widehat{u}). \end{aligned} \tag{19}$$

As $\tau > 0$, the constitutive relation $u = \widehat{u}(s, c, \chi)$ may also be made explicit with respect to s. This yields the *Gibbs-type formula*

$$\begin{aligned} &s = \widehat{s}(u, c, \chi), \\ &ds = \frac{1}{\tau} du - \frac{\mu}{\tau} dc - \frac{\lambda}{\tau} d\chi \quad \forall (u, c, \chi) \in \mathrm{Dom}(\widehat{s})^0, \end{aligned} \tag{20}$$

with $\widehat{s}$ a concave function of u, for any fixed c, χ. More generally, without assuming the differentiability of the function $\widehat{s}$, we have

$$\begin{aligned} &\frac{1}{\tau} \in \partial_u \widehat{s}(u, c, \chi), \qquad -\frac{\mu}{\tau} \in \partial_c \widehat{s}(u, c, \chi), \qquad -\frac{\lambda}{\tau} \in \partial_\chi \widehat{s}(u, c, \chi) \\ &\forall (u, c, \chi) \in \mathrm{Dom}(\widehat{s}). \end{aligned} \tag{21}$$

The relations (19)–(21) are prescribed at equilibrium. A basic postulate of *nonequilibrium thermodynamics,* assumes that (18) (and the equivalent (20)) also apply to systems that are *not too far* from equilibrium. Out of lack of a better model, here we extrapolate these relations even to the glassy phase. Actually, the limits of validity of the whole theory strongly depend on those of the Gibbs-type formula (20) and of the other constitutive relations that we introduce ahead.

[1] By ∂f we denote the subdifferential (in the sense of convex analysis) of a function f of a single variable. On the other hand, by $\partial_u f$, $\partial_v f$, ... we denote the partial subdifferentials of a function f of a two or more variables $u, v, \ldots$

[2] By $\mathrm{Dom}(\widehat{u})$ we denote the domain of $\widehat{u}$, namely the set where this function is finite. By A^0 we denote the interior of any set A.

Entropy Balance and Clausius-Duhem Inequality Let us set

$$\boldsymbol{j}_s := \frac{\boldsymbol{j}_u - \mu \boldsymbol{j}_c}{\tau} : \text{ entropy flux (per unit surface)}, \tag{22}$$

$$\pi := \boldsymbol{j}_u \cdot \nabla \frac{1}{\tau} - \boldsymbol{j}_c \cdot \nabla \frac{\mu}{\tau} - \frac{\lambda}{\tau} D_t \chi : \text{ entropy production rate (per unit volume)}. \tag{23}$$

Denoting by $\boldsymbol{q}$ the heat flux we have $\boldsymbol{j}_u = \boldsymbol{q} + \mu \boldsymbol{j}_c$, so that the two latter definitions also read

$$\boldsymbol{j}_s = \frac{\boldsymbol{q}}{\tau}, \qquad \pi = \boldsymbol{q} \cdot \nabla \frac{1}{\tau} - \frac{\boldsymbol{j}_c}{\tau} \cdot \nabla \mu - \frac{\lambda}{\tau} D_t \chi. \tag{24}$$

Multiplying (16) by $1/\tau$ and (17) by $-\mu/\tau$, by (21)–(23) we get the *entropy balance* equation

$$\begin{aligned} D_t s &= \frac{1}{\tau} D_t u - \frac{\mu}{\tau} D_t c - \frac{\lambda}{\tau} D_t \chi \\ &= -\frac{1}{\tau} \nabla \cdot \boldsymbol{j}_u + \frac{h}{\tau} + \frac{\mu}{\tau} \nabla \cdot \boldsymbol{j}_c - \frac{\lambda}{\tau} D_t \chi \\ &= -\nabla \cdot \frac{\boldsymbol{j}_u - \mu \boldsymbol{j}_c}{\tau} + \boldsymbol{j}_u \cdot \nabla \frac{1}{\tau} - \boldsymbol{j}_c \cdot \nabla \frac{\mu}{\tau} - \frac{\lambda}{\tau} D_t \chi + \frac{h}{\tau} \\ &= -\nabla \cdot \boldsymbol{j}_s + \pi + \frac{h}{\tau} \quad \text{in } Q. \end{aligned} \tag{25}$$

The quantity h/τ is the rate at which entropy is either provided to the system or extracted from it by an external source or sink of heat.

According to the local formulation of the second principle of thermodynamics (see e.g. [15, 22, 23, 38, 53]), the entropy production rate is pointwise nonnegative, and vanishes only at equilibrium. This is tantamount to the *Clausius-Duhem inequality*:

$$\begin{aligned} &\pi \geq 0 \quad \text{for any process, and} \\ &\pi = 0 \quad \text{if and only if } \nabla \boldsymbol{\tau} = \nabla \boldsymbol{\mu} = \boldsymbol{0}. \end{aligned} \tag{26}$$

Moreover, $\pi = 0$ ($\pi > 0$, resp.) corresponds to a reversible (irreversible, resp.) process.

Phenomenological Laws and Phase Relaxation The next step consists in formulating constitutive laws consistent with (26). First we introduce some further definitions:

$$z := \left(\frac{1}{\tau}, -\frac{\mu}{\tau}, -\frac{\lambda}{\tau} \right) \big(\in \mathrm{Dom}(s^*) \big) : \quad \text{dual state variables}, \tag{27}$$

$$\boldsymbol{G} := \left(\nabla \frac{1}{\tau}, -\nabla \frac{\mu}{\tau}, -\frac{\lambda}{\tau} \right) : \quad \text{generalized forces}, \tag{28}$$

$$\boldsymbol{J} := (\boldsymbol{j}_u, \boldsymbol{j}_c, D_t\chi): \quad \text{generalized fluxes.} \tag{29}$$

Along the lines of the theory of nonequilibrium thermodynamics, we assume that the generalized fluxes are functions of the dual state variables and of the generalized forces, via constitutive relations of the form

$$\boldsymbol{J} = \boldsymbol{F}(z, \boldsymbol{G}) \quad \forall z \in \mathrm{Dom}\big(s^*\big)\big(\subset \mathbb{R}^+ \times \mathbb{R}^2\big). \tag{30}$$

These relations must be consistent with the second principle, cf. (26). The mapping $\boldsymbol{F}$ must thus be positive-definite with respect to $\boldsymbol{G}$. Close to thermodynamic equilibrium, namely, for small generalized forces, one may also assume that this dependence is linear. Notice that the first two components of $\boldsymbol{J}$ and $\boldsymbol{G}$ are vectors, and the third ones are scalars. The linearized relations then uncouple, because of the *Curie principle*: "generalized forces cannot have more elements of symmetry than the generalized fluxes that they produce". Thus, denoting by $I_{[-1,1]}$ the indicator function of the interval $[-1, 1]$,

$$\begin{pmatrix} \boldsymbol{j}_u \\ \boldsymbol{j}_c \end{pmatrix} = \mathcal{L}(z) \cdot \begin{pmatrix} \nabla\frac{1}{\tau} \\ -\nabla\frac{\mu}{\tau} \end{pmatrix} \quad \text{in } Q, \tag{31}$$

$$D_t\chi + \partial I_{[-1,1]}(\chi) \ni -\ell(z)\frac{\lambda}{\tau} \quad \text{in } Q. \tag{32}$$

In (31) the dot denotes the rows-by-columns product of a tensor of $(\mathbb{R}^3)^{2\times 2}$ by a vector of $(\mathbb{R}^3)^2$. Notice that $\partial I_{[-1,1]}(-1) =]-\infty, 0]$, $\partial I_{[-1,1]}(y) = \{0\}$ for any $y \in]-1, 1[$, $\partial I_{[-1,1]}(1) = [0, +\infty[$. The linearized constitutive relations (31) are often called *phenomenological laws*; (32) is a relaxation-type dynamics. Consistently with (26), for any z the tensor $\mathcal{L}(z)$ is assumed to be positive-definite, and $\ell(z) > 0$ (whereas of course λ may change sign). A fundamental result of nonequilibrium thermodynamics due to Onsager states that the tensor $\mathcal{L}(z)$ is symmetric:

$$\mathcal{L} = \begin{pmatrix} \mathcal{L}_{11} & \mathcal{L}_{12} \\ \\ \mathcal{L}_{21} & \mathcal{L}_{22} \end{pmatrix}, \qquad \mathcal{L}_{12}(z) = \mathcal{L}_{21}(z)\big(\in \mathbb{R}^3\big) \quad \forall z \in \mathrm{Dom}\big(s^*\big). \tag{33}$$

The tensor $\mathcal{L}_{12}(z)$ accounts for mass flow induced by a temperature gradient, (*Soret effect*), whereas $\mathcal{L}_{21}(z)$ accounts for the dual phenomenon of heat flow induced by a gradient of chemical potential (*Dufour effect*).

Potential Structure of the Phenomenological Laws Let us set

$$\begin{aligned} &\boldsymbol{g} := \left(\nabla\frac{1}{\tau}, -\nabla\frac{\mu}{\tau}\right), \\ &\Phi(z, \boldsymbol{\xi}, r) := \frac{1}{2}\boldsymbol{\xi}^* \cdot \mathcal{L}(z) \cdot \boldsymbol{\xi} + \frac{1}{2}\ell(z)r^2 \\ &\forall z \in \mathrm{Dom}\big(s^*\big),\ \forall \boldsymbol{\xi} \in \big(\mathbb{R}^3\big)^2,\ \forall r \in \mathbb{R} \end{aligned} \tag{34}$$

(here by $\boldsymbol{\xi}^*$ we denote the transposed of the vector $\boldsymbol{\xi}$). Because of the Onsager relations (33), the (linearized) laws (31) and (32) may then be represented in gradient form:

$$\boldsymbol{J} \in \partial_2 \Phi(z, \boldsymbol{G}) \quad \forall z \in \mathrm{Dom}\big(s^*\big), \tag{35}$$

where by ∂_2 we denote the subdifferential with respect to the second argument, $\boldsymbol{G}$.

This representation may be extended to the nonlinear case. More specifically, within a certain range of variation of the variables, one may thus assume that the nonlinear constitutive relations (30) also have a potential structure of the form

$$\begin{aligned} &\boldsymbol{J} \in \partial_2 \Phi(z, \boldsymbol{G}) \quad \text{with} \\ &\Phi(z, \cdot) \text{ convex mapping } \big(\mathbb{R}^3\big)^2 \to \big(\mathbb{R}^3\big)^2, \quad \forall z \in \mathrm{Dom}\big(s^*\big). \end{aligned} \tag{36}$$

Even further from equilibrium, one may deal with (30) dropping the assumption of existence of a potential. As we saw, this is the case for glass formation.

In conclusion, we have represented processes in two-phase composites by the quasilinear parabolic system (16), (17), (21), coupled with phenomenological laws either of the general form (30) or (assuming existence of a potential) of the form (36).

4 Weak Formulation and Existence Theorems

In this section we formulate an initial- and boundary-value problem for phase relaxation in two-phase binary composites, and deal with existence of a weak solution.

We assume that Ω is a bounded Lipschitz domain of $\mathbb{R}^3$, denote its boundary by Γ, fix two subsets Γ_{Di} $(i = 1, 2)$ of Γ having positive bidimensional Hausdorff measure, and set $Q := \Omega \times]0, T[$ as above. We define the weighted Hilbert spaces

$$V_i := \big\{ v \in H^1(\Omega) : \gamma_0 v = 0 \text{ on } \Gamma_{Di} \big\} \quad (i = 1, 2), \tag{37}$$

and denote by $\langle \cdot, \cdot \rangle$ the pairing between V_i and the dual space V_i' for $i = 1, 2$. By identifying the space $L^2(\Omega)$ with its dual and the latter with a subspace of V_i', we get two Hilbert triplets:

$$V_i \subset L^2(\Omega) = L^2(\Omega)' \subset V_i', \quad \text{with dense and compact injections } (i = 1, 2). \tag{38}$$

We assume that

$$\begin{aligned} &\varphi : \mathbb{R} \times [0, 1] \times [-1, 1] \to \mathbb{R} \cup \{+\infty\} \\ &\text{is proper, convex and lower semicontinuous,} \end{aligned} \tag{39}$$

$$\begin{aligned} &\gamma : \mathbb{R}^2 \times \big(\mathbb{R}^3\big)^2 \to \big(\mathbb{R}^3\big)^2, \\ &\gamma(\cdot, \cdot, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \quad \text{is continuous } \forall (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \in \big(\mathbb{R}^3\big)^2, \\ &\gamma(\theta, \omega, \cdot, \cdot) \quad \text{is monotone } \forall (\theta, \omega) \in \mathbb{R}^2, \end{aligned} \tag{40}$$

$$\rho : \mathbb{R}^2 \to \mathbb{R} \quad \text{is Lipschitz continuous.} \tag{41}$$

We then fix any

$$\begin{aligned} &u^0, c^0, \chi^0 \in L^2(\Omega) \quad \text{such that } \varphi\big(u^0, c^0, \chi^0\big) < +\infty \text{ a.e. in } \Omega, \\ &f_i \in L^2\big(0, T; V_i'\big) \quad (i = 1, 2), \end{aligned} \tag{42}$$

and introduce a weak formulation.

Problem 4.1 Find $u, c, \chi, \theta, \omega, r, \boldsymbol{j}_u, \boldsymbol{j}_c$ with the regularity

$$\begin{aligned} &u \in L^2(Q) \cap H^1\big(0, T; V_1'\big), \qquad c \in L^2(Q) \cap H^1\big(0, T; V_2'\big), \\ &\chi \in H^1\big(0, T; L^2(\Omega)\big), \end{aligned} \tag{43}$$

$$\theta \in L^2(0, T; V_1), \qquad \omega \in L^2(0, T; V_2), \qquad r \in L^2(Q), \qquad \boldsymbol{j}_u, \boldsymbol{j}_c \in L^2(Q)^3, \tag{44}$$

that fulfill the constitutive relations

$$(\theta, \omega, r) \in \partial\varphi(u, c, \chi) \quad \text{a.e. in } Q, \tag{45}$$

$$(\boldsymbol{j}_u, \boldsymbol{j}_c) = -\gamma(\theta, \omega, \nabla\theta, \nabla\omega) \quad \text{a.e. in } Q, \tag{46}$$

as well as the equations

$$D_t u + \nabla \cdot \boldsymbol{j}_u = f_1 \quad \text{in } V_1', \text{ a.e. in }]0, T[, \tag{47}$$

$$D_t c + \nabla \cdot \boldsymbol{j}_c = f_2 \quad \text{in } V_2', \text{ a.e. in }]0, T[, \tag{48}$$

$$D_t \chi + r = \rho(\theta, \omega) \quad \text{a.e. in } Q, \tag{49}$$

and the initial conditions

$$u(\cdot, 0) = u^0 \quad \text{in } V_1', \qquad c(\cdot, 0) = c^0 \quad \text{in } V_2', \qquad \chi(\cdot, 0) = \chi^0 \quad \text{a.e. in } \Omega. \tag{50}$$

It is well known that by a suitable selection of the functionals f_1 and f_2, (47) and (48) respectively account for the energy balance (16) and for the mass diffusion equation (17), each one coupled with the homogeneous Dirichlet condition on Γ_{Di} and with a Neumann condition on $\Gamma \setminus \Gamma_{Di}$, for $i = 1, 2$.

Equation (49) extends (14) to a heterogeneous system.

Theorem 4.1 (Existence of a Weak Solution—I) *Assume that* (39)–(42) *are satisfied, and that*

$$\begin{aligned} &\varphi^* : \mathbb{R}^2 \times [-1, 1] \to \mathbb{R} \cup \{+\infty\} \quad \textit{is of the form} \\ &\varphi^*(\theta, \omega, \chi) = \psi_1(\theta, \omega) + \psi_2(\theta, \omega, \chi) \quad \forall (\theta, \omega, \chi), \\ &\textit{where:} \quad \psi_1 \textit{ is strictly convex and lower semicontinuous}, \\ &\psi_2(\cdot, \cdot, \chi) \textit{ is convex and lower semicontinuous } \forall \chi \in [-1, 1], \end{aligned} \tag{51}$$

$$\begin{aligned}&\exists c_1, c_2 > 0: \quad \forall (u, c, \chi) \in \mathrm{Dom}(\varphi), \forall (\theta, \omega, r) \in \partial\varphi(u, c, \chi),\\ &|\theta| \le c_1 |u| + c_2,\end{aligned} \tag{52}$$

$$\exists a_1, a_2 > 0: \quad \forall (u, c, \chi) \in \mathrm{Dom}(\varphi), \varphi(u, c, \chi) \ge a_1 |u|^2 - a_2, \tag{53}$$

$$\begin{aligned}&\gamma = \partial\Phi, \quad \textit{with } \Phi : \mathbb{R}^2 \times \left(\mathbb{R}^3\right)^2 \to \mathbb{R},\\ &\Phi(\cdot, \cdot, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \quad \textit{is continuous} \quad \forall (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \in \left(\mathbb{R}^3\right)^2,\\ &\Phi(\theta, \omega, \cdot, \cdot) \quad \textit{is convex} \quad \forall (\theta, \omega) \in \mathbb{R}^2\end{aligned} \tag{54}$$

(*in $\partial\Phi$ the subdifferential operation is applied to the two latter arguments*),

$$\begin{aligned}&\exists a_3, \dots, a_6 > 0: \quad \forall (\theta, \omega, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \in \mathbb{R}^2 \times \left(\mathbb{R}^3\right)^2,\\ &a_3\left(|\boldsymbol{\xi}_1|^2 + |\boldsymbol{\xi}_2|^2\right) - a_4 \le \Phi(\theta, \omega, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \le a_5\left(|\boldsymbol{\xi}_1|^2 + |\boldsymbol{\xi}_2|^2\right) + a_6,\end{aligned} \tag{55}$$

$$\exists a_7, a_8 > 0: \quad \forall (\theta, \omega) \in \mathbb{R}^2 \quad \left|\rho(\theta, \omega)\right| \le a_7 |\theta| + a_8. \tag{56}$$

Then Problem 4.1 has a solution such that moreover $u, c \in L^\infty(0, T; L^2(\Omega))$.

The assumptions of this theorem are consistent with the model that we illustrated in the previous section. Next we state another existence result.

Theorem 4.2 (Existence of a Weak Solution—II, [67]) *Assume that the assumptions* (39)–(42) *are satisfied, as well as the conditions* (51), (52), (56) *and*

$$\begin{aligned}&\exists C > 0: \quad \forall (u_i, c_i, \chi_i) \in \mathrm{Dom}(\varphi), \quad \forall (\theta_i, \omega_i, r_i) \in \partial\varphi(u_i, c_i, \chi_i) \quad (i = 1, 2),\\ &(u_1 - u_2)(\theta_1 - \theta_2) + (c_1 - c_2)(\omega_1 - \omega_2) + (\chi_1 - \chi_2)(r_1 - r_2)\\ &\quad \ge C\left(|\theta_1 - \theta_2|^2 + |\omega_1 - \omega_2|^2\right),\end{aligned} \tag{57}$$

$$\begin{aligned}&\exists a_9 > 0: \quad \forall (\theta, \omega) \in \mathbb{R}^2, \quad \forall (\boldsymbol{\xi}_{1i}, \boldsymbol{\xi}_{2i}) \in \left(\mathbb{R}^3\right)^2 \quad (i = 1, 2),\\ &\left[\gamma(\theta, \omega, \boldsymbol{\xi}_{11}, \boldsymbol{\xi}_{21}) - \gamma(\theta, \omega, \boldsymbol{\xi}_{12}, \boldsymbol{\xi}_{22})\right] \cdot (\boldsymbol{\xi}_{11} - \boldsymbol{\xi}_{12}, \boldsymbol{\xi}_{21} - \boldsymbol{\xi}_{22})\\ &\quad \ge a_9\left(|\boldsymbol{\xi}_{11} - \boldsymbol{\xi}_{12}|^2 + |\boldsymbol{\xi}_{21} - \boldsymbol{\xi}_{22}|^2\right),\end{aligned} \tag{58}$$

$$\begin{aligned}&\exists a_{10}, a_{11} > 0: \quad \forall (\theta, \omega) \in \mathbb{R}^2, \quad \forall (\boldsymbol{\xi}_{1i}, \boldsymbol{\xi}_{2i}) \in \left(\mathbb{R}^3\right)^2 \quad (i = 1, 2),\\ &\left|\gamma(\theta, \omega, \boldsymbol{\xi}_{11}, \boldsymbol{\xi}_{21}) - \gamma(\theta, \omega, \boldsymbol{\xi}_{12}, \boldsymbol{\xi}_{22})\right|\\ &\quad \le a_{10}\left(|\boldsymbol{\xi}_{11} - \boldsymbol{\xi}_{12}| + |\boldsymbol{\xi}_{21} - \boldsymbol{\xi}_{22}|\right) + a_{11}.\end{aligned} \tag{59}$$

Then Problem 4.1 has a solution such that moreover

$$u, c \in L^\infty\left(0, T; L^2(\Omega)\right), \qquad \theta, \omega \in H^s\left(0, T; L^2(\Omega)\right) \quad \forall s < 1/2. \tag{60}$$

The assumptions of this theorem are also consistent with the previous model. Here we just point out the main lines of the argument, which differs from that of Theorem 4.1 (see [67]) for an additional a priori estimate.

(i) First a priori estimate. Next we display the basic entropy estimate, which is also used in [67]. Let us first extend the fields u, c, χ to $t < 0$ by setting $u(\cdot, t) = u^0$, $c(\cdot, t) = c^0$, $\chi(\cdot, t) = \chi^0$ a.e. in Ω for any $t < 0$. For any $m \in \mathbb{N}$, let us also introduce the time step $h = T/m$, and define the time incremental operator δ_h by setting $\delta_h v(t) := v(t + h) - v(t)$ for any function v of t. We may then consider the approximation scheme

$$\delta_h u + h\nabla \cdot \boldsymbol{j}_u = hf_1 \quad \text{in } V_1', \text{ a.e. in }]0, T[, \tag{61}$$

$$\delta_h c + h\nabla \cdot \boldsymbol{j}_c = hf_2 \quad \text{in } V_2', \text{ a.e. in }]0, T[, \tag{62}$$

$$\delta_h \chi + hr = h\rho(\theta, \omega) \quad \text{a.e. in } Q, \tag{63}$$

and couple this system with the constitutive relations (45) and (46). It is not difficult to check that this problem has a solution (that we label by the index h) with the following regularity:

$$\begin{aligned} &u_h, c_h, \chi_h, r_h \in L^2(Q), \\ &\theta_h \in L^2(0, T; V_1), \qquad \omega_h \in L^2(0, T; V_2), \qquad (\boldsymbol{j}_u)_h, (\boldsymbol{j}_c)_h \in L^2(Q)^3. \end{aligned} \tag{64}$$

Via a standard procedure, the following uniform estimates are derived by multiplying Eqs. (61)–(63) respectively by θ_h, ω_h, r_h, and then integrating over $\Omega \times]0, t[$ for any $t \in]0, T]$:

$$\|u_h\|_{L^\infty(0,T;L^2(\Omega))\cap H^1(0,T;V_1')}, \|c_h\|_{L^\infty(0,T;L^2(\Omega))\cap H^1(0,T;V_2')} \le C_1, \tag{65}$$

$$\|\theta_h\|_{L^2(0,T;V_1)}, \|\omega_h\|_{L^2(0,T;V_2)}, \|\chi_h\|_{H^1(0,T;L^2(\Omega))} \le C_2, \tag{66}$$

$$\left\|(\boldsymbol{j}_u)_h\right\|_{L^2(Q)^3}, \left\|(\boldsymbol{j}_c)_h\right\|_{L^2(Q)^3} \le C_3. \tag{67}$$

(By $C_1, C_2, \ldots$ we denote constants independent of h.) See Sect. 7 of [67] for details.

(ii) Second a priori estimate. For any $k \in]0, T[$, further a priori estimates may be derived by multiplying the approximate equations (61)–(63) respectively by $\delta_k\theta_h, \delta_k\omega_h, \delta_k r_h$, and then integrating over $\Omega \times]k, T[$. (The reader will notice that we are not dividing these equations by k, and that two indices occur: h and k.) This

yields

$$k^{-1}\int_k^T dt\int_\Omega \big[(\delta_k u_h)(\delta_k\theta_h)+(\delta_k c_h)(\delta_k\omega_h)+(\delta_k\chi_h)(\delta_k r_h)\big]dx$$
$$\le -\int_k^T dt\int_\Omega \big[(\boldsymbol{j}_u)_h\cdot\nabla\delta_k\theta_h+(\boldsymbol{j}_c)_h\cdot\nabla\delta_k\omega_h+r_h(\delta_k r_h)\big]dx$$
$$+\int_k^T dt\int_\Omega \big[f_{1h}\delta_k\theta_h+f_{2h}\delta_k\omega_h+\rho(\theta_h,\omega_h)\delta_k r_h\big]dx\quad\forall t\in\,]0,T]. \tag{68}$$

By (57), (59) and by the previous a priori estimates, it is easily checked that the right-hand side of this inequality is uniformly bounded with respect to both h and k. Hence by (57)

$$k^{-1}\int_k^T dt\int_\Omega \big(|\delta_k\theta_h|^2+|\delta_k\omega_h|^2\big)dx\le C_5. \tag{69}$$

By Lemma 4.1 below, we then conclude that

$$\begin{aligned}&\text{the sequences } \{\theta_h\} \text{ and } \{\omega_h\}\\ &\text{are bounded in } H^s\big(0,T;L^2(\Omega)\big) \text{ for any } s<1/2.\end{aligned} \tag{70}$$

(iii) Limit procedure. The estimates (65)–(67), (70) entail that there exist (u,c,χ), (θ,ω,r) and $(\boldsymbol{j}_u,\boldsymbol{j}_c)$ as in (43) and (44) such that, up to extracting subsequences,[3]

$$u_h\overset{*}{\rightharpoonup}u\quad\text{in } L^\infty\big(0,T;L^2(\Omega)\big)\cap H^1\big(0,T;V_1'\big), \tag{71}$$
$$c_h\overset{*}{\rightharpoonup}c\quad\text{in } L^\infty\big(0,T;L^2(\Omega)\big)\cap H^1\big(0,T;V_2'\big), \tag{72}$$
$$\chi_h\overset{*}{\rightharpoonup}\chi\quad\text{in } L^\infty(Q)\cap H^1\big(0,T;L^2(\Omega)\big), \tag{73}$$
$$\theta_h\rightharpoonup\theta\quad\text{in } L^2(0,T;V_1)\cap H^s\big(0,T;L^2(\Omega)\big)\quad\forall s<1/2, \tag{74}$$
$$\omega_h\rightharpoonup\omega\quad\text{in } L^2(0,T;V_2)\cap H^s\big(0,T;L^2(\Omega)\big)\quad\forall s<1/2, \tag{75}$$
$$r_h\rightharpoonup r\quad\text{in } L^2(Q), \tag{76}$$
$$(\boldsymbol{j}_u)_h\rightharpoonup\boldsymbol{j}_u\quad\text{in } L^2(Q)^3, \tag{77}$$
$$(\boldsymbol{j}_c)_h\rightharpoonup\boldsymbol{j}_c\quad\text{in } L^2(Q)^3. \tag{78}$$

Equations (47)–(49) then follow by passing to the limit in (61)–(63). As by (74) and (75),

$$\theta_h\to\theta,\qquad\omega_h\to\omega\quad\text{in } L^2(Q), \tag{79}$$

[3] We denote the strong, weak, and weak star convergence respectively by $\to$, $\rightharpoonup$, $\overset{*}{\rightharpoonup}$.

the passage to the limit in the nonlinear terms may then be accomplished along the lines of Sect. 7 of [67].

Lemma 4.1 *Let $\{u_n\}$ be a bounded sequence of functions of $L^2(0,T)$. If*

$$\int_k^T \frac{|u_n(t)-u_n(t-k)|^2}{k}dt \le C_6: \quad \text{Constant independent of } n,k, \tag{80}$$

then the sequence $\{u_n\}$ is uniformly bounded in $H^s(0,T)$ for any $s<1/2$.

Proof For any $s\in\,]0,1/2[$ we have

$$\begin{aligned}
\|u_n\|^2_{H^s(0,T)} &= \|u_n\|^2_{L^2(Q)} + \iint_{]0,T[^2} \frac{|u_n(t')-u_n(t'')|^2}{|t'-t''|^{1+2s}}dt'dt'' \\
&= \|u_n\|^2_{L^2(Q)} + 2\int_0^T dt \int_0^t \frac{|u_n(t)-u_n(t-k)|^2}{k^{1+2s}}dk \\
&= \|u_n\|^2_{L^2(Q)} + 2\int_0^T k^{-2s}dk \int_k^T \frac{|u_n(t)-u_n(t-k)|^2}{k}dt \\
&= \|u_n\|^2_{L^2(Q)} + 2C_6\int_0^T k^{-2s}dk \\
&\overset{(80)}{\le} \text{Constant(independent of } n). \qquad \square
\end{aligned} \tag{81}$$

Remark 4.1 Theorems 4.1 and 4.2 essentially differ in the derivation of (79). More specifically, we just derived (79) by compactness, because of the a priori estimates (70). On the other hand, in the argument of Theorem 4.1 (see [67]), (79) stems from *compactness by strict convexity* (in the sense of Chap. X of [64]).

5 Fitzpatrick's Theory and Γ-Convergence

The Fitzpatrick Theorem Let V be a real Banach space, and $\alpha: V\to\mathcal{P}(V')$ a proper (multivalued) operator. In 1988 Fitzpatrick defined the convex and lower semicontinuous function

$$\begin{aligned}
f_\alpha(v,v^*) &:= \langle v^*, v\rangle + \sup\{\langle v^*-v_0^*, v_0-v\rangle : v_0^*\in\alpha(v_0)\} \\
&= \sup\{\langle v^*, v_0\rangle - \langle v_0^*, v_0-v\rangle : v_0^*\in\alpha(v_0)\} \quad \forall (v,v^*)\in V\times V',
\end{aligned} \tag{82}$$

and proved the following result.

Theorem 5.3 [30] *If $\alpha: V\to\mathcal{P}(V')$ is maximal monotone, then*

$$f_\alpha(v,v^*) \ge \langle v^*, v\rangle \quad \forall (v,v^*)\in V\times V', \tag{83}$$

$$f_\alpha(v,v^*) = \langle v^*, v\rangle \quad \Leftrightarrow \quad v^*\in\alpha(v). \tag{84}$$

Along these lines, nowadays one says that a function $f : V \times V' \to \mathbb{R} \cup \{+\infty\}$ (variationally) *represents* the operator α whenever f is convex and lower semicontinuous and fulfills the system (83), (84). We shall denote by $\mathcal{F}(V)$ the class of these *representative* functions. *Representable* operators are necessarily monotone, but need not be maximal monotone; e.g., the nonmaximal monotone operator with graph $A = \{(0,0)\}$ is represented by $f_1 = I_{\{(0,0)\}}$. On the other hand, not all monotone operators are representable; e.g., the null mapping restricted to $V \setminus \{0\}$ is not representable.

For any convex and lower semicontinuous function $\varphi : V \to \mathbb{R} \cup \{+\infty\}$, the *Fenchel function*

$$F\big(v, v^*\big) := \varphi(v) + \varphi^*\big(v^*\big) \quad \forall \big(v, v^*\big) \in V \times V' \tag{85}$$

fulfills the system (83) and (84), because of the classical *Fenchel inequality* of convex analysis (see e.g. [26, 37, 54]). Thus F represents the operator $\partial\varphi$. Other examples may be found e.g. in [69–72].

Γ-Compactness and Stability of Representative Functions Henceforth we shall assume that V' is separable, and introduce a nonlinear notion of convergence, which seems to be appropriate in this framework. For any sequence $\{(v_n, v_n^*)\}$ in $V \times V'$, let us set

$$\begin{aligned} &\big(v_n, v_n^*\big) \underset{\widetilde{\pi}}{\to} \big(v, v^*\big) \quad \text{in } V \times V' \quad \Leftrightarrow \\ &v_n \rightharpoonup v \quad \text{in } V, \quad v_n^* \overset{*}{\rightharpoonup} v^* \quad \text{in } V', \qquad \big\langle v_n^*, v_n \big\rangle \to \big\langle v^*, v \big\rangle, \end{aligned} \tag{86}$$

and similarly define the convergence of $\widetilde{\pi}$-nets. (We use the term "$\widetilde{\pi}$-convergence" since we denote by π the duality pairing between V and V', i.e., $\pi(v, v^*) := \langle v^*, v \rangle$.)

Under the assumption of equi-coerciveness, the Γ-compactness with respect to the product between the weak and weak star topologies of V and V' stems from the classical theory, see e.g. [20]. The next statement provides the Γ-compactness with respect to the $\widetilde{\pi}$-topology, which is especially relevant in the analysis of representative functions.

Theorem 5.4 [72] *Let a sequence $\{\psi_n\}$ in $\mathcal{F}(V)$ be equi-coercive in the sense that*

$$\forall C \in \mathbb{R}, \sup_{n \in \mathbb{N}} \big\{ \|v\|_V + \big\|v^*\big\|_{V'} : \big(v, v^*\big) \in V \times V', \psi_n\big(v, v^*\big) \le C \big\} < +\infty. \tag{87}$$

Then, up to extracting a subsequence, ψ_n sequentially Γ-converges to some function ψ with respect to the topology $\widetilde{\pi}$. This entails that $\psi \in \mathcal{F}(V)$.

Moreover, denoting by α_n (α, resp.) the operator $V \to \mathcal{P}(V')$ that is represented by ψ_n (ψ, resp.), for any sequence $\{(v_n, v_n^*)\}$ in $V \times V'$,

$$\begin{aligned} &v_n^* \in \alpha_n(v_n) \quad \forall n, \quad \big(v_n, v_n^*\big) \underset{\widetilde{\pi}}{\to} \big(v, v^*\big) \\ &\Rightarrow \quad v^* \in \alpha(v), \quad \psi_n\big(v_n, v_n^*\big) \to \psi\big(v, v^*\big). \end{aligned} \tag{88}$$

Representation in Spaces of Time-Dependent Functions Let us fix any $T > 0$, any $p \in]1, +\infty[$ and set $\mathcal{V} := L^p(0, T; V)$. Let us define the convergence $\widetilde{\pi}$ in $\mathcal{V} \times \mathcal{V}'$ as in (86), by replacing the space V by $\mathcal{V}$ and the associated duality pairing $\langle v^*, v\rangle$ by $\langle\langle v^*, v\rangle\rangle := \int_0^T \langle v^*(t), v(t)\rangle dt$ for any $(v, v^*) \in \mathcal{V} \times \mathcal{V}'$. Theorem 5.4 takes over to time-dependent operators and to their time-integrated representative functions, simply by replacing the space V by $\mathcal{V}$.

It is promptly seen that, whenever a function $\psi \in \mathcal{F}(V)$ is coercive in the sense that

$$\forall C \in \mathbb{R}, \quad \sup\left\{\|v\|_V + \left\|v^*\right\|_{V'} : \left(v, v^*\right) \in V \times V', \psi\left(v, v^*\right) \leq C\right\} < +\infty, \quad (89)$$

ψ represents an operator $\alpha : V \to \mathcal{P}(V')$ if and only if the functional

$$\Psi\left(v, v^*\right) := \int_0^T \psi\left(v(t), v^*(t)\right) dt \quad \forall \left(v, v^*\right) \in \mathcal{V} \times \mathcal{V}' \quad (90)$$

(which is an element of $\mathcal{F}(\mathcal{V})$) represents the operator

$$\widehat{\alpha} : \mathcal{V} \to \mathcal{P}\left(\mathcal{V}'\right), \qquad \left[\widehat{\alpha}(v)\right](t) = \alpha\left(v(t)\right) \quad \forall v \in \mathcal{V}, \text{ for a.e. } t \in]0, T[. \quad (91)$$

Next we relate the $\widetilde{\pi}$-convergence in $V \times V'$ a.e. in $]0, T[$ with the $\widetilde{\pi}$-convergence in $\mathcal{V} \times \mathcal{V}'$.

Proposition 5.1 [72] *Let $p \in]1, +\infty[$, and $\{(v_n, v_n^*)\}$ be a bounded sequence in $W^{\varepsilon,p}(0, T; V) \times W^{\varepsilon,p'}(0, T; V')$ for some $\varepsilon > 0$. If*

$$\left(v_n, v_n^*\right) \underset{\widetilde{\pi}}{\to} \left(v, v^*\right) \quad \text{in } V \times V', \text{ a.e. in }]0, T[, \quad (92)$$

then

$$\left(v_n, v_n^*\right) \underset{\widetilde{\pi}}{\to} \left(v, v^*\right) \quad \text{in } \mathcal{V} \times \mathcal{V}'. \quad (93)$$

On the other hand, (93) does not entail (92), not even for a subsequence.

For $\varepsilon = 0$ the implication (92) $\to$ (93) fails. A counterexample is provided in [72].

Compactness and Structural Stability The representation of maximal monotone operators allows one to apply variational techniques to a large class of monotone problems; one may then prove their *structural stability* via De Giorgi's notion of Γ-convergence. Here we briefly illustrate what we mean by structural stability in a general topological set-up. Let us assume that

$\mathcal{D}$ is a set of admissible data (e.g., an initial datum and/or a source term),
$\mathcal{O}$ is a set of operators (e.g., a maximal monotone operator),
$\mathcal{S}$ is a set of admissible solutions.

We also assume that each of these sets is equipped with a topology and that a (possibly multivalued) *solution operator* $R : \mathcal{D} \times \mathcal{O} \to \mathcal{S}$ is defined. We shall say that:

(i) the class of admissible operators $\mathcal{O}$ is (sequentially) compact if

$$\text{any sequence } \{o_n\} \text{ in } \mathcal{O} \text{ accumulates at some } o \in \mathcal{O}, \tag{94}$$

(ii) the problem is *structurally stable* if the operator R is (sequentially) closed, namely, for any sequence $\{(d_n, o_n, s_n)\}$ in $\mathcal{D} \times \mathcal{O} \times \mathcal{S}$,

$$s_n \in R(d_n, o_n) \quad \forall n, \qquad (d_n, o_n, s_n) \to (d, o, s) \quad \Rightarrow \quad s \in R(d, o). \tag{95}$$

It would also be desirable that any element $s \in R(\mathcal{D}, \mathcal{O})$ may be retrieved as in (95), so that the set of the limits of solutions would coincide with that of the solutions of the asymptotic problem. In general this further property seems difficult to be proved; however, it easily follows from (95) if the limit problem has only one solution.

6 Variational Formulation and Structural Stability of Monotone Flows

In this section we apply the Fitzpatrick theory to monotone flows of the form $D_t u + \alpha(u) \ni h$, along the lines of Sects. 7 and 8 of [72].

Maximal Monotone Flows Let us assume that we are given a Gelfand triplet of (real) Hilbert spaces

$$V \subset H = H' \subset V' \quad \text{with continuous and dense injections.} \tag{96}$$

Let $\alpha : V \to \mathcal{P}(V')$ be a maximal monotone operator, $h \in L^2(0, T; V')$, and consider the Cauchy problem

$$\begin{cases} u \in X := \{v \in L^2(0, T; V) \cap H^1(0, T; V') : v(0) = 0\}, \\ D_t u + \alpha(u) \ni h \quad \text{in } V', \text{ a.e. in }]0, T[. \end{cases} \tag{97}$$

Here we embed the homogeneous initial condition into the space, so that

$$X \to L^2\big(0, T; V'\big)\big(\subset X'\big) : v \mapsto D_t v \quad \text{is monotone.}$$

The condition $u(0) = 0$ is not really restrictive, since it may be retrieved by shifting the unknown function u. More specifically, if $u^0 \in V$ then the initial condition $u(0) = u^0$ may be dealt with by replacing u by $\widetilde{u} := u - u^0$ and α by $\widetilde{\alpha} := \alpha(\cdot + u^0)$. (The case of $u^0 \in H$ is more delicate.)

We shall assume that

$$\exists a, b > 0 : \quad \forall \big(v, v^*\big) \in \text{graph}(\alpha), \quad \big\langle v^*, v \big\rangle \geq a \|v\|_V^2 - b, \tag{98}$$

$$\exists c, d > 0: \quad \forall (v, v^*) \in \text{graph}(\alpha), \quad \|v^*\|_{V'} \leq c\|v\|_V + d. \tag{99}$$

It is known that the problem (97) then has one and only one solution, see e.g. [6, 9, 74].

Variational Formulations Next we introduce several variational formulations of the problem (97). Let us define the Hilbert spaces $\mathcal{H} := L^2(0, T; H)$ and $\mathcal{V} := L^2(0, T; V)$, so that we have the Gelfand triplet

$$\mathcal{V} \subset \mathcal{H} = \mathcal{H}' \subset \mathcal{V}' \quad \text{with continuous and dense injections.} \tag{100}$$

Let the operator α be represented by a function $f \in \mathcal{F}(V)$, and set

$$F(v, v^*) := \int_0^T f(v, v^*)dt \quad \forall (v, v^*) \in \mathcal{V} \times \mathcal{V}'. \tag{101}$$

Notice that $F \in \mathcal{F}(\mathcal{V})$; actually, F represents the operator $\widehat{\alpha} : \mathcal{V} \to \mathcal{P}(\mathcal{V}')$, cf. (91).

By (84), the inclusion $(97)_2$ is equivalent to

$$f(u, h - D_t u) = \langle h - D_t u, u \rangle \quad \text{a.e. in }]0, T[.$$

For any $v \in X$ the mapping $t \mapsto \|v(t)\|_H^2$ is absolutely continuous and differentiable a.e. in $]0, T[$, and $D_t\|v(t)\|_H^2 = 2\langle D_t v, v\rangle$ a.e. The latter equation then also reads

$$f(u, h - D_t u) + \frac{1}{2} D_t \|u\|_H^2 = \langle h, u \rangle \quad \text{a.e. in }]0, T[. \tag{102}$$

As f fulfills (83), this is also equivalent to the family of equations that is obtained by time integration

$$\int_0^\tau f(u, h - D_t u)dt + \frac{1}{2}\|u(\tau)\|_H^2 = \int_0^\tau \langle h, u\rangle dt \quad \forall \tau \in]0, T[, \tag{103}$$

and also to the single equation

$$J(u, h) := F(u, h - D_t u) + \frac{1}{2}\|u(T)\|_H^2 = \int_0^T \langle h, u\rangle dt. \tag{104}$$

(Notice that $u(T) \in H$, as by a standard identification $X \subset C^0([0, T]; H)$, see e.g. Chap. I of [40].)

Let us next define the Hilbert spaces

$$\widetilde{\mathcal{H}} := \left\{ v :]0, T[\to H \text{ measurable: } \|v\|_{\widetilde{\mathcal{H}}}^2 := \int_0^T (T - t)\|v\|_H^2 dt < +\infty \right\}, \tag{105}$$

$$\widetilde{\mathcal{V}} := \left\{ v :]0, T[\to V \text{ measurable: } \|v\|_{\widetilde{\mathcal{V}}}^2 := \int_0^T (T - t)\|v\|_V^2 dt < +\infty \right\}, \tag{106}$$

and the corresponding Gelfand triplet

$$\widetilde{\mathcal{V}} \subset \widetilde{\mathcal{H}} = \widetilde{\mathcal{H}}' \subset \widetilde{\mathcal{V}}' \quad \text{with continuous and dense injections.} \tag{107}$$

Let us next set

$$\widetilde{F}\big(v, v^*\big) := \int_0^T (T-t) f\big(v, v^*\big) dt \quad \forall \big(v, v^*\big) \in \widetilde{\mathcal{V}} \times \widetilde{\mathcal{V}}', \tag{108}$$

which represents the operator

$$\widetilde{\alpha} : \widetilde{\mathcal{V}} \to \mathcal{P}\big(\widetilde{\mathcal{V}}'\big), \qquad \big[\widetilde{\alpha}(v)\big](t) = \alpha\big(v(t)\big) \quad \forall v \in \widetilde{\mathcal{V}}, \text{ for a.e. } t \in]0, T[. \tag{109}$$

Notice that the system (97) is also equivalent to the twice time-integrated equation

$$\widetilde{J}(u, h) := \widetilde{F}(u, h - D_t u) + \frac{1}{2} \int_0^T \big\|u(t)\big\|_H^2 dt = \int_0^T (T-t) \langle h, u \rangle dt. \tag{110}$$

Thus $\widetilde{J}$ represents the operator $D_t + \widetilde{\alpha}$ (in a space of time dependent functions that here we do not specify). Because of (83), (110) is equivalent to

$$\widetilde{J}(u, h) \le \int_0^T (T-t) \langle h, u \rangle dt, \tag{111}$$

and thus also to what we label as a *null-minimization problem*:

$$\widetilde{K}(u, h) := \widetilde{J}(u, h) - \int_0^T (T-t) \langle h, u \rangle dt = \inf \widetilde{K} = 0. \tag{112}$$

(The vanishing of the infimum is crucial.) It is easily seen that each of the other equivalent equations (102), (103), (104) may also be formulated as a null-minimization problem.

Conclusion as for the Variational Formulation of (97) We exhibited four equivalent variational formulations of the problem (97), namely (102), (103), (104), (110). Each of them is tantamount to a null-minimization problem.

These formulations are only formally (i.e., nonrigorously) equivalent, since they involve different function spaces. We shall refer to the equivalence between (97) and (104) as the *extended B.E.N. principle*, since it generalizes an approach that was pioneered by Brezis and Ekeland [10, 11] and by Nayroles [50] in 1976; see [66]. More specifically, the original B.E.N. principle assumes that α is cyclically monotone and selects f equal to the Fenchel function. This is here extended to any maximal monotone operator α on the basis of Fitzpatrick's Theorem 5.3.

Compactness of Representative Functions Let us now consider a $V \times V'$-equicoercive sequence $\{f_n\}$ in $\mathcal{F}(V)$, in the sense that

$$\forall C \in \mathbb{R}, \quad \sup_{n \in \mathbb{N}} \big\{ \|v\|_V + \big\|v^*\big\|_{V'} : \big(v, v^*\big) \in V \times V',\ f_n\big(v, v^*\big) \le C \big\} < +\infty, \tag{113}$$

and assume that

$$h_n \to h \quad \text{in } \mathcal{V}'. \tag{114}$$

For any n let us define the functionals F_n, $\widetilde{F}_n$ and $\widetilde{J}_n$ as above, with f_n in place of f. Next we are concerned with the Γ-compactness of these sequences in the respective function spaces with respect to the corresponding $\widetilde{\pi}$-convergence.

By (113) and the Γ-compactness Theorem 5.4, there exists f such that, up to extracting a subsequence,

$$f_n \xrightarrow{\Gamma} f \quad \text{sequentially w.r.t. the topology } \widetilde{\pi} \text{ of } V \times V'; \tag{115}$$

this entails that $f \in \mathcal{F}(V)$. Thus f represents an operator $\alpha : V \to \mathcal{P}(V')$.

By (113), the sequence $\{F_n\}$ is $\mathcal{V} \times \mathcal{V}'$-equi-coercive; there exists then $F \in \mathcal{F}$ such that, up to extracting a subsequence,

$$F_n \xrightarrow{\Gamma} F \quad \text{sequentially w.r.t. the topology } \widetilde{\pi} \text{ of } \mathcal{V} \times \mathcal{V}'; \tag{116}$$

hence $F \in \mathcal{F}(\mathcal{V})$. Let us denote by $\widehat{\alpha} : \mathcal{V} \to \mathcal{P}(\mathcal{V}')$ the operator that is represented by F.

The same applies to the sequence $\{\widetilde{F}_n\}$ in $\mathcal{F}(\widetilde{\mathcal{V}})$: by (113) this sequence is $\widetilde{\mathcal{V}} \times \widetilde{\mathcal{V}}'$-equi-coercive. There exists $\widetilde{F}$ then such that, up to extracting a subsequence,

$$\widetilde{F}_n \xrightarrow{\Gamma} \widetilde{F} \quad \text{sequentially w.r.t. the topology } \widetilde{\pi} \text{ of } \widetilde{\mathcal{V}} \times \widetilde{\mathcal{V}}', \tag{117}$$

and this entails that $\widetilde{F} \in \mathcal{F}(\widetilde{\mathcal{V}})$. Let us denote by $\widetilde{\alpha} : \mathcal{V} \to \mathcal{P}(\widetilde{\mathcal{V}}')$ the operator that is represented by $\widetilde{F}$.

We emphasize that the convergences (115)–(117) do not infer that f, F and $\widetilde{F}$ are related as in (101) and (108), and not even that F and $\widetilde{F}$ are integral functionals. Thus (91) and (109) need not hold in the limit; actually, a priori $[\widehat{\alpha}(v)](t)$ and $[\widetilde{\alpha}(v)](t)$ might also depend on $v(\tau)$ for $0 < \tau < t$, as we shall see ahead.

Besides the asymptotic behavior of the operators $\{\alpha_\varepsilon\}$, we must study that of the corresponding solutions of the monotone flow (97).

Tartar's Example The flow (97) may not be stable with respect to variations of the operator α_n, even within the class of linear maximal monotone operators that fulfill (98) and (99). We show this by means of a simple but illuminating example due to Tartar [60], who also investigated the onset of long memory in (linear) homogenization in [61] and [62], pp. 249–264. Let us assume that

$$\begin{aligned} &a_n : \Omega \to \mathbb{R} \quad \text{is measurable,} \quad \forall n, \\ &\exists c_1, c_2 > 0 : \quad \forall n, \quad c_1 \le a_n \le c_2 \quad \text{a.e. in } \Omega. \end{aligned} \tag{118}$$

The Cauchy problem

$$\begin{cases} D_t u_n + a_n(x) u_n = 0 & \text{a.e. in } \Omega, \text{ for } t > 0, \\ u(x,0) = u^0(x) & \text{a.e. in } \Omega \end{cases} \tag{119}$$

is associated with a linear and continuous semigroup in $H = L^2(\Omega)$:

$$S_n(t) : L^2(\Omega) \to L^2(\Omega) : u^0 \mapsto u_n(x,t) = e^{-a_n(x)t} u^0(x). \tag{120}$$

(Equation $(97)_2$ might also be regarded as an O.D.E. parameterized by x, but this would not be equivalent to the present approach.)

If $a_n \rightharpoonup a$ but $a_n \not\to a$ in $L^1_{loc}(\Omega)$ (that is, a_n converges weakly but not strongly), then it is easily seen that the exponential form of (120) is lost in the limit. Indeed, for any $u^0 \in L^2(\Omega)$,

$$u_n(x,t) = e^{-a_n(x)t} u^0(x) \overset{*}{\rightharpoonup} u(x,t) \neq e^{-a(x)t} u^0(x) \quad \text{in } BV\big(0,T;L^2(\Omega)\big). \tag{121}$$

We may thus conclude that the asymptotic linear operator $u^0 \mapsto u$ defines no semigroup: u does not solve any problem of the form (119), for any $a(x)$. The same conclusion may also be attained from a different viewpoint: as $a_n \rightharpoonup a$ and apparently one cannot prove more than $u_n(\cdot,t) \rightharpoonup u(\cdot,t)$ in $L^2(\Omega)$ for a.e. t, there is no way to pass to the limit in Eq. $(119)_1$.

This phenomenon may be interpreted as the onset of *long memory* from a sequence of flows with *short memory*.

Asymptotic Short Memory Let us assume that a sequence $\{\alpha_n\}$ of operators $V \to \mathcal{P}(V')$ fulfills (98) and (99) uniformly in n. For any n let u_n be the solution of (97) that corresponds to f_n and h_n; it is easily seen that this sequence is bounded in the space X (which we defined in (97)). There exists then $u \in X$ such that, up to extracting a subsequence,

$$u_n \rightharpoonup u \quad \text{in } L^2(0,T;V) \cap H^1\big(0,T;V'\big); \tag{122}$$

hence $u(0) = 0$, thus $u \in X$. Let us now assume that

$$\text{the injection } V \to H \text{ is compact,} \tag{123}$$

so that the function

$$q : X \to \mathbb{R} : v \mapsto \frac{1}{2} \int_0^T \big\| v(t) \big\|_H^2 dt \text{ is weakly continuous.} \tag{124}$$

The asymptotic mapping $\widetilde{J}$ then has the form (110).

If one were able to show that (124) entails $\|u_n(T)\|_H^2 \to \|u(T)\|_H^2$, then the form of (104) would also be preserved in the limit—but this convergence is not obvious: a priori, (122) just entails $u_n(T) \rightharpoonup u(T)$ in H. At this point this author is just able to say that Eq. (110) defines a (monotone) representable relation between u and h. This need not be representable via a short-memory monotone flow of the form (97), since f and F need not fulfill (101), as we saw for Tartar's example above.

In order to identify $\widetilde{F}(u, f - D_t u)$, some further compactness property is in order, besides (123). Let us first notice that, under further assumptions on the data,

the sequence u_n is bounded in

$$X^s := H^s(0,T;V) + H^{1+s}\big(0,T;V'\big) \quad (0 < s \le 1). \tag{125}$$

More specifically, for $s = 1$ this holds if the sequence $\{h_n\}$ is bounded in $H^1(0,T;V')$ and $\{h_n\}$ and $\{\alpha_n\}$ are such that the sequence $\{D_t u_n(0)\} = \{h_n(0) - \alpha_n(0)\}$ is bounded in V. This rests on a standard argument, that is based on multiplying the inclusion $D_t u_n + \alpha_n(u_n) \ni h_n$ by the time increment $\delta_k u_n$ for any $k > 0$, see e.g. [72]. This may easily be extended to any $s \in]0,1[$.

By Proposition 5.1, the boundedness of $\{u_n\}$ in X^s entails that

$$\begin{aligned} F(u, h - D_t u) &= \int_0^T f(u, h - D_t u)dt \\ \widetilde{F}(u, h - D_t u) &= \int_0^T (T - t) f(u, h - D_t u)dt, \end{aligned} \tag{126}$$

as in (101) and (108). The function u thus fulfills the asymptotic gradient flow.

Conclusions as for the Compactness and Structural Stability of (97) Under the equi-coerciveness assumption (113), a subsequence of the representative functions $\widetilde{F}_n$ Γ-converges in the sense of (117).

Under the convergences (114) and (117) of the data and of the operator, the associated solutions u_n weakly converge in X. The asymptotic pair (u,h) fulfills a monotone relation, that may exhibit long memory. However, if (123) holds and the sequence $\{u_n\}$ is bounded in X^s for some $s > 0$, then the short-memory form (97) is preserved in the limit.

Remark 6.1 (i) Onset of long memory in the limit is also excluded if, in alternative to assuming compactness, we replace the initial condition $u(0) = 0$ by time-periodicity: $u(0) = u(T)$; see [72].

(ii) In Tartar's example above $V = H = L^2(\Omega)$. In this case the lack of compactness in the injection $V \to H$ is at the basis of onset of long memory.

(iii) In a work in progress, the structural stability of Eq. (102) is directly studied without time integration, defining a notion of *time-dependent Γ-convergence.*

7 Variational Formulation of a Class of Nonmonotone Flows

In this section we discuss the extension of the above analysis to some classes of nonmonotone flows, partially along the lines of [69].

Variational Formulations of a Doubly Nonlinear Flow Let us now assume that

$$\begin{cases} \alpha : V \to \mathcal{P}\big(V'\big) & \text{is maximal monotone,} \\ \psi : H \to \mathbb{R} \cup \{+\infty\} & \text{is proper, convex and lower semicontinuous,} \\ h \in L^2\big(0,T;V'\big), \quad w^0 \in H, \end{cases} \tag{127}$$

and consider a problem with two nonlinearities:

$$\begin{cases} u \in L^2(0,T;V), \quad w \in L^2(0,T;H) \cap H^1\big(0,T;V'\big), \\ D_t w + \alpha(u) \ni h \quad \text{in } V', \text{ a.e. in }]0,T[, \\ u \in \partial\psi(w) \quad \text{in } H, \text{ a.e. in }]0,T[, \\ w(0) = w^0. \end{cases} \tag{128}$$

Of course, if $\psi(w) = \frac{1}{2}\|w\|_H^2$ we retrieve (97). (We might prescribe a vanishing initial value, as we did in (97); however in this case this would not provide the space-time monotonicity.)

If (98) and (99) are fulfilled and ψ is coercive, i.e.,

$$\forall C \in \mathbb{R}, \quad \big\{v \in H : \psi(v) \le C\big\} \quad \text{is bounded}, \tag{129}$$

then it is known that the above problem has a solution, see e.g. [3, 24]. Let the operators α and $\partial\psi$ be respectively represented by $f \in \mathcal{F}(V)$ and by the Fenchel function $g \in \mathcal{F}(H)$ (that is, $g(v_1, v_2) = \psi(v_1) + \psi^*(v_2)$ for any $v_1, v_2 \in H$). The system (128) is then equivalent to

$$\begin{cases} u \in L^2(0,T;V), \quad w \in L^2(0,T;H) \cap H^1\big(0,T;V'\big), \\ f(u, h - D_t w) = \langle h - D_t w, u\rangle \quad \text{a.e. in }]0,T[, \\ \psi(w) + \psi^*(u) = (u,w)_H \quad \text{a.e. in }]0,T[, \\ w(0) = w^0. \end{cases} \tag{130}$$

Because of $(128)_3$, the mapping $t \mapsto \psi(w(t))$ is absolutely continuous and differentiable a.e. in $]0,T[$, and

$$D_t\psi(w) = \langle D_t u, z\rangle \quad \text{a.e. in }]0,T[, \ \forall z \in \partial\psi(w).$$

Equation $(130)_2$ is then equivalent to

$$f(u, h - D_t w) + D_t\psi(w) = \langle h, u\rangle \quad \text{a.e. in }]0,T[. \tag{131}$$

As f fulfills (83), this equality is also equivalent to

$$\int_0^\tau f(u, h - D_t w)dt + \psi\big(w(\tau)\big) - \psi\big(w^0\big) = \int_0^\tau \langle h, u\rangle dt \quad \forall \tau \in]0,T[. \tag{132}$$

By the same token, the latter is equivalent to the single equation

$$\int_0^T f(u, h - D_t w)dt + \psi\big(w(T)\big) - \psi\big(w^0\big) = \int_0^T \langle h, u\rangle dt, \tag{133}$$

and also to the twice time-integrated equation

$$\int_0^T \big[(T-t) f(u, h - D_t w) + \psi\big(w(t)\big)\big]dt - T\psi\big(w^0\big) = \int_0^T \langle h, u\rangle dt. \tag{134}$$

Defining F and $\widetilde{F}$ as in (101) and (108), the two latter equations also read

$$J(u,h) := F(u, h - D_t w) + \psi\big(w(T)\big) - \psi\big(w^0\big) = \int_0^T \langle h, u\rangle dt, \tag{135}$$

$$\widetilde{J}(u,h) := \widetilde{F}(u, h - D_t w) + \int_0^T \psi\big(w(t)\big)dt - T\psi\big(w^0\big) = \int_0^T (T-t)\langle h, u\rangle dt. \tag{136}$$

Each one of these equations is equivalent to a null-minimization problem. For instance, (136) is equivalent to

$$\widetilde{K}(u,h) := \widetilde{J}(u,h) - \int_0^T (T-t)\langle h, u\rangle dt = \inf \widetilde{K} = 0. \tag{137}$$

On the other hand $(130)_3$ is equivalent to

$$\int_0^T \big[\psi(w) + \psi^*(u)\big]dt = \int_0^T (u, w)_H dt, \tag{138}$$

which is also equivalent to a null-minimization problem:

$$\widetilde{H}(u,h) := \int_0^T \big[\psi(w) + \psi^*(u)\big]dt - \int_0^T (u, w)_H dt = \inf \widetilde{H} = 0. \tag{139}$$

Finally, each of these systems either of two equations or of two null-minimization problems is equivalent to a single null-minimization problem. For instance, the system (137), (139) is equivalent to

$$\widetilde{K}(u,h) + \widetilde{H}(u,h) = \inf(\widetilde{K} + \widetilde{H}) = 0. \tag{140}$$

(Of course, these equivalences rest on the two conditions (83) and (84) of representation.)

Conclusions as for the Variational Formulation of (128) The system (128) is equivalent to (130), and this is tantamount to a null-minimization problem.

Each of Eqs. (131)–(134) coupled with (138) is formally equivalent to the system of the two Eqs. $(130)_2$ and $(130)_3$. Each of these systems may be formulated as a null-minimization problem.

The structural stability of the problem (128) may be proved by using tools analogous to those of Sect. 6; hopefully, this issue will be addressed in a work apart.

Remark 7.1 The present discussion may be extended to doubly-nonlinear problems of the form

$$\begin{cases} u \in L^2(0,T;V) \cap H^1(0,T;H), \quad w \in L^2(0,T;H), \\ w + \alpha(u) \ni h \quad \text{in } V', \text{ a.e. in }]0,T[, \\ w \in \partial\psi(D_t u) \quad \text{in } H, \text{ a.e. in }]0,T[. \end{cases} \tag{141}$$

A variational formulation may also be given for this problem, and structural stability may be studied.

8 Variational Formulation of the Heat and Mass Diffusion Problem

In this section we address the variational formulation of the problem of Sect. 4.

Variational Formulation of the Single-Phase Problem Let us first consider the problem of heat and mass diffusion without phase transition

$$D_t u + \nabla \cdot \boldsymbol{j}_u = f_1 \quad \text{in } V_1', \text{ a.e. in }]0, T[, \tag{142}$$

$$D_t c + \nabla \cdot \boldsymbol{j}_c = f_2 \quad \text{in } V_2', \text{ a.e. in }]0, T[, \tag{143}$$

$$(\theta, \omega) \in \partial\varphi(u, c) \quad \text{a.e. in } Q, \tag{144}$$

$$(\boldsymbol{j}_u, \boldsymbol{j}_c) = -\gamma(\theta, \omega, \nabla\theta, \nabla\omega) \quad \text{a.e. in } Q. \tag{145}$$

By setting

$$\begin{aligned} &U := (u, c), \qquad \Theta := (\theta, \omega), \\ &J := (\boldsymbol{j}_u, \boldsymbol{j}_c), \qquad \Lambda J := \nabla \cdot J = (\nabla \cdot \boldsymbol{j}_u, \nabla \cdot \boldsymbol{j}_c), \\ &V := V_1 \times V_2, \qquad f := (f_1, f_2) \in V', \end{aligned} \tag{146}$$

the system (142)–(145) also reads

$$\Theta \in \partial\varphi(U) \quad \text{a.e. in } Q, \tag{147}$$

$$J = -\gamma(\Theta, \nabla\Theta) \quad \text{a.e. in } Q, \tag{148}$$

$$D_t U + \Lambda J = f \quad \text{in } V', \text{ a.e. in }]0, T[. \tag{149}$$

Denoting by F the Fenchel function $\varphi + \varphi^*$, the relation (147) is clearly equivalent to

$$F(U, \Theta) = U \cdot \Theta \quad \text{a.e. in } Q. \tag{150}$$

Next we shall formulate the relation (148) in $V \times V'$ a.e. in $]0, T[$, rather than pointwise in Q. Let us first denote by g_Θ a representative function of the maximal monotone mapping $\gamma(\Theta, \cdot) : (\mathbb{R}^3)^2 \to \mathcal{P}((\mathbb{R}^3)^2)$, so that (148) also reads

$$g_\Theta(\nabla\Theta, -J) = -J \cdot \nabla\Theta \quad \text{a.e. in } Q \tag{151}$$

(here $\Theta \in (\mathbb{R}^3)^2$ just plays the role of a parameter). Let us assume that

$$\forall C \in \mathbb{R}, \sup\left\{\|S\|_V + \left\|S^*\right\|_{V'} : \left(S, S^*\right) \in V \times V', g_\Theta\left(S, S^*\right) \le C\right\} < +\infty, \tag{152}$$

uniformly with respect to Θ, and define the function

$$G_\Theta(S, S^*) = \inf\left\{\int_\Omega g_\Theta(\nabla S, Z)dx : Z \in \left(L^2(\Omega)^3\right)^2, -\nabla \cdot Z = S^* \text{ in } \mathcal{D}'(\Omega)^2\right\}$$
$$\forall (S, S^*) \in V \times V'. \tag{153}$$

By (152) this infimum is attained at some $\hat{Z} \in (L^2(\Omega)^3)^2$. The function G_Θ is convex and lower semicontinuous, and

$$G_\Theta(S, S^*) = \int_\Omega g_\Theta(\nabla S, \hat{Z})dx \overset{g_\Theta \in \mathcal{F}((\mathbb{R}^3)^2)}{\geq} \int_\Omega \hat{Z} \cdot \nabla S dx = -\langle \nabla \cdot \hat{Z}, S\rangle = \langle S^*, S\rangle; \tag{154}$$

thus $G_\Theta \in \mathcal{F}(V)$. Moreover, as $g_\Theta(\nabla S, \hat{Z}) \geq \hat{Z} \cdot \nabla S$ pointwise in Ω, equality holds in (154) if and only if $g_\Theta(\nabla S, \hat{Z}) = \hat{Z} \cdot \nabla S$ a.e. in Ω. As the function g_Θ represents $\gamma(\Theta, \cdot)$, this is equivalent to $\hat{Z} = \gamma(\Theta, \nabla S)$ a.e. in Ω, whence

$$\hat{S}^* = -\nabla \cdot \hat{Z} = -\nabla \cdot \gamma(\Theta, \nabla S) \quad \text{ in } \left(H^{-1}(\Omega)^3\right)^2.$$

Denoting by $\langle \cdot, \cdot \rangle$ the duality between V' and V, we may then replace (148) by the equation

$$G_\Theta(\Theta, \nabla \cdot J) = \langle \Lambda \cdot J, \Theta\rangle \quad \text{a.e. in }]0, T[. \tag{155}$$

By eliminating Eq. (149), we then infer that the system (147)–(149) is equivalent to

$$F(U, \Theta) = U \cdot \Theta \quad \text{a.e. in } Q, \tag{156}$$
$$G_\Theta(\Theta, f - D_t U) + \langle D_t U, \Theta\rangle = \langle f, \Theta\rangle \quad \text{a.e. in }]0, T[. \tag{157}$$

By (147) (or equivalently (156)), we have $D_t\varphi(U) = \langle D_t U, \Theta\rangle$. Assuming the initial condition $U(0) = U^0$, Eq. (157) is then also equivalent to either of the following equations

$$G_\Theta(J, f - D_t U)dt + D_t\varphi(U) = \langle f, \Theta\rangle dt \quad \text{a.e. in }]0, T[, \tag{158}$$
$$\int_0^T G_\Theta(J, f - D_t U)dt + \varphi\big(U(T)\big) - \varphi\big(U^0\big) = \int_0^T \langle f, \Theta\rangle dt, \tag{159}$$
$$\int_0^T (T - t)G_\Theta(J, f - D_t U)dt + \int_0^T \varphi\big(U(T)\big)dt - T\varphi\big(U^0\big)$$
$$= \int_0^T (T - t)\langle f, \Theta\rangle dt. \tag{160}$$

Therefore the system (147)–(149) is equivalent to either of these equations coupled with

$$\iint_Q F(U, \Theta)dxdt = \iint_Q U \cdot \Theta dxdt. \tag{161}$$

Each of these equations is equivalent to a null-minimization problem; therefore the whole system is equivalent to a single null-minimization, in analogy with (140).

Variational Formulation of the Glass-Formation Problem In Sect. 3 we derived the model at the basis of Problem 4.1, i.e.,

$$(\theta, \omega, r) \in \partial\varphi(u, c, \chi) \quad \text{a.e. in } Q, \tag{162}$$

$$(\boldsymbol{j}_u, \boldsymbol{j}_c) = -\gamma(\theta, \omega, \nabla\theta, \nabla\omega) \quad \text{a.e. in } Q, \tag{163}$$

$$D_t u + \nabla \cdot \boldsymbol{j}_u = f_1 \quad \text{in } V_1', \text{ a.e. in }]0, T[, \tag{164}$$

$$D_t c + \nabla \cdot \boldsymbol{j}_c = f_2 \quad \text{in } V_2', \text{ a.e. in }]0, T[, \tag{165}$$

$$D_t \chi + r = \rho(\theta, \omega) \quad \text{a.e. in } Q. \tag{166}$$

Next we replace the definitions (146) by

$$\begin{aligned} &U := (u, c, \chi), \qquad \Theta := (\theta, \omega, r), \\ &J := (\boldsymbol{j}_u, \boldsymbol{j}_c, r), \qquad \Lambda J := (\nabla \cdot \boldsymbol{j}_u, \nabla \cdot \boldsymbol{j}_c, r), \\ &V := V_1 \times V_2 \times L^2(\Omega), \qquad f(\Theta) := \big(f_1, f_2, \rho(\theta, \omega)\big) \in V'. \end{aligned} \tag{167}$$

The system (162)–(166) then also reads:

$$\Theta \in \partial\varphi(U) \quad \text{a.e. in } Q, \tag{168}$$

$$J = -\gamma(\Theta, \nabla\Theta) \quad \text{a.e. in } Q, \tag{169}$$

$$D_t U + \Lambda J = f(\Theta) \quad \text{in } V', \text{a.e. in }]0, T[. \tag{170}$$

Defining F and G_Θ as above, we may then repeat the analysis of (150)–(161), with the proviso of replacing the prescribed source term f by $f(\Theta)$. However, despite of the formal analogy, this problem differs from that of the first part of this section: for instance, this problem also includes the ODE (166).

Conclusions as for the Variational Formulation of (162)–(166) This system is equivalent to either of Eqs. (158)–(160) coupled with (161) (here with $f(\Theta)$ in place of f). Each of these formulations is tantamount to a single null-minimization problem.

The analysis of the structural stability of this problem is here left open.

Acknowledgement This research was partially supported by the P.R.I.N. project "Phase transitions, hysteresis and multiscaling" of Italian M.I.U.R.

References

1. Alexiades, V., Solomon, A.D.: Mathematical Modeling of Melting and Freezing Processes. Hemisphere, Washington (1993)

2. Alexiades, V., Wilson, D.G., Solomon, A.D.: Macroscopic global modeling of binary alloy solidification processes. Q. Appl. Math. **43**, 143–158 (1985)
3. Alt, H.W., Luckhaus, S.: Quasilinear elliptic-parabolic differential equations. Math. Z. **183**, 311–341 (1983)
4. Ansini, N., Dal Maso, G., Zeppieri, C.I.: New results on Gamma-limits of integral functionals. Preprint, SISSA, Trieste (2012)
5. Attouch, H.: Variational Convergence for Functions and Operators. Pitman, Boston (1984)
6. Barbu, V.: Nonlinear Differential Equations of Monotone Types in Banach Spaces. Springer, Berlin (2010)
7. Braides, A.: Γ-Convergence for Beginners. Oxford University Press, Oxford (2002)
8. Braides, A., Defranceschi, A.: Homogenization of Multiple Integrals. Oxford University Press, Oxford (1998)
9. Brezis, H.: Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert. North-Holland, Amsterdam (1973)
10. Brezis, H., Ekeland, I.: Un principe variationnel associé à certaines équations paraboliques. I. Le cas indépendant du temps. C. R. Acad. Sci. Paris, Sér. A–B **282**, 971–974 (1976)
11. Brezis, H., Ekeland, I.: Un principe variationnel associé à certaines équations paraboliques. II. Le cas dépendant du temps. C. R. Acad. Sci. Paris, Sér. A–B **282**, 1197–1198 (1976)
12. Brokate, M., Sprekels, J.: Hysteresis and Phase Transitions. Springer, Heidelberg (1996)
13. Burachik, R.S., Svaiter, B.F.: Maximal monotone operators, convex functions, and a special family of enlargements. Set-Valued Anal. **10**, 297–316 (2002)
14. Burachik, R.S., Svaiter, B.F.: Maximal monotonicity, conjugation and the duality product. Proc. Am. Math. Soc. **131**, 2379–2383 (2003)
15. Callen, H.B.: Thermodynamics and an Introduction to Thermostatistics. Wiley, New York (1985)
16. Chalmers, B.: Principles of Solidification. Wiley, New York (1964)
17. Chiadò Piat, V., Dal Maso, G., Defranceschi, A.: G-convergence of monotone operators. Ann. Inst. Henri Poincaré, Anal. Non Linéaire **7**, 123–160 (1990)
18. Christian, J.W.: The Theory of Transformations in Metals and Alloys. Part 1: Equilibrium and General Kinetic Theory. Pergamon Press, London (2002)
19. Colli, P.L., Visintin, A.: On a class of doubly nonlinear evolution problems. Commun. Partial Differ. Equ. **15**, 737–756 (1990)
20. Dal Maso, G.: An Introduction to Γ-Convergence. Birkhäuser, Boston (1993)
21. De Giorgi, E., Franzoni, T.: Su un tipo di convergenza variazionale. Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat. (8) **58**, 842–850 (1975)
22. De Groot, S.R.: Thermodynamics of Irreversible Processes. North-Holland, Amsterdam (1961)
23. De Groot, S.R., Mazur, P.: Non-equilibrium Thermodynamics. North-Holland, Amsterdam (1962)
24. DiBenedetto, E., Showalter, R.E.: Implicit degenerate evolution equations and applications. SIAM J. Math. Anal. **12**, 731–751 (1981)
25. Donnelly, J.D.P.: A model for non-equilibrium thermodynamic processes involving phase changes. J. Inst. Math. Appl. **24**, 425–438 (1979)
26. Ekeland, I., Temam, R.: Analyse Convexe et Problèmes Variationnelles. Dunod/Gauthier-Villars, Paris (1974)
27. Eckart, C.: The thermodynamics of irreversible processes I: The simple fluid, Phys. Rev. **58** (1940)
28. Eckart, C.: The thermodynamics of irreversible processes II. Fluid mixtures, Phys. Rev. **58** (1940)
29. Elliott, C.M., Ockendon, J.R.: Weak and Variational Methods for Moving Boundary Problems. Pitman, Boston (1982)
30. Fitzpatrick, S.: Representing monotone operators by convex functions. In: Workshop/Miniconference on Functional Analysis and Optimization, Canberra, 1988. Proc. Centre Math. Anal. Austral. Nat. Univ., vol. 20, pp. 59–65. Austral. Nat. Univ., Canberra (1988)

31. Flemings, M.C.: Solidification Processing. McGraw-Hill, New York (1973)
32. Francfort, G., Murat, F., Tartar, L.: Homogenization of monotone operators in divergence form with x-dependent multivalued graphs. Ann. Mat. Pura Appl. (4) **118**, 631–652 (2009)
33. Frémond, M.: Phase Change in Mechanics. Springer, Berlin (2012)
34. Ghoussoub, N.: A variational theory for monotone vector fields. J. Fixed Point Theory Appl. **4**, 107–135 (2008)
35. Ghoussoub, N.: Self-dual Partial Differential Systems and Their Variational Principles. Springer, Berlin (2009)
36. Gupta, S.C.: The classical Stefan problem. In: Basic Concepts, Modelling and Analysis. North-Holland Series. Elsevier, Amsterdam (2003)
37. Hiriart-Urruty, J.-B., Lemarechal, C.: Convex Analysis and Optimization Algorithms. Springer, Berlin (1993)
38. Kondepudi, D., Prigogine, I.: Modern Thermodynamics. Wiley, New York (1998)
39. Kurz, W., Fisher, D.J.: Fundamentals of Solidification. Trans Tech, Aedermannsdorf (1989)
40. Lions, J.L., Magenes, E.: Non-homogeneous Boundary Value Problems and Applications, vols. I, II. Springer, Berlin (1972). French edition: Dunod, Paris (1968)
41. Luckhaus, S.: Solidification of alloys and the Gibbs-Thomson law. Preprint (1994)
42. Luckhaus, S., Visintin, A.: Phase transition in a multicomponent system. Manuscr. Math. **43**, 261–288 (1983)
43. Martinez-Legaz, J.-E., Svaiter, B.F.: Monotone operators representable by l.s.c. convex functions. Set-Valued Anal. **13**, 21–46 (2005)
44. Martinez-Legaz, J.-E., Svaiter, B.F.: Minimal convex functions bounded below by the duality product. Proc. Am. Math. Soc. **136**, 873–878 (2008)
45. Martinez-Legaz, J.-E., Théra, M.: A convex representation of maximal monotone operators. J. Nonlinear Convex Anal. **2**, 243–247 (2001)
46. Marques Alves, M., Svaiter, B.F.: Brøndsted-Rockafellar property and maximality of monotone operators representable by convex functions in non-reflexive Banach spaces. J. Convex Anal. **15**, 693–706 (2008)
47. Müller, I.: A History of Thermodynamics. Springer, Berlin (2007)
48. Müller, I., Weiss, W.: Entropy and Energy. A Universal Competition. Springer, Berlin (2005)
49. Müller, I., Weiss, W.: A history of thermodynamics of irreversible processes (in preparation)
50. Nayroles, B.: Deux théorèmes de minimum pour certains systèmes dissipatifs. C. R. Acad. Sci. Paris, Sér. A–B **282**, A1035–A1038 (1976)
51. Penrose, O., Fife, P.C.: Thermodynamically consistent models of phase-field type for the kinetics of phase transitions. Physica D **43**, 44–62 (1990)
52. Penrose, O., Fife, P.C.: On the relation between the standard phase-field model and a "thermodynamically consistent" phase-field model. Physica D **69**, 107–113 (1993)
53. Prigogine, I.: Thermodynamics of Irreversible Processes. Wiley-Interscience, New York (1967)
54. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1969)
55. Stefanelli, U.: The Brezis-Ekeland principle for doubly nonlinear equations. SIAM J. Control Optim. **8**, 1615–1642 (2008)
56. Stefan, J.: Über einige Probleme der Theorie der Wärmeleitung. Sitzungber., Wien, Akad. Mat. Natur. **98**, 473–484 (1889)
57. Stefan, J.: Über einige Probleme der Theorie der Wärmeleitung. Sitzungber., Wien, Akad. Mat. Natur. **98**, 614–634 (1889)
58. Stefan, J.: Über einige Probleme der Theorie der Wärmeleitung. Sitzungber., Wien, Akad. Mat. Natur. **98**, 965–983 (1889)
59. Stefan, J.: Über einige Probleme der Theorie der Wärmeleitung. Sitzungber., Wien, Akad. Mat. Natur. **98**, 1418–1442 (1889)
60. Tartar, L.: Nonlocal effects induced by homogenization. In: Colombini, F., Marino, A., Modica, L., Spagnolo, S. (Eds.) Partial Differential Equations and the Calculus of Variations, vol. II, pp. 925–938. Birkhäuser, Boston (1989)

61. Tartar, L.: Memory effects and homogenization. Arch. Ration. Mech. Anal. **111**, 121–133 (1990)
62. Tartar, L.: The General Theory of Homogenization. A Personalized Introduction. Springer/UMI, Berlin/Bologna (2009)
63. Tarzia, D.A.: A Bibliography on Moving-Free Boundary Problems for the Heat-Diffusion Equation. The Stefan and Related Problems. Universidad Austral, Departamento de Matematica, Rosario (2000)
64. Visintin, A.: Models of Phase Transitions. Birkhäuser, Boston (1996)
65. Visintin, A.: Introduction to Stefan-type problems. In: Dafermos, C., Pokorny, M. (Eds.) Handbook of Differential Equations: Evolutionary Differential Equations, vol. IV, pp. 377–484. North-Holland, Amsterdam (2008), Chap. 8
66. Visintin, A.: Extension of the Brezis-Ekeland-Nayroles principle to monotone operators. Adv. Math. Sci. Appl. **18**, 633–650 (2008)
67. Visintin, A.: Phase transitions and glass formation. SIAM J. Math. Anal. **41**, 1725–1756 (2009)
68. Visintin, A.: Scale-transformations in the homogenization of nonlinear magnetic processes. Arch. Ration. Mech. Anal. **198**, 569–611 (2010)
69. Visintin, A.: Structural stability of doubly nonlinear flows. Boll. Unione Mat. Ital. **IV**, 363–391 (2011)
70. Visintin, A.: On the structural stability of monotone flows. Boll. Unione Mat. Ital. **IV**, 471–479 (2011)
71. Visintin, A.: Structural stability of rate-independent nonpotential flows. Discrete Contin. Dyn. Syst. (in press)
72. Visintin, A.: Variational formulation and structural stability of monotone equations. Calc. Var. Partial Differ. Equ. (in press)
73. Woodruff, D.P.: The Solid-Liquid Interface. Cambridge University Press, Cambridge (1973)
74. Zeidler, E.: Nonlinear Functional Analysis and Its Applications. II/B. Nonlinear Monotone Operators. Springer, New York (1990)