International Series of Numerical Mathematics

166

Aldo Pratelli
Günter Leugering
Editors

New Trends in Shape Optimization

Birkhäuser

ISNM

International Series of Numerical Mathematics

Volume 166

More information about this series at www.birkhauser-science.com/series/4819

Aldo Pratelli · Günter Leugering
Editors

New Trends in Shape Optimization

Birkhäuser

Editors
Aldo Pratelli
Department Mathematik
Universität Erlangen-Nürnberg
Erlangen
Germany

Günter Leugering
Department Mathematik
Universität Erlangen-Nürnberg
Erlangen
Germany

ISSN 0373-3149 ISSN 2296-6072 (electronic)
International Series of Numerical Mathematics
ISBN 978-3-319-17562-1 ISBN 978-3-319-17563-8 (eBook)
DOI 10.1007/978-3-319-17563-8

Library of Congress Control Number: 2015950019

Mathematics Subject Classification (2010): 35B20, 35B40, 35J25, 35J40, 35J57, 35P15, 35R11, 35R35, 35Q61, 45C05, 47A75, 49J45, 49R05, 49R50, 49Q10, 49Q12, 49Q20, 53A10, 65D15, 65N30, 74K20, 76D07, 76Q20, 78A50, 81Q05, 82B10

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.birkhauser-science.com)

Contents

Introduction

Shape optimization has emerged from calculus of variations and from the theory of mathematical optimization in topological vector spaces together with the field of numerical methods for optimization and can be said to have established itself as one of the most fruitful applications of mathematics to problems in science and engineering. The particular mathematical challenge that distinguishes this field from standard calculus of variations or mathematical optimization is the fact the optimization variables are now represented by shapes or sets. One, thus, has to topologize these sets in a way such that the convergence of sets is consistent with the topological properties of the performance criterion and possibly side constraints. Indeed, a typical shape optimization problem consists of a so-called cost function depending on the shape of a body and state variables. The state variables are governed by a partial differential equation, the state equation, and specific equality or inequality constraints, depending also on the shape of the body. The problem then is to choose a shape from the set of admissible shapes such that the cost function is minimized over all admissible states and shapes.

While one branch of the field is concerned with set convergence, proper topologies, and existence results for the optimization problems together with optimality conditions, another branch investigates the sensitivity of the above-mentioned quantities with respect to shape changes. Here shape differentiability of first and second order comes into play and shape gradient descent as well as shape Newton methods are investigated and numerically implemented.

A third development emerges into the direction of applications in science, engineering, and industry, where the discretization plays a major role along with modern tools in supercomputing.

The goal of the workshop on "Trends in shape optimization," held in Erlangen in the fall of 2013, was to bring together experts in these different areas of shape optimization in order to provide a platform for intense mathematical discussions. It was felt that in order to further develop and foster this field at the frontiers of modern mathematics, an intensified interaction and exchange of methods and tools would be indispensable.

As a result, we brought together scientists from Canada, France, Germany, Italy, Poland, and USA focusing on general minimization problems, spectral problems, Cheeger problems, metric spaces of shapes and geometries, phase field approaches and crack propagation control for shape and topology problems, approximate shape gradients for interface problems, Lagrange–Newton methods for PDE-constrained problems, electromagnetic applications, nonlinear problems and regularity results for the half-Laplacian, and many other topics not covered in this volume.

Most speakers of the workshop provided survey articles, research articles. as well as notes on works in progress. The present volume very much reflects the atmosphere of the workshop, where purely theoretical results and novel approaches interfered with more application-oriented developments and industrial applications.

The editors sincerely hope that this book on shape optimization will have an impact on this exciting field of mathematics.

We are grateful to Lukas Pflug for his administrative support during the process of editing this volume.

Erlangen
July 2015

Prof. Aldo Pratelli
Prof. Günter Leugering

On the Minimization of Area Among Chord-Convex Sets

Beatrice Acciaio and Aldo Pratelli

Abstract In this paper we study the problem of minimizing the area for the chord-convex sets of given size, that is, the sets for which each bisecting chord is a segment of length at least 2. This problem has been already studied and solved in the framework of convex sets, though nothing has been said in the non-convex case. We introduce here the relevant concepts and show some first properties.

Keywords Area-minimizing sets · Chord convex

1 Introduction and Setting of the Problem

Consider the convex planar sets with the property that all the bisecting chords (i.e., the segments dividing the set in two parts of equal area) have length at least 2. A very simple question is which set in this class minimizes the area. Surprisingly enough, the answer is not the unit disk, as one would immediately guess, but the so-called "Auerbach triangle", shown in Fig. 1 left.

The story of this problem is quite old: already in the 1920s Zindler posed the question whether the disk is the unique planar convex set having all the bisecting chords of the same length (see [4]), while few years later Ulam asked if there are other planar convex sets, besides the disk, which have the floating property, that can be described as follows. Assume that the set has density 1/2 and it is immersed in the water (hence, half of the set remains immersed while the other half stays out of

Both the authors have been supported through the ERC St.G. AnOpt- SetCon.

B. Acciaio (✉)
Department of Statistics, London School of Economics, 10 Houghton Street, London WC2A 2AE, UK
e-mail: b.acciaio@lse.ac.uk

A. Pratelli
Department Mathematik, Universität Erlangen, Cauerstrasse 11, 91056 Erlangen, Germany
e-mail: pratelli@math.fau.de

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*, International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_1

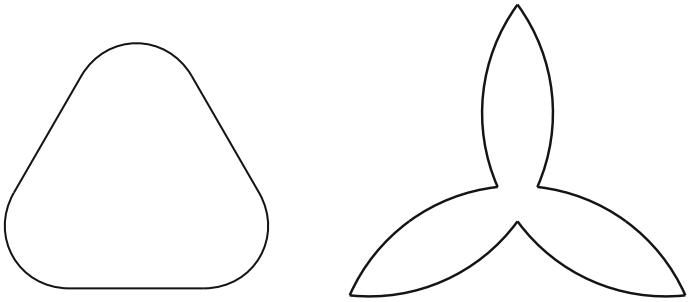


Fig. 1 The "Auerbach triangle" and the "Zindler flower"

the water): the set is said to have the "Ulam floating property" if the floating position is of equilibrium, and if this remains true after an arbitrary rotation of the set. For instance, of course the disk has the Ulam floating property, while any other ellipsis does not: indeed, only two floating positions of the cylinder are of equilibrium (those for which the water is parallel to one of the two symmetry axes of the ellipsis), while all the other positions are not of equilibrium. In the 1930s, Auerbach showed that the two problems above are equivalent (see [1]), and that there is a whole class of sets having these two properties (we call such sets "Zindler sets"). In his paper, Auerbach considered also the question of which Zindler set minimizes the area, among those for which the length of every bisecting chord is the same, say 2: he was able to show that the answer is not the disk, and he conjectured the solution to be a "triangle" -named after him the Auerbach triangle—whose area is ≈ 3.11, thus just about 1 % less than that of the unit disk.

In the last years the problem addressed above was finally solved in [2, 3]. In particular, Fusco and the second author proved in [3] the Auerbach conjecture, that is, the Auerbach triangle minimizes the area among the Zindler sets. Then, Esposito, Ferone, Kawohl, Nitsch and Trombetti in [2] proved that the convex set with minimal area (among those with all the bisecting chords of length at least 2) must be a Zindler set, and thus it is the Auerbach triangle.

Up to now, nothing has been said in the non-convex case, and the aim of this paper is to start working on this more general problem. We can immediately notice that the Auerbach triangle is no longer the solution if we allow other sets to be considered: for instance, consider the "Zindler flower", shown in Fig. 1 right. As it appears evident from the figure, the boundary of this set is contained in the union of three equal arcs of circle, each of which covers an angle of 120°. It can be easily calculated that the area of this set is ≈ 2.54, then much smaller than both the area of the unit disk, and that of the Auerbach triangle.

In this paper, we consider the class of the "chord-convex sets", see Definition 2.1: roughly speaking, these sets are not necessarily convex, but have the property that all the bisecting chords are actually segments. We will be able to prove some preliminary interesting properties of these sets, and also to give some counterexamples.

2 Definitions and Results

In this section we introduce all the relevant concepts and we prove our results. It would be impossible to first give all the definitions and then present and prove the results, because most of the definitions would not make sense if some properties have not been preliminarily proved, hence we have to give them in parallel. This is done in three parts: the uniqueness of the bisecting chords, the properties of the extremes of the bisecting chords and of the intersections between chords, and the Zindler sets. First of all, we introduce the class of sets that we will consider.

Definition 2.1 Let $E \subseteq \mathbb{R}^2$ be an open set of finite measure with the property that $E = \text{Int}\,\overline{E}$. A line r is called a *bisecting line* if the intersection of E with each of the two half-spaces in which $\mathbb{R}^2$ is subdivided by r has area exactly $|E|/2$. The set E is called *chord-convex* if the intersection of $\overline{E}$ with each bisecting line is a closed segment, which will be called *bisecting chord*. The *size* of a chord-convex set is the minimal length of a bisecting chord.

The main problem that one wants to consider is the minimization of the area among the chord-convex sets of size at least 2 (or, equivalently, of size 2). For instance, the unit disk of area π belongs to this class, as well as the Auerbach triangle, which has area $\approx$3.11, and as the Zindler flower, which has area $\approx$2.54: the first two sets are also convex, while the third is only chord-convex.

2.1 Uniqueness of the Bisecting Chord of Given Direction

The first property that we want to investigate is the uniqueness of the bisecting chord of a given direction, to which we will devote the present section. Indeed, it is obvious by continuity that for every direction there is some bisecting chord of that direction, but the uniqueness is not clear, since it is not obvious that a chord-convex set is connected. Actually, as we will see in Theorem 2.6 and Example 2.1, the closure of a chord-convex set is always connected (and even simply connected), but a chord-convex set needs not be connected. However, the uniqueness of the bisecting chord of any given direction is ensured by Theorem 2.6. Before proving that, we need a couple of technical results and of definitions. Throughout this section, E will always denote a chord-convex set.

Definition 2.2 Let $x,\ y \in \overline{E}$. We say that x and y are *connected* if there is a path in $\overline{E}$ connecting x and y. If b is a bisecting chord, we say that x and b are connected if there is some $y \in b$ such that x and y are connected. Notice that, by definition, if x and b are connected then x is connected to every point $y \in b$.

Lemma 2.3 *Let E be a chord-convex set, b be a bisecting chord of direction $\bar{\theta} \in \mathbb{S}^1$, and let $x \in E \cap b$. Then there exists $\varepsilon > 0$ such that for all $\theta \in (\bar{\theta} - \varepsilon, \bar{\theta} + \varepsilon)$ there is a unique bisecting chord $b(\theta)$ of direction θ, and this chord is connected to x.*

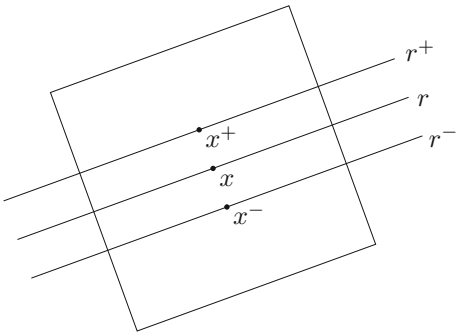


Fig. 2 The situation in Lemma 2.3

Proof Let r be the bisecting line containing b. Since E is open and $x \in E$, we can take a square centered in x, with two sides parallel to r, entirely contained in E. Let now $x^{\pm}$ be two points in the interior of the square, contained in the two opposite halfspaces defined by the bisecting line r containing the chord b, and let also $r^{\pm}$ be the two lines parallel to r passing through $x^{\pm}$. The situation is depicted in Fig. 2. Let us now define H_r the halfspace "below" the line r, that is, the one containing r^-. In the same way, we call H_γ the halfspace "below" γ: this makes sense for every line γ with a direction close to that of r. Since r is a bisecting line, we have $|E \cap H_r| = 1/2$, thus $|E \cap H_{r^-}| < 1/2 < |E \cap H_{r^+}|$ because the square is entirely contained in E. By continuity, there is $\varepsilon > 0$ such that $|E \cap H_{r^-(\theta)}| < 1/2 < |E \cap H_{r^+(\theta)}|$ for all $\theta \in (\bar\theta - \varepsilon, \bar\theta + \varepsilon)$, where $r^+(\theta)$ and $r^-(\theta)$ are the lines of direction θ passing through x^+ and x^- respectively. Again by continuity and using the fact that the square is contained in E, we deduce that there is a unique bisecting line of direction θ, which lies between $r^+(\theta)$ and $r^-(\theta)$, and thus intersects the square. Since E is chord-convex, we also deduce the existence and uniqueness of a bisecting chord $b(\theta)$. Since this chord intersects the square, we derive the fact that $b(\theta)$ is connected to x. □

Lemma 2.4 *Let r, s be two different bisecting lines of a chord-convex set E, let T be one of the four corresponding open regions, and let $x \in T \cap E$. Then, there exists a bisecting line passing through x, whose direction belongs to the open interval in $\mathbb{S}^1$ corresponding to T.*

Proof Let us call, as in Fig. 3, r_x and s_x the two lines passing through x and parallel to r and s respectively, and let us denote, as in the proof of Lemma 2.3, by H_γ the half-space "below" γ for any line γ having direction between those of r and of s. Then, by construction $|E \cap H_{r_x}| \le |E \cap H_r| = 1/2$, since $x \in T$ and r is a bisecting line. Moreover, since x belongs to E, then in fact it must be $|E \cap H_{r_x}| < 1/2$, and in the very same way $|E \cap H_{s_x}| > 1/2$. By continuity, there is clearly a line passing through x, with direction in the open interval of $\mathbb{S}^1$ corresponding to T, and which is bisecting. □

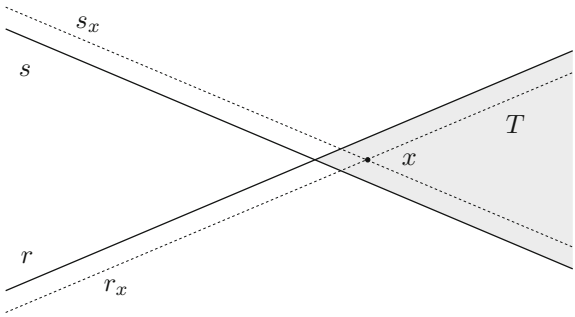


Fig. 3 The situation in Lemma 2.4

Definition 2.5 We say that the sequence of lines $\{r_n\}_{n\in\mathbb{N}}$ *converges* to the line r if the directions of r_n converge in $\mathbb{S}^1$ to the direction of r, and for any ball B big enough the segments $r_n \cap B$ converge in the Hausdorff sense to the segment $r \cap B$.

We are finally in the position to prove the main result of this section.

Theorem 2.6 *Let E be a chord-convex set. Then $\overline{E}$ is connected and simply connected, and there is a unique bisecting chord for every direction in $\mathbb{S}^1$.*

Proof We will divide the proof of this result in four steps, for the sake of simplicity.

Step I. Every sequence of bisecting lines converges to a bisecting line up to a subsequence.

Let $\{r_n\}$ be a sequence of bisecting lines: first of all, up to a subsequence we can assume that the directions of the lines r_n converge to some $\theta \in \mathbb{S}^1$. Then, we will obtain the existence of a line r such that the sequence $\{r_n\}$ converges to r (up to a subsequence, of course) as soon as we show the existence of a ball B which has non-empty intersection with all the lines r_n. Let then B be a ball centered at the origin and with the property that $|E \cap B| > |E|/2$, which clearly exists since this is true if the radius of the ball is big enough. We have that $B \cap r_n \neq \emptyset$, which follows immediately from the fact that r_n is a bisecting line and thus, as pointed out above, we derive the existence of a line r such that a suitable subsequence of $\{r_n\}$ converges to r. To conclude this step, we only have to show that r is a bisecting line, but this is in turn obvious by continuity and since so are all the lines r_n.

Step II. If r and s are two bisecting lines such that there exists $x \in E \cap r \backslash s$, then for each of the four open regions T_i, $1 \leq i \leq 4$ determined by r and s one has $|T_i \cap E| > 0$.

Let us assume, without loss of generality, that x belongs to the closures of T_1 and T_2, and that T_3 and T_4 are the regions opposite to T_1 and T_2 respectively. Then we have by construction that $|E \cap T_1| = |E \cap T_3|$, and in turn $|E \cap T_1| > 0$ because x belongs to the open set E. The same argument shows also $|E \cap T_2| = |E \cap T_4| > 0$, hence the step is completed.

Step III. The set $\overline{E}$ is connected and there is a unique bisecting chord for each direction.

Take a generic point $x \in E$, and let r be a bisecting line passing through x. Without loss of generality, let us assume that the line r is horizontal. Define now

$$\bar{\theta} := \max\{\, \nu \in [0,\pi] : \forall\, 0 \le \theta < \nu,\ \exists!\, \text{bisecting chord } b(\theta) \text{ of direction } \theta, \text{ and } b(\theta) \text{ is connected to } x \,\}.$$

Of course, if we show that $\bar{\theta} = \pi$ then we have proved at once the uniqueness of the bisecting chords for any direction, and the connectedness of $\overline{E}$ (since any two points of $\overline{E}$ are connected to x, and then they are connected among themselves).

Let us then assume by contradiction that $\bar{\theta} < \pi$, and notice that Lemma 2.3 ensures that $\bar{\theta} > 0$. Pick now a bisecting line s of direction $\bar{\theta}$—we still do not know whether this bisecting line is unique, or connected to x, but the existence is obvious. The point x cannot belong to s, because otherwise Lemma 2.3 would give a contradiction to the maximality of $\bar{\theta}$. Let us call T the region (shaded in Fig. 4) determined by r and s corresponding to the angles between 0 and $\bar{\theta}$ and not containing x in its closure. By Step II we have $|T \cap E| > 0$, so in particular there is some point $y \in T \cap E$. Applying Lemma 2.4 to this point y and the region T, we find a bisecting line passing through y and with direction $\hat{\theta} \in (0, \bar{\theta})$. By definition of $\bar{\theta}$, we know that this line is the unique bisecting line with direction $\hat{\theta}$, and that its intersection $b(\hat{\theta})$ with $\overline{E}$ is connected to x. Thus, there is some path $\gamma \subseteq \overline{E}$ starting from y and ending at x, and this path must clearly intersect s.

As in Fig. 4, let us call $s^{\pm}$ two lines parallel to s, lying on the opposite hyperplanes defined by s. We can choose the lines very close to s, so in particular neither y nor x is between them, and hence γ must cross both $s^{\pm}$. Recalling that γ is contained in $\overline{E}$, and calling again $H_{s^{\pm}}$ the half-space "below" $s^{\pm}$, we deduce

$$\left|E \cap H_{s^-}\right| < \left|E \cap H_s\right| = \frac{1}{2} < \left|E \cap H_{s^+}\right|.$$

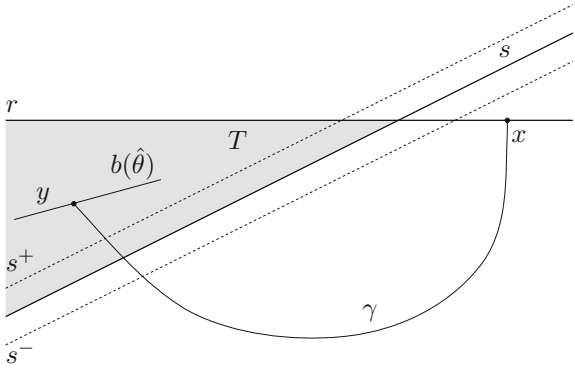


Fig. 4 The situation in Step III: the region T is shaded

Arguing exactly as in Lemma 2.3, we obtain then the existence and uniqueness of a bisecting line of direction θ for any $\theta \in (\bar{\theta} - \varepsilon, \bar{\theta} + \varepsilon)$, and the corresponding bisecting chord $b(\theta)$ must be connected to x, since it intersects the path γ. Since this is in contrast with the definition of $\bar{\theta}$, we have obtained $\bar{\theta} = \pi$ which—as noticed above—concludes this step.

Step IV. The set $\overline{E}$ is simply connected.

To conclude the proof of the theorem, we only need to check that the set $\overline{E}$ is simply connected. If this were not true, there would exist a closed curve $\gamma \subseteq \overline{E}$ enclosing some small ball $B \subseteq \mathbb{R}^2 \setminus \overline{E}$. Pick any point $x \in B$, and take any bisecting line r passing through x: by construction, each of the two halflines contained in r and having x as endpoint intersects γ, thus $\overline{E}$. Since this implies that $r \cap \overline{E}$ is not a segment, the contradiction comes from the fact that E is chord-convex. □

We can immediately observe two simple consequences of the above theorem.

Corollary 2.7 *The claim of Lemma 2.4 is valid for any $x \in T$, not only for the points $x \in T \cap E$.*

Proof Let us call, as in the proof of Lemma 2.4, H_r, H_s, H_{r_x} and H_{s_x} the four half-spaces determined by the lines r and s, and by the lines r_x and s_x parallel to r and s and passing through x.

Again, we know that $|E \cap H_{r_x}| \leq |E \cap H_r| = 1/2$ since $H_{r_x} \subseteq H_r$. There are now two possibilities: either $|E \cap H_{r_x}| = 1/2$, or $|E \cap H_{r_x}| < 1/2$. The first case can be excluded because otherwise r and r_x would be two different parallel bisecting lines, which is impossible by Theorem 2.6; then, $|E \cap H_{r_x}| < 1/2$, and in the very same way $|E \cap H_{s_x}| > 1/2$. The conclusion now follows exactly as in Lemma 2.4. □

Corollary 2.8 *In any chord-convex set E, every two bisecting chords intersect.*

Proof Suppose that there exist two bisecting lines, r and s, such that the corresponding bisecting chords $b(r)$ and $b(s)$ do not intersect. As in Fig. 5, let us then call T and $\widetilde{T}$ two of the regions in which r and s divide the plane, so that $b(r)$ and $b(s)$ belong to the closure of $\widetilde{T}$, and T is opposite to $\widetilde{T}$. Since both r and s are bisecting lines, we have $|T \cap E| = |\widetilde{T} \cap E|$.

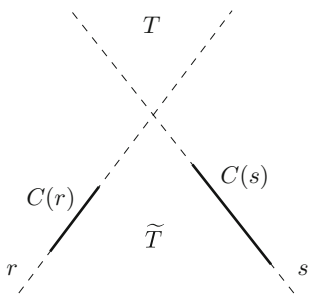


Fig. 5 Situation in Corollary 2.8

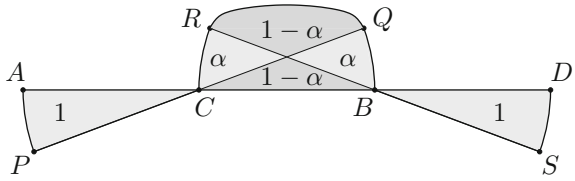


Fig. 6 Example 2.1: the set E is shaded

Now, if $|T \cap E| > 0$ then there is some $x \in T \cap E$, but this is impossible because this x would not be connected with the two chords: indeed, a curve connecting x with $b(r)$ should somewhere exit from the region T, and this would happen at some point in $r \backslash b(r)$, or in $s \backslash b(s)$, in contradiction with the definition of bisecting chords. On the other hand, if $|T \cap E| = 0$, then also $|\widetilde{T} \cap E| = 0$, and we would run into the same contradiction, because then the two chords $b(r)$ and $b(s)$ would not be connected with each other. □

In the above Theorem 2.6, to get the simple connectedness it was necessary to consider the closure $\overline{E}$ of E. In fact, there exist chord-convex sets which are not simply connected (but their closure is of course simply connected, by Theorem 2.6). An example is shown below.

Example Here we provide an example of a chord-convex set which is not simply connected. As in Fig. 6, let AD be a segment, and let the points C and B divide it in three equal parts. Then, let P and Q be two points on the circle centered at C and passing through A and B, such that the segment PQ passes through C. Analogously, let R and S be two points on the circle centered at B passing through C and D, so that the segment RS passes through B. Let finally E be the bounded set whose boundary is the union of the segments AC, BD, PC and BS, the arcs of circle $\widehat{AP}$, $\widehat{BQ}$, $\widehat{CR}$ and $\widehat{SD}$, and some convex curve connecting R and Q as in the Figure. Of course, a suitable choice of this curve and the other parameters allows us to consider that the different parts of E have size either 1, or α, or $1-\alpha$ for some $0 < \alpha < 1$, as indicated in Fig. 6. As a consequence, it is easy to see that this set E is chord-convex, but it is not connected (as already said, of course $\overline{E}$ is connected, according to Theorem 2.6).

2.2 Properties of the Extreme Points and of the Intersections Between Chords

Thanks to Theorem 2.6, we have the uniqueness of the bisecting chord for any direction. This allows us to give the following definition.

Definition 2.9 Let E be a chord-convex set. For any $\theta \in \mathbb{S}^1$, we denote by $r(\theta)$ the unique bisecting line of E with direction θ, and by $L(\theta)$, $M(\theta)$ and $R(\theta)$ the left extreme, the center and the right extreme, respectively, of the corresponding bisecting chord, that we denote by $b(\theta)$. As a consequence, $b(\theta) = r(\theta) \cap \overline{E} = [L(\theta), R(\theta)]$.

Even though the functions L, M, R take value in $\mathbb{R}^2$, throughout the paper we use the following abuse of notation: when a direction θ is specified and the points under consideration all belong to $r(\theta)$, then the line $r(\theta)$ is identified with $\mathbb{R}$, taken with direction θ.

Notice that of course

$$M(\theta+\pi) = M(\theta), \qquad L(\theta+\pi) = R(\theta), \qquad R(\theta+\pi) = L(\theta).$$

Moreover, for any $\theta \in \mathbb{S}^1$, we will call π_θ the projection on the bisecting line $r(\theta)$, and define

$$L^+(\theta) := \overline{\lim_{\alpha\to\theta}}\ \pi_\theta(L(\alpha)), \qquad R^-(\theta) := \underline{\lim_{\alpha\to\theta}}\ \pi_\theta(R(\alpha)),$$

where the limits make sense in view of the argument after Definition 2.9. Then, by definition,

$$L(\theta) = \underline{\lim_{\alpha\to\theta}}\ \pi_\theta(L(\alpha)), \qquad R(\theta) = \overline{\lim_{\alpha\to\theta}}\ \pi_\theta(R(\alpha)).$$

Proposition 2.10 *Let E be a chord-convex set of size 2. Then, for every $\theta \in \mathbb{S}^1$,*

$$\overline{L(\theta)R^-(\theta)} \geq 2, \qquad \overline{L^+(\theta)R(\theta)} \geq 2.$$

Proof By symmetry, it is enough to show the first inequality. Let us also assume for simplicity of notation that $\theta = 0$, and assume that $\overline{L(0)R^-(0)} < 2$. By definition of R^-, we can find directions ξ arbitrarily close to 0 with $\pi_0(R(\xi)) \leq R^-(0) + \varepsilon$; on the other hand, if ξ is close enough to 0, then $\pi_0(L(\xi)) \geq L(0) - \varepsilon$. Let then $\xi \ll 1$ be a direction for which both the inequalities hold: one has then

$$\overline{L(\xi)R(\xi)} = \frac{\pi_0(R(\xi)) - \pi_0(L(\xi))}{\cos\xi} \leq \frac{\overline{L(0)R^-(0)} + 2\varepsilon}{\cos\xi} < 2,$$

where the last inequality is true as soon as both ε and $|\xi|$ are small enough. This is in contradiction with the fact that the size of E is 2, and this concludes the thesis.□

Let us now prove that the intersection between any two bisecting chords is always in the "internal part" of both chords, that is, between L^+ and R^-.

Lemma 2.11 *Let E be a chord-convex set. Then, for any $\theta \neq \xi \in \mathbb{S}^1$, one has*

$$b(\theta) \cap b(\xi) \in [L^+(\theta), R^-(\theta)].$$

Proof Let us assume for simplicity that $\theta = 0$, and assume also that the claim is false. Hence, there exists some $\xi \neq 0$ such that $b(0) \cap b(\xi) \in [L(0), L^+(0))$. By definition of L^+, we can find an arbitrarily small α with $L(\alpha)$ very close to $L^+(0)$, in

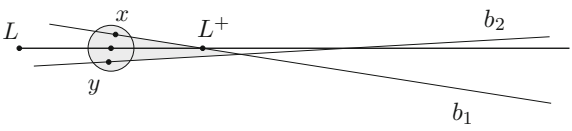


Fig. 7 The situation in Lemma 2.12

particular $\pi_0(L(\alpha)) > b(0) \cap b(\xi)$. Since we can choose such a direction satisfying $|\alpha| < |\xi|$, we obtain that the bisecting chords $b(\alpha)$ and $b(\xi)$ do not intersect, which is absurd by Corollary 2.8. □

We can now deduce that the segments $[L, L^+]$ belong to the boundary of E.

Lemma 2.12 *For any $\theta \in \mathbb{S}^1$, we have $b(\theta) \cap E \subseteq (L^+(\theta), R^-(\theta))$, which in particular implies that the interval $[L(\theta), L^+(\theta)]$ belongs to ∂E.*

Proof As usual, let us take $\theta = 0$ and write $L = L(0)$ and $L^+ = L^+(0)$ for simplicity of notations. By definition, it is clear that $L \in \partial E$, hence there is nothing to prove if $L = L^+$; moreover, since ∂E is closed, it is enough to exclude the presence of some point of E in the open interval (L, L^+). Suppose then that such a point exists, thus there exists some small ball contained in E and centered at a point in $(L, L^+) \cap E$, as in Fig. 7. Any point of this ball is of course contained in some bisecting chord; if we take points arbitrarily close to the center, we get that the corresponding bisecting chords become very close to be horizontal: otherwise, we would find a bisecting chord $b(\xi)$ for some $\xi \neq 0$ which intersects $b(0)$ in the center of the ball, in contradiction to Lemma 2.11.

Let us then take two points of the ball, x and y, respectively above and below $b(0)$, and consider two bisecting chords b_1 and b_2 passing through x and y, which have a negative and a positive direction, respectively, since they must intersect $b(0)$. Again by Lemma 2.11, we know that both these chords intersect $b(0)$ at some point in $[L^+, R^-]$; since $\overline{E}$ is simply connected by Theorem 2.6, we obtain that E contains the open region $\mathcal{R}$ shaded in Fig. 7, which is the union of the ball and the triangle with extremes x, y and the intersection between b_1 and b_2. Since L^+ is the limit of left extremes, it belongs to ∂E. Thus, at least one of the two chords b_1 and b_2 must pass through L^+, for instance the figure depicts a situation where $L^+ \in b_1$ but $L^+ \notin b_2$.

Let us use again the fact that L^+ is the limit of points $L(\theta_i)$ for a suitable sequence $\theta_i \to 0$. Take some θ_i such that $|\theta_i|$ is smaller than both the directions of b_1 and of b_2, and consider the point $L(\theta_i)$: it cannot belong to $\mathcal{R}$, because it belongs to ∂E while $\mathcal{R} \subseteq E$. On the other hand, if it does not belong to $\partial\mathcal{R}$ then the chord $b(\theta_i)$ cannot intersect both b_1 and b_2, which is absurd: we deduce that $L(\theta_i)$ belongs to one of the two segments xL^+ and yL^+ (in particular, one which is part of b_1 or b_2). However, this also leads to a contradiction, because since $|\theta_i| \ll 1$ then the bisecting line $r(\theta_i)$

would enter inside $\mathcal{R}$ before the left extreme $L(\theta_i)$, which is in contradiction with the definition of bisecting chord. We have thus concluded the proof. □

It is now useful to introduce some further concept.

Definition 2.13 A point $x \in [L(\theta), R(\theta)]$ is said to be *above* E (resp. *below* E) if there is a ball $B(x, \rho)$ of center x and radius ρ such that $B(x, \rho) \cap H_{r(\theta)} \subseteq \overline{E}$ (resp. $B(x, \rho) \cap H^c_{r(\theta)} \subseteq \overline{E}$). An interval $I \subseteq [L(\theta), R(\theta)]$ is said to be above E (resp. below E) if is made of points which are all above E (resp. below E).

Lemma 2.14 *Let E be a chord-convex set and assume that for some $\theta \in \mathbb{S}^1$ there exists a point $z \in [L(\theta), L^+(\theta))$ which is below E. Then the whole segment $(z, L^+(\theta))$ is below E, and there is no point in $[L(\theta), L^+(\theta))$ which is above E.*

Proof The proof follows with the very same argument as in Lemma 2.12. Indeed, assume as usual that $\theta = 0$ and take a point $z \in [L(0), L^+(0))$ below E: by definition, this means that there is a small ball centered at z whose upper half ball belongs to E. As in the proof of Lemma 2.12, we can take some point x in this ball very close to z, so that a bisecting chord passing through x must have a negative slope close to 0 and cross $b(0)$ in a point which belongs to $[L^+(0), R^-(0)]$. As a consequence, the simple connectedness of $\overline{E}$ given by Theorem 2.6 ensures that the whole open triangle $xzL^+(0)$ is contained in E, and this implies that every point of the segment $(z, L^+(0))$ is below E, which concludes the first part of the proof.

The second part easily follows: if some point in $[L(0), L^+(0))$ were above E, by the first part there would be points in $[L(0), L^+(0))$ which are at the same time above and below E, hence inside E. And in turn, this gives a contradiction to Lemma 2.12, because the whole segment $[L(0), L^+(0)]$ must be in ∂E. □

We can now characterize the intersection $b(\xi) \cap E$ for any angle ξ. Let us be more precise: fix for simplicity $\xi = 0$, and write again L, L^+, R^- and R in place of $L(0)$, $L^+(0)$, $R^-(0)$ and $R(0)$. Define then, for any $\theta \neq 0$, $P(\theta)$ as the intersection between $b(0)$ and $b(\theta)$; moreover, define the following points in $[L^+, R^-]$:

$$Q_l^+ := \overline{\lim_{\theta \nearrow 0}} P(\theta), \quad Q_l^- := \underline{\lim_{\theta \nearrow 0}} P(\theta), \quad Q_r^+ := \overline{\lim_{\theta \searrow 0}} P(\theta), \quad Q_r^- := \underline{\lim_{\theta \searrow 0}} P(\theta).$$

Arguing exactly as in Lemmas 2.12 and 2.14, we can prove that

$$\begin{aligned} &(L^+, Q_l^+) \text{ and } (Q_r^-, R^-) \text{ are below } E, \\ &\text{while } (L^+, Q_r^+) \text{ and } (Q_l^-, R^-) \text{ are above } E. \end{aligned} \tag{2.1}$$

Indeed, for any $\varepsilon > 0$ we can find some $0 < -\theta \ll 1$ such that $\pi_0(L(\theta)) < L^+ + \varepsilon$ and $\pi_0(P(\theta)) > Q_l^+ - \varepsilon$. Since the function $\beta(\theta) = \pi_0^{-1}(L^+ + \varepsilon) \cap b(\theta)$ is well defined and continuous near 0, we deduce that the vertical segment connecting $L^+ + \varepsilon$ to $\beta(\theta)$ is entirely contained in E, thus again Theorem 2.6 ensures that the triangle of vertices $L^+ + \varepsilon$, $\beta(\theta)$ and $P(\theta)$ is inside E, and in turn this implies that all the

points of $(L^+ + \varepsilon, Q_l^+ - \varepsilon)$ are below E. Now, by letting $\varepsilon \to 0$, we get that the interval (L^+, Q_l^+) is below E. The very same argument shows also the claims about the other three intervals, so (2.1) is established. As a consequence, if we set

$$Q^+ = \min\{Q_l^+,\ Q_r^+\}, \qquad\qquad Q^- = \max\{Q_l^-,\ Q_r^-\},$$

we know that E contains the open intervals (L^+, Q^+) and (Q^-, R^-). There are now two possible cases: if $Q^+ > Q^-$, then the whole interval (L^+, Q^-) is inside E. Instead, if $Q^+ \le Q^-$, then we only know that E contains (L^+, Q^+) and (Q^-, R^-); the points in (Q^+, Q^-) are then all above E but not necessarily below E (if $Q_l^- \le Q_l^+ = Q^+ < Q^- = Q_r^- \le Q_r^+$), or all below E but not necessarily above E (if $Q_r^- \le Q_r^+ = Q^+ < Q^- = Q_l^- \le Q_l^+$). All these observations become particularly useful in a specific case, namely, if the functions L and R are continuous at $\theta = 0$: indeed, in this case obviously $L = L^+$ and $R = R^-$, and it easily follows that the four points $Q_{r,l}^{\pm}$ coincide all with the middle point of $b(0)$ (this follows from Lemma 2.16 below). We can then summarize what we found in the following result.

Lemma 2.15 *Let E be a chord-convex set such that the functions L and R are continuous. Then, the interior of any bisecting chord $b(\theta)$ is contained inside E, except possibly the middle point $M(\theta)$.*

Let us now show what we just mentioned, that is, the intersection between bisecting chords converges to their middle point when L and R are continuous.

Lemma 2.16 *Let E be a chord-convex set such that L and R are continuous. Then, for any $\theta \in \mathbb{S}^1$, the point $b(\theta) \cap b(\xi)$ converges to $M(\theta)$ when $\xi \to \theta$.*

Proof Let us call ℓ the length of the chord $b(\theta)$. For any $\xi \in \mathbb{S}^1$, since both $b(\theta)$ and $b(\xi)$ are bisecting chords, we know that $|T \cap E| = |T' \cap E|$, where T and T' are two opposite regions in which $\mathbb{R}^2$ is divided by the two lines $r(\theta)$ and $r(\xi)$. In particular, let $\xi = \theta + \eta$ be very close to θ, and let T and T' be the two opposite "small" regions, that is, those corresponding to the small corner $\eta \ll 1$. Since L and R are continuous, we know that the extremes of any bisecting chord of direction between θ and ξ are at most a distance ε apart from those of $b(\theta)$; as a consequence, if the point $b(\theta) \cap b(\xi)$ is at distances d and $\ell - d$ from the extremes of $b(\theta)$, we have

$$\frac{(d-\varepsilon)^2}{2}\,|\eta| \le |T \cap E| \le \frac{(d+\varepsilon)^2}{2}\,|\eta|,$$
$$\frac{(\ell-d-\varepsilon)^2}{2}\,|\eta| \le |T' \cap E| \le \frac{(\ell-d+\varepsilon)^2}{2}\,|\eta|,$$

and it follows that d converges to $\ell/2$ when $\eta \to 0$, that is the thesis. □

We conclude this section with an important result, which states that the intersection point between any two bisecting chords cannot be an extreme point for both of them.

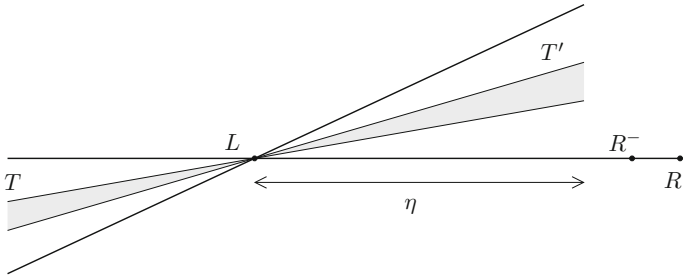


Fig. 8 The situation in Theorem 2.17

Theorem 2.17 *Let E be a chord-convex set. Then, two bisecting chords cannot intersect at a point which is an extreme point for both of them.*

Proof Let us suppose that the claim is false. In particular, we can assume that $L := L(0) = L(\theta)$ for some $0 < \theta < \pi$ (if for such a θ one has $L(0) = R(\theta)$ then a very similar argument would work).

Then we observe that, for every $0 < \xi < \theta$, the bisecting chord $b(\xi)$ must pass through L, since it must cross both $b(0)$ and $b(\theta)$. By definition of R^-, we can now take some $0 < \bar{\xi} < \theta$ such that $\pi_0(R(\xi)) > R^- - \varepsilon$ for every $0 < \xi < \bar{\xi}$. As a consequence, for every $0 < \xi < \bar{\xi}$ we have that

$$\overline{LR(\xi)} \geq \overline{L\pi_0(R(\xi))} > R^- - L - \varepsilon > 2 - \varepsilon,$$

where the last inequality follows by Proposition 2.10, assuming without loss of generality that the size of E is 2. For brevity, we set $\eta := R^- - L - \varepsilon > 0$. Let us now fix any two directions $0 < \xi_1 < \xi_2 < \bar{\xi}$, and call T and T' the two regions determined by the bisecting lines $r(\xi_1)$ and $r(\xi_2)$, as in Fig. 8. By construction we have

$$|E \cap T| = |E \cap T'| \geq \frac{\eta^2(\xi_2 - \xi_1)}{2},$$

and this implies that there is some point $x \in E \cap T$ with $\pi_0(x) < L - \eta$. By construction, a bisecting chord passing through x must have direction between 0 and θ; but then, it must pass through L, and so its direction is actually between ξ_1 and ξ_2. Summarizing, for any choice of $0 < \xi_1 < \xi_2 < \theta$ we have found a direction $\xi \in (\xi_1, \xi_2)$ such that $\pi_0(L(\xi)) \leq L - \eta$. If we now let both ξ_1 and ξ_2 go to 0, so does also ξ, and this implies that

$$L = \lim_{\xi \to 0} \pi_0(L(\xi)) \leq L - \eta,$$

which is absurd. This concludes the proof. □

Notice that the above theorem does not state that the intersection point of two bisecting chords is always in the interior of both of them, but only that it cannot be an extreme for both of them. For instance, in the case of the Zindler flower of Fig. 1, the left extreme $L(\pi/2)$ is the intersection point of $b(\pi/2)$ with $b(0)$: in particular, this point is the left extreme of a bisecting chord, and the middle point of the other one. We can immediately observe that this is always the case, at least when L and R are continuous functions.

Corollary 2.18 *Let E be a chord-convex set such that L and R are continuous. Then, if the intersection of two bisecting chords is an extreme point of one of them, it must be the middle point of the other one.*

Proof This immediately follows from Theorem 2.17 and Lemma 2.15. Indeed, assume that $x := b(\theta) \cap b(\xi)$ coincides with $L(\theta)$ for some $\theta \neq \xi \in \mathbb{S}^1$. Theorem 2.17 ensures that x is neither $L(\xi)$ nor $R(\xi)$, hence $x \in (L(\xi), R(\xi))$. On the other hand, $x \in \partial E$, and the only point of the interval $(L(\xi), R(\xi))$ which can be in ∂E is the middle point, according to Lemma 2.15. □

2.3 Zindler Sets and Their Properties

In this last section we define the Zindler sets and we prove their main properties.

Definition 2.19 Let E be a chord-convex set. We say that E is a *Zindler set* if all the bisecting chords have the same length.

As we said in the Introduction, the Zindler sets play an important role in the problem of minimizing the area among the *convex* set of given size: roughly speaking, it is very easy to guess (but hard to show!) that a minimizer of the area must be a Zindler set. In fact, the proof of the minimality of the Auerbach triangle was done in two steps, by different authors: it was first proved that the Auerbach triangle minimizes the area among the (convex) Zindler sets [3], and then that a minimizer among the convex sets (which trivially exists by compactness) must be Zindler [2].

The general situation of the chord-convex sets that we are considering now seems even more complicated, though there are common points. First of all, it is not obvious whether a minimizer of the area among the chord-convex sets exists, since the class of the chord-convex sets is not compact, while so are the convex sets. Moreover, it is again extremely reasonable to guess that, if a minimizer exists, then it must be a Zindler set: we are not able to show this result in full generality, but we can prove a particular case in Theorem 2.23.

We can immediately observe that, for a Zindler set, the functions L and R are automatically continuous, hence in particular Lemma 2.15 and Corollary 2.18 always apply for a Zindler set.

Lemma 2.20 *Let E be a Zindler set. Then, the functions L and R are continuous.*

Proof This readily follows from Proposition 2.10. Indeed, assuming without loss of generality that the size of E is 2, for any $\theta \in \mathbb{S}^1$ we know on one side that $\overline{L(\theta)R(\theta)} = 2$ because E is a Zindler set, and on the other hand that $\overline{L(\theta)R^-(\theta)} \geq 2$ by Proposition 2.10. It follows that $R^-(\theta) = R(\theta)$, and similarly that $L^+(\theta) = L(\theta)$. By definition of L^+ and R^-, the continuity of L and R trivially follows. □

Now, thanks to Corollary 2.18, for a Zindler set we have that the intersection between two bisecting chords has two possibilities: either it is an internal point of both chords, or it is at the same time an extreme point of one of them and the middle point of the other one. Of course the second case is more peculiar, and we will call "edge angle" any of the two directions. More precisely, it is useful to give the next definition.

Definition 2.21 For any chord-convex set E, the sets $\mathcal{L}$, $\mathcal{R}$, $\mathcal{ML}$ and $\mathcal{MR}$ are defined as

$$\begin{aligned}
\mathcal{L} &:= \{\, \theta \in \mathbb{S}^1 : \exists\, \eta \in [\theta - \pi/2, \theta + \pi/2] \text{ such that } L(\theta) = M(\eta) \,\},\\
\mathcal{R} &:= \{\, \theta \in \mathbb{S}^1 : \exists\, \eta \in [\theta - \pi/2, \theta + \pi/2] \text{ such that } R(\theta) = M(\eta) \,\},\\
\mathcal{ML} &:= \{\, \theta \in \mathbb{S}^1 : \exists\, \eta \in [\theta - \pi/2, \theta + \pi/2] \text{ such that } M(\theta) = L(\eta) \,\},\\
\mathcal{MR} &:= \{\, \theta \in \mathbb{S}^1 : \exists\, \eta \in [\theta - \pi/2, \theta + \pi/2] \text{ such that } M(\theta) = R(\eta) \,\}.
\end{aligned}$$

If θ belongs to any of the above sets, we call it an *edge angle*.

We can immediately notice a technical property of the intervals which are contained in one of the above sets. We state it for $\mathcal{ML}$, but of course the analogous results for the other sets are also valid.

Lemma 2.22 *Let E be a chord-convex set such that L and R are continuous, assume that $I \subseteq \mathcal{ML}$ for some interval $I \subseteq \mathbb{S}^1$, and define $\psi : I \to \mathbb{S}^1$ the function such that $M(\theta) = L(\psi(\theta))$ for any $\theta \in I$. Then, the function ψ is decreasing.*

Proof First of all, observe that the function ψ is well-defined, since by Theorem 2.17 it is not possible that two different directions have the same left extreme. Let us assume without loss of generality that $I = [0, \bar{\theta})$ and $0 < \psi(0) < \pi$.

We claim that for any $\theta \in (0, \bar{\theta})$ the middle point $M(\theta)$ is below $r(0)$: indeed, since the function ψ is clearly continuous, then $0 < \psi(\theta) < \pi$, and then if $M(\theta) = L(\psi(\theta))$ is above $r(0)$ the two bisecting chords $b(0)$ and $b(\psi(\theta))$ do not intersect, which is absurd by Corollary 2.8. Analogously, since $b(\theta)$ must intersect $b(\psi(0))$, then by construction $b(\theta) \cap b(\psi(0)) \in (M(\theta), R(\theta))$. Finally, since $M(\theta) = L(\psi(\theta))$ and $b(\psi(\theta))$ must intersect $b(\psi(0))$, then $0 < \psi(\theta) < \psi(0)$ holds. The monotonicity of the function ψ then follows. □

As we said above, it is reasonable to expect that the intersection of two bisecting chords is usually an internal point for both of them, and that the edge angles are quite rare: for instance, the Zindler flower of Fig. 1 has six edge angles (corresponding to three "bad" pairs of chords), and a simple modification—namely, a flower with n

petals instead of 3—gives an example with $2n$ edge angles. In particular, if the edge angles are finite or countably many, we can show that a minimizer of the area—if it exists—must be a Zindler set. It is actually enough something even weaker, namely, that the edge angles do not fill any open interval.

Theorem 2.23 *Assume that E minimizes the area among the chord-convex sets of size 2. Assume in addition that L and R are continuous, and that the directions which are not edge angles are dense (for instance, this is true if the edge angles have zero length in $\mathbb{S}^1$). Then, E is a Zindler set.*

Proof Let us assume that E is not a Zindler set: then, there must be a direction such that the corresponding bisecting chord has length strictly more than 2. Actually, the same remains true for all the directions in a small neighborhood, because L and R are continuous, and so is then also the length of the bisecting chords. Since the non-edge angles are dense, we can then assume the existence of a direction (say 0) which is a non-edge angle and for which the bisecting chord has length $2\ell > 2 + 6a$, for some strictly positive a.

Let us now apply Lemma 2.16 to get the continuity of the function $\tau : \mathbb{S}^1 \times \mathbb{S}^1 \to \mathbb{R}^2$ defined as $\tau(\theta, \xi) = b(\theta) \cap b(\xi)$ for $\theta \neq \xi$, and $\tau(\theta, \theta) = M(\theta)$. Since 0 is not an edge angle, $L(0)$ and $R(0)$ do not belong to the image of τ, hence we can assume, possibly taking a smaller a, that

$$\tau(\theta, \xi) \notin B\big(L(0), 6a\big) \cup B\big(R(0), 6a\big) \qquad \forall\, \theta,\ \xi \in \mathbb{S}^1. \tag{2.2}$$

By the continuity of L and R, there exists $\bar\theta > 0$ such that

$$\max\left\{\overline{L(\theta)L(0)},\ \overline{R(\theta)R(0)}\right\} < a \qquad \forall\, \theta \in (-\bar\theta, \bar\theta). \tag{2.3}$$

Let us now call $x = \tau(-\bar\theta, \bar\theta)$: again by Lemma 2.16, up to taking a smaller $\bar\theta$, we also have

$$\overline{xM(0)} < a. \tag{2.4}$$

Let us now consider the four regions in which $\mathbb{R}^2$ is subdivided by the bisecting lines $r(\bar\theta)$ and $r(-\bar\theta)$, and call T and T' the two small ones, corresponding to the directions between $-\bar\theta$ and $\bar\theta$: as usual, we know that $|T \cap E| = |T' \cap E|$. Putting together (2.3) and (2.4), we have that for any $\theta \in (-\bar\theta, \bar\theta)$ both the points $L(\theta)$ and $R(\theta)$ have distance at least $\ell - 2a$ from x. Thus, recalling that $\overline{E}$ is simply connected by Theorem 2.6,

$$B(x, \ell - 2a) \cap \big(T \cup T'\big) \subseteq E. \tag{2.5}$$

We now define the competitor set

$$\widetilde{E} := \big(E \backslash (T \cup T')\big) \cup \big(B(x, \ell - 3a) \cap (T \cup T')\big).$$

By (2.5) we know that $\widetilde{E}$ is strictly contained in E, so it has a strictly smaller volume: we will conclude the proof by showing that $\widetilde{E}$ is also a chord-convex set of size at least 2.

First of all, observe that by construction $|T \cap \widetilde{E}| = |T' \cap \widetilde{E}|$, hence

$$\left|\left(E \backslash \widetilde{E}\right) \cap T\right| = \left|\left(E \backslash \widetilde{E}\right) \cap T'\right|.$$

This implies that the lines $r(\bar{\theta})$ and $r(-\bar{\theta})$ are bisecting lines also for $\widetilde{E}$, and in turn this ensures that, for every $\theta \in (-\bar{\theta}, \bar{\theta})$, the unique bisecting line of direction θ is the one crossing x, whose intersection with $\widetilde{E}$ is a segment of length $2(\ell - 3a) > 2$. To conclude that $\widetilde{E}$ is chord-convex and has size at least 2, it is then sufficient to show that for any $\theta \notin (-\bar{\theta}, \bar{\theta})$ the line $r(\theta)$ is a bisecting line also for $\widetilde{E}$, and its intersection with the closure of $\widetilde{E}$ coincides with $b(\theta)$ (and it is then a segment of length at least 2). Actually, since we already checked that $r(\pm\bar{\theta})$ are bisecting lines for $\widetilde{E}$, it is enough to consider directions $\theta \notin [-\bar{\theta}, \bar{\theta}]$.

Let then θ be such an angle; notice that the intersection of $r(\theta)$ with the region $T \cup T'$ is a segment PQ, and the points P and Q coincide by definition with $\tau(\theta, \bar{\theta})$ and $\tau(\theta, -\bar{\theta})$, so they are both in $b(\theta)$ and in $b(\pm\bar{\theta})$. By (2.2) and a trivial geometrical argument, both P and Q are inside the ball $B(x, \ell - 3a)$, so the segment PQ is entirely inside $\widetilde{E}$ and the proof is concluded. □

References

1. H. Auerbach, Sur un problème de M. Ulam concernant l'équilibre des corps flottants. Studi. Math. **7**, 121–142 (1938)
2. L. Esposito, V. Ferone, B. Kawohl, C. Nitsch, C. Trombetti, The longest shortest fence and sharp Poincaré–sobolev inequalities. Arch. Ration. Mech. Anal. **206**(3), 821–851 (2012)
3. N. Fusco, A. Pratelli, On a conjecture by Auerbach. J. Eur. Math. Soc. **13**(6), 1633–1676 (2011)
4. K. Zindler, Über konvexe Gebilde, II. Teil Monatsh. Math. Phys. **31**, 25–56 (1921)

Optimization Problems Involving the First Dirichlet Eigenvalue and the Torsional Rigidity

Michiel van den Berg, Giuseppe Buttazzo and Bozhidar Velichkov

Abstract We present some open problems and obtain some partial results for spectral optimization problems involving measure, torsional rigidity and first Dirichlet eigenvalue.

Keywords Torsional rigidity · Dirichlet eigenvalues · Spectral optimization

Mathematics Subject Classification (2010) Primary 49J45 · 49R05 · Secondary 35P15 · 47A75 · 35J25

1 Introduction

A shape optimization problem can be written in the very general form

$$\min\big\{F(\Omega)\ :\ \Omega\in\mathcal{A}\big\},$$

where $\mathcal{A}$ is a class of admissible domains and F is a cost functional defined on $\mathcal{A}$. We consider in the present paper the case where the cost functional F is related to the solution of an elliptic equation and involves the spectrum of the related elliptic operator. We speak in this case of *spectral optimization problems*. Shape optimization problems of spectral type have been widely considered in the literature; we mention

M. van den Berg (✉)
School of Mathematics, University of Bristol University Walk, Bristol Bs8 1tw, UK
e-mail: mamvdb@bristol.ac.uk

G. Buttazzo
Dipartimento di Matematica, Università di Pisa Largo B. Pontecorvo 5, 56127 Pisa, Italy
e-mail: buttazzo@dm.unipi.it

B. Velichkov
Laboratoire Jean Kuntzmann (LJK), Université Joseph Fourier Tour
IRMA, BP 53, 51 Rue des Mathématiques, 38041 Grenoble Cedex 9, France
e-mail: bozhidar.velichkov@imag.fr
URL: http://www.velichkov.it

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_2

for instance the papers [7, 9, 10, 12–15, 22], and we refer to the books [8, 19, 20], and to the survey papers [2, 11, 18], where the reader can find a complete list of references and details.

In the present paper we restrict ourselves for simplicity to the Laplace operator $-\Delta$ with Dirichlet boundary conditions. Furthermore we shall assume that the admissible domains Ω are a priori contained in a given *bounded* domain $D \subset \mathbb{R}^d$. This assumption greatly simplifies several existence results that otherwise would require additional considerations in terms of concentration-compactness arguments [7, 32].

The most natural constraint to consider on the class of admissible domains is a bound on their Lebesgue measure. Our admissible class $\mathcal{A}$ is then

$$\mathcal{A} = \big\{\Omega \subset D \ : \ |\Omega| \leq 1\big\}.$$

Other kinds of constraints are also possible, but we concentrate here to the one above, referring the reader interested in possible variants to the books and papers quoted above.

The following two classes of cost functionals are the main ones considered in the literature.

Integral functionals. Given a right-hand side $f \in L^2(D)$, for every $\Omega \in \mathcal{A}$ let u_Ω be the unique solution of the elliptic PDE

$$-\Delta u = f \text{ in } \Omega, \qquad u \in H_0^1(\Omega).$$

The integral cost functionals are of the form

$$F(\Omega) = \int_\Omega j\big(x, u_\Omega(x), \nabla u_\Omega(x)\big)\, dx,$$

where j is a suitable integrand that we assume convex in the gradient variable. We also assume that j is bounded from below by

$$j(x, s, z) \geq -a(x) - c|s|^2,$$

with $a \in L^1(D)$ and c smaller than the first Dirichlet eigenvalue of the Laplace operator $-\Delta$ in D. For instance, the energy $\mathcal{E}_f(\Omega)$ defined by

$$\mathcal{E}_f(\Omega) = \inf\left\{\int_D \Big(\frac{1}{2}|\nabla u|^2 - f(x)u\Big)\, dx \ : \ u \in H_0^1(\Omega)\right\},$$

belongs to this class since, integrating by parts its Euler-Lagrange equation, we have that

$$\mathcal{E}_f(\Omega) = -\frac{1}{2}\int_D f(x)u_\Omega\, dx,$$

which corresponds to the integral functional above with

$$j(x, s, z) = -\frac{1}{2} f(x)s.$$

The case $f = 1$ is particularly interesting for our purposes. We denote by w_Ω the *torsion function*, that is the solution of the PDE

$$-\Delta u = 1 \text{ in } \Omega, \qquad u \in H_0^1(\Omega),$$

and by the *torsional rigidity* $T(\Omega)$ the L_1 norm of w_Ω,

$$T(\Omega) = \int_\Omega w_\Omega \, dx = -2\mathcal{E}_1(\Omega).$$

Spectral functionals. For every admissible domain $\Omega \in \mathcal{A}$ we consider the spectrum $\Lambda(\Omega)$ of the Laplace operator $-\Delta$ on $H_0^1(\Omega)$. Since Ω has a finite measure, the operator $-\Delta$ has a compact resolvent and so its spectrum $\Lambda(\Omega)$ is discrete:

$$\Lambda(\Omega) = \big(\lambda_1(\Omega), \lambda_2(\Omega), \dots\big),$$

where $\lambda_k(\Omega)$ are the eigenvalues counted with their multiplicity. The spectral cost functionals we may consider are of the form

$$F(\Omega) = \Phi\big(\Lambda(\Omega)\big),$$

for a suitable function $\Phi : \mathbb{R}^{\mathbb{N}} \to \overline{\mathbb{R}}$. For instance, taking $\Phi(\Lambda) = \lambda_k(\Omega)$ we obtain

$$F(\Omega) = \lambda_k(\Omega).$$

We take the torsional rigidity $T(\Omega)$ and the first eigenvalue $\lambda_1(\Omega)$ as prototypes of the two classes above and we concentrate our attention on cost functionals that depend on both of them. We note that, by the maximum principle, when Ω increases $T(\Omega)$ increases, while $\lambda_1(\Omega)$ decreases.

2 Statement of the Problem

The optimization problems we want to consider are of the form

$$\min\left\{\Phi\big(\lambda_1(\Omega), T(\Omega)\big) \ : \ \Omega \subset D, \ |\Omega| \le 1\right\}, \tag{2.1}$$

where we have normalized the constraint on the Lebesgue measure of Ω, and where Φ is a given continuous (or lower semi-continuous) and non-negative function. In

the rest of this paper we often take for simplicity $D = \mathbb{R}^d$, even if most of the results are valid in the general case. For instance, taking $\Phi(a, b) = ka + b$ with k a fixed positive constant, the quantity we aim to minimize becomes

$$k\lambda_1(\Omega) + T(\Omega) \quad \text{with} \quad \Omega \subset D, \quad \text{and} \ |\Omega| \leq 1.$$

Remark 2.1 If the function $\Phi(a, b)$ is increasing with respect to a and decreasing with respect to b, then the cost functional

$$F(\Omega) = \Phi\big(\lambda_1(\Omega), T(\Omega)\big)$$

turns out to be decreasing with respect to the set inclusion. Since both the torsional rigidity and the first eigenvalue are γ-continuous functionals and the function Φ is assumed lower semi-continuous, we can apply the existence result of [13], which provides the existence of an optimal domain.

In general, if the function Φ does not verify the monotonicity property of Remark 2.1, then the existence of an optimal domain is an open problem, and the aim of this paper is to discuss this issue. For simplicity of the presentation we limit ourselves to the two-dimensional case $d = 2$. The case of general d does not present particular difficulties but requires the use of several d-dependent exponents.

Remark 2.2 The following facts are well known.

(i) If B is a disk in $\mathbb{R}^2$ we have

$$T(B) = \frac{1}{8\pi}|B|^2.$$

(ii) If $j_{0,1} \approx 2.405$ is the first positive zero of the Bessel functions $J_0(x)$ and B is a disk of $\mathbb{R}^2$ we have

$$\lambda_1(B) = \frac{\pi}{|B|} j_{0,1}^2.$$

(iii) The torsional rigidity $T(\Omega)$ scales as

$$T(t\Omega) = t^4 T(\Omega), \qquad \forall t > 0.$$

(iv) The first eigenvalue $\lambda_1(\Omega)$ scales as

$$\lambda_1(t\Omega) = t^{-2}\lambda_1(\Omega), \qquad \forall t > 0.$$

(v) For every domain Ω of $\mathbb{R}^2$ and any disk B we have

$$|\Omega|^{-2}T(\Omega) \leq |B|^{-2}T(B) = \frac{1}{8\pi}.$$

(vi) For every domain Ω of $\mathbb{R}^2$ and any disk B we have (Faber-Krahn inequality)

$$|\Omega|\lambda_1(\Omega) \geq |B|\lambda_1(B) = \pi j_{0,1}^2.$$

(vii) A more delicate inequality is the so-called Kohler-Jobin inequality (see [21], [3]): for any domain Ω of $\mathbb{R}^2$ and any disk B we have

$$\lambda_1^2(\Omega)T(\Omega) \geq \lambda_1^2(B)T(B) = \frac{\pi}{8} j_{0,1}^4.$$

This had been previously conjectured by G. Pólya and G.Szegö [23].

We recall the following inequality, well known for planar regions (Sect. 5.4 in [23]), between torsional rigidity and first eigenvalue.

Proposition 2.3 *For every domain $\Omega \subset \mathbb{R}^d$ we have*

$$\lambda_1(\Omega)T(\Omega) \leq |\Omega|.$$

Proof By definition, $\lambda_1(\Omega)$ is the infimum of the Rayleigh quotient

$$\int_\Omega |\nabla u|^2\,dx \Big/ \int_\Omega u^2\,dx \qquad \text{over all } u \in H_0^1(\Omega),\ u \neq 0.$$

Taking as u the torsion function w_Ω, we have

$$\lambda_1(\Omega) \leq \int_\Omega |\nabla w_\Omega|^2\,dx \Big/ \int_\Omega w_\Omega^2\,dx.$$

Since $-\Delta w_\Omega = 1$, an integration by parts gives

$$\int_\Omega |\nabla w_\Omega|^2\,dx = \int_\Omega w_\Omega\,dx = T(\Omega),$$

while the Hölder inequality gives

$$\int_\Omega w_\Omega^2\,dx \geq \frac{1}{|\Omega|}\left(\int_\Omega w_\Omega\,dx\right)^2 = \frac{1}{|\Omega|}\big(T(\Omega)\big)^2.$$

Summarizing, we have

$$\lambda_1(\Omega) \leq \frac{|\Omega|}{T(\Omega)}$$

as required. □

Remark 2.4 The infimum of $\lambda_1(\Omega)T(\Omega)$ over open sets Ω of prescribed measure is zero. To see this, let Ω_n be the disjoint union of one ball of volume $1/n$ and $n(n-1)$

balls of volume $1/n^2$. Then the radius R_n of the ball of volume $1/n$ is $(n\omega_d)^{-1/d}$ while the radius r_n of the balls of volume $1/n^2$ is $(n^2\omega_d)^{-1/d}$, so that $|\Omega_n| = 1$,

$$\lambda_1(\Omega_n) = \lambda_1(B_{R_n}) = \frac{1}{R_n^2}\lambda_1(B_1) = (n\omega_d)^{2/d}\lambda_1(B_1),$$

and

$$\begin{aligned} T(\Omega_n) &= T(B_{R_n}) + n(n-1)T(B_{r_n}) = T(B_1)\left(R_n^{d+2} + n(n-1)r_n^{d+2}\right) \\ &= T(B_1)\omega_d^{-1-2/d}\left(n^{-1-2/d} + (n-1)n^{-1-4/d}\right). \end{aligned}$$

Therefore

$$\lambda_1(\Omega_n)T(\Omega_n) = \frac{\lambda_1(B_1)T(B_1)}{\omega_d}\frac{n^{2/d}+n-1}{n^{1+2/d}},$$

which vanishes as $n \to \infty$.

In the next section we investigate the inequality of Proposition 2.3.

3 A Sharp Inequality Between Torsion and First Eigenvalue

We define the constant

$$\mathcal{K}_d = \sup\left\{\frac{\lambda_1(\Omega)T(\Omega)}{|\Omega|} \ : \ \Omega \text{ open in } \mathbb{R}^d, \ |\Omega| < \infty\right\}.$$

We have seen in Proposition 2.3 that $\mathcal{K}_d \leq 1$. The question is if the constant 1 can be improved.

Consider a ball B; performing the shape derivative as in [20], keeping the volume of the perturbed shapes constant, we obtain for every field $V(x)$

$$\begin{aligned} \partial[\lambda_1(B)T(B)](V) &= T(B)\partial[\lambda_1(B)](V) + \lambda_1(B)\partial[T(B)](V) \\ &= C_B\int_{\partial B} V\cdot n\, d\mathcal{H}^{d-1} \end{aligned}$$

for a suitable constant C_B. Since the volume of the perturbed shapes is constant, we have

$$\int_{\partial B} V\cdot n\, d\mathcal{H}^{d-1} = 0,$$

where $\mathcal{H}^{d-1}$ denotes $(d-1)$-dimensional Hausdorff measure. This shows that balls are stationary for the functional

$$F(\Omega) = \frac{\lambda_1(\Omega)T(\Omega)}{|\Omega|}.$$

Below we will show, by considering rectangles, that balls are not optimal. To do so we shall obtain a lower bound for the torsional rigidity of a rectangle.

Proposition 3.1 *In a rectangle* $R_{a,b} = (-b/2, b/2) \times (-a/2, a/2)$ *with* $a \le b$ *we have*

$$T(R_{a,b}) \ge \frac{a^3 b}{12} - \frac{11a^4}{180}.$$

Proof Let us estimate the energy

$$\mathcal{E}_1(R_{a,b}) = \inf\left\{\int_{R_{a,b}} \left(\frac{1}{2}|\nabla u|^2 - u\right) dx\, dy \ : \ u \in H_0^1(R_{a,b})\right\}$$

by taking the function

$$u(x, y) = \frac{a^2 - 4y^2}{8}\theta(x),$$

where $\theta(x)$ is defined by

$$\theta(x) = \begin{cases} 1, & \text{if } |x| \le (b-a)/2 \\ (b - 2|x|)/a, & \text{otherwise.} \end{cases}$$

We have

$$|\nabla u|^2 = \left(\frac{a^2 - 4y^2}{8}\right)^2 |\theta'(x)|^2 + y^2|\theta(x)|^2,$$

so that

$$\begin{aligned}
\mathcal{E}_1(R_{a,b}) &\le 2\int_0^{a/2} \left(\frac{a^2 - 4y^2}{8}\right)^2 dy \int_0^{b/2} |\theta'(x)|^2\, dx \\
&\quad + 2\int_0^{a/2} y^2\, dy \int_0^{b/2} |\theta(x)|^2\, dx \\
&\quad - 4\int_0^{a/2} \frac{a^2 - 4y^2}{8}\, dy \int_0^{b/2} \theta(x)\, dx \\
&= \frac{a^4}{60} + \frac{a^3}{12}\left(\frac{b-a}{2} + \frac{a}{6}\right) - \frac{a^3}{6}\left(\frac{b-a}{2} + \frac{a}{4}\right) \\
&= -\frac{a^3 b}{24} + \frac{11a^4}{360}.
\end{aligned}$$

The desired inequality follows since $T(R_{a,b}) = -2\mathcal{E}_1(R_{a,b})$. □

In d-dimensions we have the following.

Proposition 3.2 *If $\Omega_\varepsilon = \omega \times (-\varepsilon/2, \varepsilon/2)$, where ω is a convex set in $\mathbb{R}^{d-1}$ with $|\omega| < \infty$, then*

$$T(\Omega_\varepsilon) = \frac{\varepsilon^3}{12}|\omega| + O(\varepsilon^4), \qquad \epsilon \downarrow 0.$$

We defer the proof to Sect. 5.

For a ball of radius R we have

$$\lambda_1(B) = \frac{j^2_{d/2-1,1}}{R^2}, \qquad T(B) = \frac{\omega_d R^{d+2}}{d(d+2)}, \qquad |B| = \omega_d R^d, \tag{3.1}$$

so that

$$F(B) = \frac{\lambda_1(B)T(B)}{|B|} = \frac{j^2_{d/2-1,1}}{d(d+2)} := \alpha_d$$

For instance, we have

$$\alpha_2 \approx 0.723, \qquad \alpha_3 \approx 0.658, \qquad \alpha_4 \approx 0.612.$$

Moreover, since $j_{\nu,1} = \nu + O(\nu^{1/3})$, $\nu \to \infty$, we have that $\lim_{d\to\infty} \alpha_d = \frac{1}{4}$. A plot of α_d is given in Fig. 1.

We now consider a slab $\Omega_\varepsilon = \omega \times (0, \varepsilon)$ of thickness $\varepsilon \to 0$. We have by separation of variables and Proposition 3.2 that

$$\lambda_1(\Omega_\varepsilon) = \frac{\pi^2}{\varepsilon^2} + \lambda_1(\omega) \approx \frac{\pi^2}{\varepsilon^2}, \qquad T(\Omega_\varepsilon) \approx \frac{\varepsilon^3|\omega|}{12}, \qquad |\Omega_\varepsilon| = \varepsilon|\omega|,$$

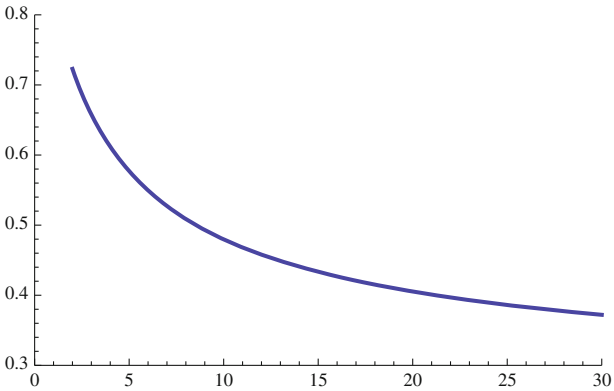


Fig. 1 The plot of α_d for $2 \le d \le 30$

so that

$$F(\Omega_\varepsilon) \approx \frac{\pi^2}{12} \approx 0.822.$$

This shows that in any dimension the slab is better than the ball. Using domains in $\mathbb{R}^d$ with k small dimensions and $d-k$ large dimensions does not improve the value of the cost functional F. In fact, if ω is a convex domain in $\mathbb{R}^{d-k}$ and $B_k(\varepsilon)$ a ball in $\mathbb{R}^k$, then by Theorem 5.1 with $\Omega_\varepsilon = \omega \times B_k(\varepsilon)$ we have that

$$\lambda_1(\Omega_\varepsilon) \approx \frac{1}{\varepsilon^2}\lambda_1\big(B_k(1)\big), \quad T(\Omega_\varepsilon) \approx \varepsilon^{k+2}|\omega|T(B_k(1)), \quad |\Omega_\varepsilon| = \varepsilon^k|\omega||B_k(1)|,$$

so that

$$F(\Omega_\varepsilon) \approx \frac{j^2_{k/2-1,1}}{k(k+2)} \le \frac{\pi^2}{12}.$$

This supports the following.

Conjecture 3.3 For any dimension d we have $\mathcal{K}_d = \pi^2/12$, and no domain in $\mathbb{R}^d$ maximizes the functional F for $d > 1$. The maximal value $\mathcal{K}_d$ is asymptotically reached by a thin slab $\Omega_\varepsilon = \omega \times (0, \varepsilon)$, with $\omega \subset \mathbb{R}^{d-1}$, as $\varepsilon \to 0$.

4 The Attainable Set

In this section we bound the measure by $|\Omega| \le 1$. Our goal is to plot the subset of $\mathbb{R}^2$ whose coordinates are the eigenvalue $\lambda_1(\Omega)$ and the torsion $T(\Omega)$. It is convenient to change coordinates and to set for any admissible domain Ω,

$$x = \lambda_1(\Omega), \qquad y = \big(\lambda_1(\Omega)T(\Omega)\big)^{-1}.$$

In addition, define

$$E = \left\{(x, y) \in \mathbb{R}^2 \ : \ x = \lambda_1(\Omega),\ y = \big(\lambda_1(\Omega)T(\Omega)\big)^{-1} \text{for some } \Omega, \quad |\Omega| \le 1\right\}.$$

Therefore, the optimization problem (2.1) can be rewritten as

$$\min\left\{\Phi\big(x, 1/(xy)\big) \ : \ (x, y) \in E\right\}.$$

Conjecture 4.1 The set E is closed.

We remark that the conjecture above, if true, would imply the existence of a solution of the optimization problem (2.1) for many functions Φ. Below we will analyze the variational problem in case $\Phi(x, y) = kx + \frac{1}{xy}$, where $k > 0$.

Theorem 4.2 *Let $d = 2, 3, \ldots,$ and let*

$$k_d^* = \frac{1}{2d\omega_d^{4/d} j_{d/2-1,1}^2}.$$

Consider the optimization problem

$$\min \{k\lambda_1(\Omega) + T(\Omega) \ : \ |\Omega| \leq 1\}. \tag{4.1}$$

If $0 < k \leq k_d^*$ then the ball with radius

$$R_k = \left(\frac{2kdj_{d/2-1,1}^2}{\omega_d}\right)^{1/(d+4)} \tag{4.2}$$

is the unique minimizer (modulo translations and sets of capacity 0). If $k > k_d^*$ then the ball B with measure 1 is the unique minimizer.

Proof Consider the problem (4.1) without the measure constraint

$$\min \left\{k\lambda_1(\Omega) + T(\Omega) \ : \ \Omega \subset \mathbb{R}^d\right\}. \tag{4.3}$$

Taking $t\Omega$ instead of Ω gives that

$$k\lambda_1(t\Omega) + T(t\Omega) = kt^{-2}\lambda_1(\Omega) + t^{d+2}T(\Omega).$$

The optimal t which minimizes this expression is given by

$$t = \left(\frac{2k\lambda_1(\Omega)}{(d+2)T(\Omega)}\right)^{1/(d+4)}.$$

Hence (4.3) equals

$$\min\left\{(d+4)\left(\frac{k^{d+2}}{4(d+2)^{d+2}}T^2(\Omega)\lambda_1^{d+2}(\Omega)\right)^{1/(d+4)} \ : \ \Omega \subset \mathbb{R}^d\right\}. \tag{4.4}$$

By the Kohler-Jobin inequality in $\mathbb{R}^d$, the minimum in (4.4) is attained by any ball. Therefore the minimum in (4.3) is given by a ball B_R such that

$$\left(\frac{2k\lambda_1(B_R)}{(d+2)T(B_R)}\right)^{1/(d+4)} = 1.$$

This gives (4.2). We conclude that the measure constrained problem (4.1) admits the ball B_{R_k} as a solution whenever $\omega_d R_k^d \leq 1$. That is $k \leq k_d^*$.

Next consider the case $k > k_d^*$. Let B be the open ball with measure 1. It is clear that

$$\min\{k\lambda_1(\Omega) + T(\Omega) \ : \ |\Omega| \le 1\} \le k\lambda_1(B) + T(B).$$

To prove the converse we note that for $k > k_d^*$,

$$\begin{aligned} \min \left\{k\lambda_1(\Omega) + T(\Omega) \ : \ |\Omega| \le 1\right\} & \\ \ge \min \left\{(k - k_d^*)\lambda_1(\Omega) \ : \ |\Omega| \le 1\right\} & \\ + \min \left\{k_d^*\lambda_1(\Omega) + T(\Omega) \ : \ |\Omega| \le 1\right\}. & \end{aligned} \tag{4.5}$$

The minimum in the first term in the right hand side of (4.5) is attained for B by Faber-Krahn, whereas the minimum in second term is attained for $B_{R_{k_d^*}}$ by our previous unconstrained calculation. Since $|B_{R_{k_d^*}}| = |B| = 1$ we have by (4.5) that

$$\begin{aligned} \min \left\{k\lambda_1(\Omega) + T(\Omega) \ : \ |\Omega| \le 1\right\} & \\ \ge (k - k_d^*)\lambda_1(B) + k_d^*\lambda_1(B) + T(B) & \\ = k\lambda_1(B) + T(B). & \end{aligned}$$

Uniqueness of the above minimizers follows by uniqueness of Faber-Krahn and Kohler-Jobin. □

It is interesting to replace the first eigenvalue in (4.1) be a higher eigenvalue. We have the following for the second eigenvalue.

Theorem 4.3 *Let $d = 2, 3, \ldots,$ and let*

$$l_d^* = \frac{1}{2d(2\omega_d)^{4/d} j_{d/2-1,1}^2}.$$

Consider the optimization problem

$$\min \{l\lambda_2(\Omega) + T(\Omega) \ : \ |\Omega| \le 1\}. \tag{4.6}$$

If $0 < l \le l_d^*$ then the union of two disjoint balls with radii

$$R_l = \left(\frac{ldj_{d/2-1,1}^2}{\omega_d}\right)^{1/(d+4)} \tag{4.7}$$

is the unique minimizer (modulo translations and sets of capacity 0).

If $l > l_d^*$ then union of two disjoint balls with measure $1/2$ each is the unique minimizer.

Proof First consider the unconstrained problem

$$\min\left\{l\lambda_1(\Omega)+T(\Omega)\ :\ \Omega\subset\mathbb{R}^d\right\}. \tag{4.8}$$

Taking $t\Omega$ instead of Ω gives that

$$l\lambda_2(t\Omega)+T(t\Omega)=lt^{-2}\lambda_2(\Omega)+t^{d+2}T(\Omega).$$

The optimal t which minimizes this expression is given by

$$t=\left(\frac{2l\lambda_2(\Omega)}{(d+2)T(\Omega)}\right)^{1/(d+4)}.$$

Hence (4.8) equals

$$\min\left\{(d+4)\left(\frac{l^{d+2}}{4(d+2)^{d+2}}T^2(\Omega)\lambda_2^{d+2}(\Omega)\right)^{1/(d+4)}\ :\ \Omega\subset\mathbb{R}^d\right\}. \tag{4.9}$$

It follows by the Kohler-Jobin inequality, see for example Lemma 6 in [31], that the minimizer of (4.9) is attained by the union of two disjoint balls B_R and B_R' with the same radius. Since $\lambda_2(B_R\cup B_R')=\lambda_1(B_R)$ and $T(B_R\cup B_R')=2T(B_R)$ we have, using (3.1), that the radii of these balls are given by (4.7). We conclude that the measure constrained problem (4.6) admits the union of two disjoint balls with equal radius R_l as a solution whenever $2\omega_d R_l^d\le 1$. That is $l\le l_d^*$.

Next consider the case $l>l_d^*$. Let Ω be the union of two disjoint balls B and B' with measure $1/2$ each. Then

$$\min\{l\lambda_2(\Omega)+T(\Omega)\ :\ |\Omega|\le 1\}\le l\lambda_1(B)+2T(B).$$

To prove the converse we note that for $l>l_d^*$,

$$\begin{aligned}\min\left\{l\lambda_2(\Omega)+T(\Omega)\ :\ |\Omega|\le 1\right\}&\\ \ge\min\left\{(l-l_d^*)\lambda_2(\Omega)\ :\ |\Omega|\le 1\right\}&\\ +\min\left\{l_d^*\lambda_2(\Omega)+T(\Omega)\ :\ |\Omega|\le 1\right\}.&\end{aligned} \tag{4.10}$$

The minimum in the first term in the right hand side of (4.10) is attained for $B\cup B'$ by the Krahn-Szegö inequality, whereas the minimum in second term is attained for the union of two disjoint balls with radius $R_{l_d^*}$ by our previous unconstrained calculation. Since $|B_{R_{l_d^*}}|=1/2=|B|=|B'|$ we have by (4.10) that

$$\begin{aligned}\min\{l\lambda_2(\Omega)+T(\Omega)\ :\ |\Omega|\le 1\}&\ge(l-l_d^*)\lambda_1(B)+l_d^*\lambda_1(B)+2T(B)\\ &=l\lambda_1(B)+2T(B).\end{aligned}$$

Uniqueness of the above minimizers follows by uniqueness of Krahn-Szegö and Kohler-Jobin for the second eigenvalue. □

To replace the first eigenvalue in (4.1) be the jth eigenvalue ($j > 2$) is a very difficult problem since we do not know the minimizers of the jth Dirichlet eigenvalue with a measure constraint nor the minimizer of the jth Dirichlet eigenvalue a torsional rigidity constraint. However, if these two problems have a common minimizer then information similar to the above can be obtained.

Putting together the facts listed in Remark 2.2 we obtain the following inequalities.

(i) By Faber-Krahn inequality we have $x \geq \pi j_{0,1}^2 \approx 18.168$.
(ii) By Conjecture 3.3 (if true) we have $y \geq 12/\pi^2 \approx 1.216$.
(iii) By the bound on the torsion of Remark 2.2 v) we have $xy \geq 8\pi \approx 25.133$.
(iv) By the Kohler-Jobin inequality we have $y/x \leq 8/(\pi j_{0,1}^4) \approx 0.076$.
(v) The set E is *conical*, that is if a point (x_0, y_0) belongs to E, then all the half-line $\{(tx_0, ty_0) : t \geq 1\}$ in contained in E. This follows by taking $\Omega_t = \Omega/t$ and by the scaling properties iii) and iv) of Remark 2.2.
(vi) The set E is *vertically convex*, that is if a point (x_0, y_0) belongs to E, then all points (x_0, ty_0) with $1 \leq t \leq 8/(\pi j_{0,1}^4)$ belong to E. To see this fact, let Ω be a domain corresponding to the point $(x_0, y_0) \in E$. The *continuous Steiner symmetrization* path Ω_t (with $t \in [0, 1]$) then continuously deforms the domain $\Omega = \Omega_0$ into a ball $B = \Omega_1$, preserving the Lebesgue measure and decreasing $\lambda_1(\Omega_t)$ (see [5] where this tool has been developed, and Sect. 6.3 of [8] for a short survey). The curve

$$x(t) = \lambda_1(\Omega_t), \qquad y(t) = \big(\lambda_1(\Omega_t)T(\Omega_t)\big)^{-1}$$

then connects the point (x_0, y_0) to the Kohler-Jobin line $\{y = 8x/(\pi j_{0,1}^4)\}$, having $x(t)$ decreasing. Since $\big(x(t), y(t)\big) \in E$, the conicity of E then implies vertical convexity.

A plot of the constraints above is presented in Fig. 2. Some particular cases can be computed explicitly. Consider $d = 2$, and let

$$\Omega = B_R \cup B_r, \text{ with } B_R \cap B_r = \emptyset, r \leq R, \text{ and } \pi(R^2 + r^2) = 1.$$

An easy computation gives that

$$\lambda_1(\Omega) = \frac{j_{0,1}^2}{R^2}, \qquad T(\Omega) = \frac{2\pi^2 R^4 - 2\pi R^2 + 1}{8\pi},$$

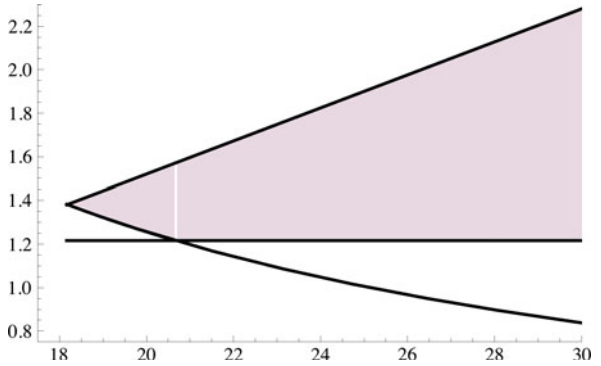


Fig. 2 The admissible region E is contained in the *dark* area

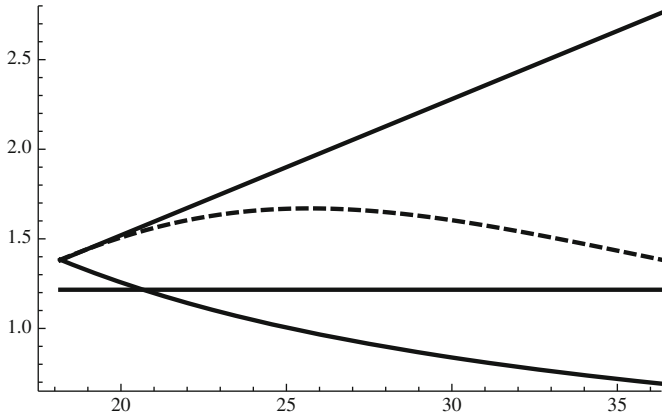


Fig. 3 The *dashed line* corresponds to two disks of variable radii

so that the curve

$$y = \frac{8\pi x}{x^2 - 2\pi j_{0,1}^2 x + 2\pi^2 j_{0,1}^4}, \qquad \pi j_{0,1}^2 \le x \le 2\pi j_{0,1}^2$$

is contained in E (see Fig. 3).

If we consider the rectangle

$$\Omega = (0, b) \times (0, a) \text{ with } a \le b, \text{ and } ab = 1,$$

we have by Proposition 3.1

$$\lambda_1(\Omega) = \pi^2\left(\frac{1}{a^2}+\frac{1}{b^2}\right) = \pi^2\left(\frac{1}{a^2}+a^2\right),$$
$$T(\Omega) \geq \frac{a^3 b}{12} - \frac{11a^4}{180} = \frac{a^2}{12} - \frac{11a^4}{180}.$$

Therefore $y \leq h\big(x/(2\pi^2)\big)$, where

$$h(t) = \frac{90}{\pi^2 t\left(11+15t-22t^2-(15+2t)\sqrt{t^2-1}\right)}, \qquad t \geq 1.$$

By E being conical the curve

$$y = h\big(x/(2\pi^2)\big), \qquad \pi^2 \leq x < +\infty$$

is contained in E (see Fig. 4).

Besides the existence of optimal domains for problem (2.1), the regularity of optimal shapes is another very delicate and important issue. Very little is known about the regularity of optimal domains for spectral optimization problems (see for instance [4, 6, 17, 32]); the cases where only the first eigenvalue $\lambda_1(\Omega)$ and the torsion $T(\Omega)$ are involved could be simpler and perhaps allow to use the free boundary methods developed in [1].

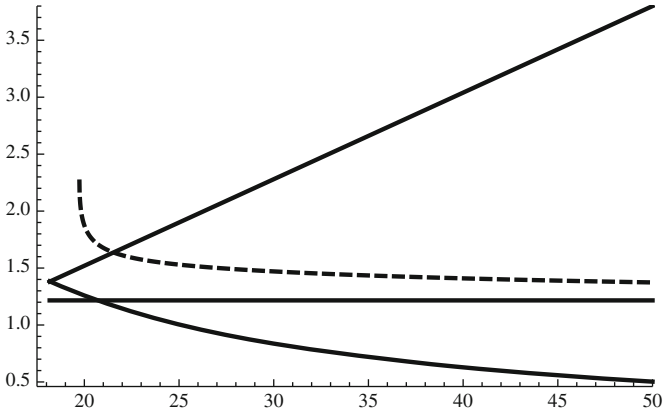


Fig. 4 The *dashed line* is an upper bound to the line corresponding to rectangles

5 Torsional Rigidity and the Heat Equation

It is well known that the rich interplay between elliptic and parabolic partial differential equations provide tools for obtaining results in one field using tools from the other. See for example the monograph by E. B. Davies [16], and [24–29] for some more recent results. In this section we use some heat equation tools to obtain new estimates for the torsional rigidity. Before we do so we recall some basic facts relating the torsional rigidity to the heat equation. For an open set Ω in $\mathbb{R}^d$ with boundary $\partial\Omega$ we denote the Dirichlet heat kernel by $p_\Omega(x, y; t),\ x \in \Omega,\ y \in \Omega,\ t > 0$. So

$$u_\Omega(x; t) := \int_\Omega p_\Omega(x, y; t)\, dy,$$

is the unique weak solution of

$$\begin{cases} \dfrac{\partial u}{\partial t} = \Delta u & x \in \Omega,\ t > 0, \\ \lim_{t \downarrow 0} u(x; t) = 1 & \text{in } L^2(\Omega), \\ u(x; t) = 0 & x \in \partial\Omega,\ t > 0. \end{cases}$$

The latter boundary condition holds at all regular points of $\partial\Omega$. We denote the heat content of Ω at time t by

$$Q_\Omega(t) = \int_\Omega u_\Omega(x; t)\, dx.$$

Physically the heat content represents the amount of heat in Ω at time t if Ω has initial temperature 1, while $\partial\Omega$ is kept at temperature 0 for all $t > 0$. Since the Dirichlet heat kernel is non-negative, and monotone in Ω we have that

$$0 \le p_\Omega(x, y; t) \le p_{\mathbb{R}^d}(x, y; t) = (4\pi t)^{-d/2} e^{-|x-y|^2/(4t)}. \tag{5.1}$$

It follows by either (5.1) or by the maximum principle that

$$0 \le u_\Omega(x; t) \le 1,$$

and that if $|\Omega| < \infty$ then

$$0 \le Q_\Omega(t) \le |\Omega|. \tag{5.2}$$

In the latter situation we also have an eigenfunction expansion for the Dirichlet heat kernel in terms of the Dirichlet eigenvalues $\lambda_1(\Omega) \le \lambda_2(\Omega) \le \dots$, and a corresponding orthonormal set of eigenfunctions $\{\varphi_1, \varphi_2, \dots\}$,

$$p_\Omega(x, y; t) = \sum_{j=1}^{\infty} e^{-t\lambda_j(\Omega)} \varphi_j(x) \varphi_j(y).$$

We note that the eigenfunctions are in $L^p(\Omega)$ for all $1 \leq p \leq \infty$. It follows by Parseval's formula that

$$\begin{aligned} Q_\Omega(t) &= \sum_{j=1}^{\infty} e^{-t\lambda_j(\Omega)} \left(\int_\Omega \varphi_j \, dx \right)^2 \\ &\leq e^{-t\lambda_1(\Omega)} \sum_{j=1}^{\infty} \left(\int_\Omega \varphi_j \, dx \right)^2 \\ &= e^{-t\lambda_1(\Omega)} |\Omega|. \end{aligned} \tag{5.3}$$

Since the torsion function is given by

$$w_\Omega(x) = \int_0^\infty u_\Omega(x; t) \, dt,$$

we have that

$$T(\Omega) = \sum_{j=1}^{\infty} \lambda_j(\Omega)^{-1} \left(\int_\Omega \varphi_j \, dx \right)^2.$$

We recover Proposition 2.3 by integrating (5.3) with respect to t over $[0, \infty)$:

$$T(\Omega) \leq \lambda_1(\Omega)^{-1} \sum_{j=1}^{\infty} \left(\int_\Omega \varphi_j \, dx \right)^2 = \lambda_1(\Omega)^{-1} |\Omega|.$$

Let M_1 and M_2 be two open sets in Euclidean space with finite Lebesgue measures $|M_1|$ and $|M_2|$ respectively. Let $M = M_1 \times M_2$. We have that

$$p_{M_1 \times M_2}(x, y; t) = p_{M_1}(x_1, y_1; t) p_{M_2}(x_2, y_2; t),$$

where $x = (x_1, x_2)$, $y = (y_1, y_2)$. It follows that

$$Q_M(t) = Q_{M_1}(t) Q_{M_2}(t), \tag{5.4}$$

and

$$T(M) = \int_0^\infty Q_{M_1}(t) Q_{M_2}(t) \, dt. \tag{5.5}$$

Integrating (5.4) with respect to t, and using (5.2) for M_2 we obtain that

$$T(M) \leq T(M_1)|M_2|. \tag{5.6}$$

This upper bound should be "sharp" if the decay of $Q_{M_2}(t)$ with respect to t is much slower than the decay of $Q_{M_1}(t)$. The result below makes this assertion precise

in the case where M_2 is a convex set with $\mathcal{H}^{d_2-1}(\partial M_2) < \infty$. The latter condition is for convex sets equivalent to requiring that M_2 is bounded. Here $\mathcal{H}^{d_2-1}$ denotes the (d_2-1)-dimensional Hausdorff measure.

Theorem 5.1 *Let $M = M_1 \times M_2$, where M_1 is an arbitrary open set in $\mathbb{R}^{d_1}$ with finite d_1-measure and M_2 is a bounded convex open set in $\mathbb{R}^{d_2}$. Then there exists a constant $\mathcal{C}_{d_2}$ depending on d_2 only such that*

$$T(M) \geq T(M_1)|M_2| - \mathcal{C}_{d_2}\lambda_1(M_1)^{-3/2}|M_1|\mathcal{H}^{d_2-1}(\partial M_2). \tag{5.7}$$

For the proof of Theorem 5.1 we need the following lemma (proved as Lemma 6.3 in [30]).

Lemma 5.2 *For any open set Ω in $\mathbb{R}^d$,*

$$u_\Omega(x;t) \geq 1 - 2\int_{\{y\in\mathbb{R}^d:|y-x|>d(x)\}} p_{\mathbb{R}^d}(x,y;t)\,dy, \tag{5.8}$$

where

$$d(x) = \min\{|x-z| \ : \ z \in \partial\Omega\}.$$

Proof of Theorem 5.1 With the notation above we have that

$$\begin{aligned} T(M) &= T(M_1)|M_2| - \int_0^\infty Q_{M_1}(t)(|M_2| - Q_{M_2}(t))\,dt \\ &= T(M_1)|M_2| - \int_0^\infty Q_{M_1}(t)\int_{M_2}(1-u_{M_2}(x_2;t))\,dx_2\,dt. \end{aligned}$$

Define for $r > 0$,

$$\partial M_2(r) = \{x \in M_2 : d(x) = r\}.$$

It is well known that (Proposition 2.4.3 in [8]) if M_2 is convex then

$$\mathcal{H}^{d_2-1}(\partial M_2(r)) \leq \mathcal{H}^{d_2-1}(\partial M_2). \tag{5.9}$$

By (5.3), (5.8) and (5.9) we obtain that

$$\begin{aligned} &\int_0^\infty Q_{M_1}(t)\int_{M_2}(1-u_{M_2}(x_2;t))\,dx_2\,dt \\ &\quad\leq 2|M_1|\mathcal{H}^{d_2-1}(\partial M_2)\int_0^\infty dt\,e^{-t\lambda_1(M_1)}\int_0^\infty dr\int_{\{z\in\mathbb{R}^{d_2}:|z-x|>r\}} p_{\mathbb{R}^{d_2}}(x,z;t)\,dz \\ &\quad= 2d_2\omega_{d_2}|M_1|\mathcal{H}^{d_2-1}(\partial M_2)\int_0^\infty dt\,e^{-t\lambda_1(M_1)}(4\pi t)^{-d_2/2}\int_0^\infty dr\,r^{d_2}e^{-r^2/(4t)} \\ &\quad= \mathcal{C}_{d_2}\lambda_1(M_1)^{-3/2}|M_1|\mathcal{H}^{d_2-1}(\partial M_2), \end{aligned} \tag{5.10}$$

where

$$\mathcal{C}_{d_2} = \frac{\pi^{1/2} d_2 \Gamma((d_2+1)/2)}{\Gamma((d_2+2)/2)}.$$

This concludes the proof. □

Proof of Proposition 3.2 Let $M_1 = (0, \epsilon) \subset \mathbb{R}$, $M_2 = \omega \subset \mathbb{R}^{d-1}$. Since the torsion function for M_1 is given by $x(\epsilon - x)/2$, $0 \le x \le \epsilon$ we have that $T(M_1) = \epsilon^3/12$. Then (5.6) proves the upper bound. The lower bound follows from (5.7) since $\lambda_1(M_1) = \pi^2/\epsilon^2$, $|M_1| = \epsilon$. □

It is of course possible, using the Faber-Krahn inequality for $\lambda_1(M_1)$, to obtain a bound for the right-hand side of (5.10) in terms of the quantity $|M_1|^{(d_1+3)/d_1} \mathcal{H}^{d_2-1}(\partial M_2)$.

Our next result is an improvement of Proposition 3.1. The torsional rigidity for a rectangle follows by substituting the formulae for $Q_{(0,a)}(t)$ and $Q_{(0,b)}(t)$ given in (5.12) below into (5.5). We recover the expression given on p.108 in [23]:

$$T(R_{a,b}) = \frac{64ab}{\pi^6} \sum_{k=1,3,\ldots} \sum_{l=1,3,\ldots} k^{-2} l^{-2} \left(\frac{k^2}{a^2} + \frac{l^2}{b^2} \right)^{-1}.$$

Nevertheless the following result is not immediately obvious.

Theorem 5.3

$$\left| T(R_{a,b}) - \frac{a^3 b}{12} + \frac{31\zeta(5) a^4}{2\pi^5} \right| \le \frac{a^5}{15b}, \tag{5.11}$$

where

$$\zeta(5) = \sum_{k=1}^{\infty} \frac{1}{k^5}.$$

Proof A straightforward computation using the eigenvalues and eigenfunctions of the Dirichlet Laplacian on the interval together with the first identity in (5.3) shows that

$$Q_{(0,a)}(t) = \frac{8a}{\pi^2} \sum_{k=1,3,\ldots} k^{-2} e^{-t\pi^2 k^2/a^2}. \tag{5.12}$$

We write

$$Q_{(0,b)}(t) = b - \frac{4t^{1/2}}{\pi^{1/2}} + \left(Q_{(0,b)}(t) + \frac{4t^{1/2}}{\pi^{1/2}} - b \right). \tag{5.13}$$

The constant term b in the right-hand side of (5.13) gives, using (5.12), a contribution

$$\begin{aligned}\frac{8ab}{\pi^2}\int_{[0,\infty)} dt \sum_{k=1,3,\dots} k^{-2}e^{-t\pi^2k^2/a^2} &= \frac{8a^3b}{\pi^4}\sum_{k=1,3,\dots} k^{-4}\\ &= \frac{8a^3b}{\pi^4}\left(\sum_{k=1}^{\infty}k^{-4} - \sum_{k=2,4,\dots} k^{-4}\right) = \frac{15a^3b}{2\pi^4}\zeta(4)\\ &= \frac{a^3b}{12},\end{aligned}$$

which jibes with the corresponding term in (5.11). In a very similar calculation we have that the $-\frac{4t^{1/2}}{\pi^{1/2}}$ term in the right-hand side of (5.13) contributes

$$-\frac{32a}{\pi^{5/2}}\int_{[0,\infty)} dt\, t^{1/2} \sum_{k=1,3,\dots} k^{-2}e^{-t\pi^2k^2/a^2} = -\frac{31\zeta(5)a^4}{2\pi^5},$$

which jibes with the corresponding term in (5.11). It remains to bound the contribution from the expression in the large round brackets in (5.11). Applying formula (5.12) to the interval $(0, b)$ instead and using the fact that $\sum_{k=1,3,\dots} k^{-2} = \pi^2/8$ gives that

$$\begin{aligned}Q_{(0,b)}(t) - b + \frac{4t^{1/2}}{\pi^{1/2}} &= \frac{8b}{\pi^2}\sum_{k=1,3,\dots} k^{-2}\left(e^{-t\pi^2k^2/b^2} - 1\right) + \frac{4t^{1/2}}{\pi^{1/2}}\\ &= -\frac{8}{b}\sum_{k=1,3,\dots}\int_{[0,t]} d\tau e^{-\tau\pi^2k^2/b^2} + \frac{4t^{1/2}}{\pi^{1/2}}\\ &= -\frac{8}{b}\int_{[0,t]} d\tau\left(\sum_{k=1}^{\infty} e^{-\tau\pi^2k^2/b^2} - \sum_{k=1}^{\infty} e^{-4\tau\pi^2k^2/b^2}\right)\\ &\quad + \frac{4t^{1/2}}{\pi^{1/2}}.\end{aligned} \tag{5.14}$$

In order to bound the right-hand side of (5.14) we use the following instance of the Poisson summation formula.

$$\sum_{k\in\mathbb{Z}} e^{-t\pi k^2} = t^{-1/2}\sum_{k\in\mathbb{Z}} e^{-\pi k^2/t},\ t > 0.$$

We obtain that

$$\sum_{k=1}^{\infty} e^{-t\pi k^2} = \frac{1}{(4t)^{1/2}} - \frac{1}{2} + t^{-1/2}\sum_{k=1}^{\infty} e^{-\pi k^2/t},\ t > 0.$$

Applying this identity twice (with $t = \pi\tau/b^2$ and $t = 4\pi\tau/b^2$ respectively) gives that the right-hand side of (5.14) equals

$$-\frac{8}{\pi^{1/2}}\int_{[0,t]} d\tau\left(\tau^{-1/2}\sum_{k=1}^{\infty} e^{-k^2b^2/\tau} - (4\tau)^{-1/2}\sum_{k=1}^{\infty} e^{-k^2b^2/(4\tau)}\right).$$

Since $k \mapsto e^{-k^2b^2/\tau}$ is non-negative and decreasing,

$$\sum_{k=1}^{\infty} \tau^{-1/2} e^{-k^2b^2/\tau} \le \tau^{-1/2}\int_{[0,\infty)} dk e^{-k^2b^2/\tau} = \pi^{1/2}(2b)^{-1}.$$

It follows that

$$\left|Q_{(0,b)}(t) - b + \frac{4t^{1/2}}{\pi^{1/2}}\right| \le \frac{8t}{b},\ t > 0.$$

So the contribution of the third term in (5.13) to $T(R_{a,b})$ is bounded in absolute value by

$$\begin{aligned}\frac{64a}{\pi^2 b}\int_{[0,\infty)} dt\, t \sum_{k=1,3,\ldots} k^{-2} e^{-t\pi^2k^2/a^2} &= \frac{64a^5}{\pi^6 b}\sum_{k=1,3,\ldots} k^{-6}\\ &= \frac{63a^5}{\pi^6 b}\zeta(6)\\ &= \frac{a^5}{15b}.\end{aligned}$$

This completes the proof of Theorem 5.3. □

The Kohler-Jobin theorem mentioned in Sect. 2 generalizes to d-dimensions: for any open set Ω with finite measure the ball minimizes the quantity $T(\Omega)\lambda_1(\Omega)^{(d+2)/2}$. Moreover, in the spirit of Theorem 5.1, the following inequality is proved in [31] through an elementary heat equation proof.

Theorem 5.4 *If $T(\Omega) < \infty$ then the spectrum of the Dirichlet Laplacian acting in $L^2(\Omega)$ is discrete, and*

$$T(\Omega) \ge \left(\frac{2}{d+2}\right)\left(\frac{4\pi d}{d+2}\right)^{d/2}\sum_{k=1}^{\infty}\lambda_k(\Omega)^{-(d+2)/2}.$$

We obtain, using the Ashbaugh-Benguria theorem (p.86 in [19]) for $\lambda_1(\Omega)/\lambda_2(\Omega)$, that

$$T(\Omega)\lambda_1(\Omega)^{(d+2)/2} \geq \left(\frac{2}{d+2}\right)\left(\frac{4\pi d}{d+2}\right)^{d/2}\Gamma\left(1+\frac{d}{2}\right)\left(1+\left(\frac{\lambda_1(B)}{\lambda_2(B)}\right)^{(d+2)/2}\right). \tag{5.15}$$

The constant in the right-hand side of (5.15) is for $d = 2$ off by a factor $\frac{j_{0,1}^4 j_{1,1}^4}{8(j_{0,1}^4+j_{1,1}^4)} \approx 3.62$ if compared with the sharp Kohler-Jobin constant. We also note the missing factor $m^{m/(m+2)}$ in the right-hand side of (57) in [31].

Acknowledgments A large part of this paper was written during a visit of the first two authors at the Isaac Newton Institute for Mathematical Sciences of Cambridge (UK). GB and MvdB gratefully acknowledge the Institute for the excellent working atmosphere provided. The authors also wish to thank Pedro Antunes helpful discussions. The work of GB is part of the project 2010A2TFX2 *"Calcolo delle Variazioni"* funded by the Italian Ministry of Research and University.

References

1. H.W. Alt, L.A. Caffarelli, Existence and regularity for a minimum problem with free boundary. J. Reine. Angew. Math. **325**, 105–144 (1981)
2. M.S. Ashbaugh, Open problems on eigenvalues of the Laplacian, in *Analytic and Geometric Inequalities and Applications*. Mathematics and Its Applications (Kluwer Academic Publisher, Dordrecht, 1999), pp. 13–28
3. L. Brasco, On torsional rigidity and principal frequencies: an invitation to the Kohler-Jobin rearrangement technique. ESAIM Control Optim. Calc. Var. **20**(2), 315–338 (2014)
4. T. Briançon, J. Lamboley, Regularity of the optimal shape for the first eigenvalue of the Laplacian with volume and inclusion constraints. Ann. Inst. H. Poincaré Anal. Non Linéaire **26**, 1149–1163 (2009)
5. F. Brock, Continuous Steiner symmetrization. Math. Nachr. **172**, 25–48 (1995)
6. D. Bucur, Regularity of optimal convex shapes. J. Convex Anal. **10**, 501–516 (2003)
7. D. Bucur, Minimization of the kth eigenvalue of the Dirichlet Laplacian. Arch. Ration. Mech. Anal. **206**, 1073–1083 (2012)
8. D. Bucur, G. Buttazzo, *Variational Methods in Shape Optimization Problems*. Progress in Nonlinear Differential Equations, vol. 65 (Springer, Birkhäuser, 2005)
9. D. Bucur, G. Buttazzo, I. Figueiredo, On the attainable eigenvalues of the Laplace operator. SIAM J. Math. Anal. **30**, 527–536 (1999)
10. D. Bucur, G. Buttazzo, B. Velichkov, Spectral optimization problems with internal constraint. Ann. Inst. H. Poincaré Anal. Non Linéaire **30**, 477–495 (2013)
11. G. Buttazzo, Spectral optimization problems. Rev. Mat. Complut. **24**, 277–322 (2011)
12. G. Buttazzo, G. Dal Maso, Shape optimization for Dirichlet problems: relaxed formulation and optimality conditions. Appl. Math. Optim. **23**, 17–49 (1991)
13. G. Buttazzo, G. Dal Maso, An existence result for a class of shape optimization problems. Arch. Ration. Mech. Anal. **122**, 183–195 (1993)
14. G. Buttazzo, B. Velichkov, Shape optimization problems on metric measure spaces. J. Funct. Anal. **264**, 1–33 (2013)
15. G. Buttazzo, B. Velichkov, Some new problems in spectral optimization, in *Calculus of Variations and PDEs*, vol. 101 (Banach Center Publications, Polish Academy of Sciences, Warszawa, 2014), pp. 19–35
16. E.B. Davies, *Heat Kernels and Spectral Theory* (Cambridge University Press, Cambridge, 1989)

17. G. De Philippis, B. Velichkov, Existence and regularity of minimizers for some spectral optimization problems with perimeter constraint. Appl. Math. Optim. **69**, 199–231 (2014)
18. A. Henrot, Minimization problems for eigenvalues of the Laplacian. J. Evol. Equ. **3**, 443–461 (2003)
19. A. Henrot, *Extremum Problems for Eigenvalues of Elliptic Operators*. Frontiers in Mathematics (Verlag, Birkhäuser Basel, 2006)
20. A. Henrot, M. Pierre, *Variation et Optimisation de Formes. Une Analyse Géométrique*. Mathématiques & Applications (Springer, Berlin, 2005)
21. M.T. Kohler-Jobin, Une méthode de comparaison isopérimétrique de fonctionnelles de domaines de la physique mathématique. I. Une démonstration de la conjecture isopérimétrique $P\lambda^2 \geq \pi j_0^4/2$ de Pólya et Szegö Z. Angew. Math. Phys. **29**, 757–766 (1978)
22. D. Mazzoleni, A. Pratelli, Existence of minimizers for spectral problems. J. Math. Pures Appl. **100**, 433–453 (2013)
23. G. Pólya, G. Szegö, *Isoperimetric Inequalities in Mathematical Physics*. Annals of Mathematics Studies (Princeton University Press, Princeton, 1951)
24. M. van den Berg, On the spectrum of the Dirichlet Laplacian for horn-shaped regions in $\mathbb{R}^n$ with infinite volume. J. Funct. Anal. **58**, 150–156 (1984)
25. M. van den Berg, Estimates for the torsion function and Sobolev constants. Potential Anal. **36**, 607–616 (2012)
26. M. van den Berg, R. Bañuelos, Dirichlet eigenfunctions for horn-shaped regions and Laplacians on cross sections. J. Lond. Math. Soc. **53**, 503–511 (1996)
27. M. van den Berg, E. Bolthausen, Estimates for Dirichlet eigenfunctions. J. Lond. Math. Soc. **59**, 607–619 (1999)
28. M. van den Berg, D. Bucur, On the torsion function with Robin or Dirichlet boundary conditions. J. Funct. Anal. **266**, 1647–1666 (2014)
29. M. van den Berg, T. Carroll, Hardy inequality and L^p estimates for the torsion function. Bull. Lond. Math. Soc. **41**, 980–986 (2009)
30. M. van den Berg, E.B. Davies, Heat flow out of regions in $\mathbb{R}^m$. Math. Z. **202**, 463–482 (1989)
31. M. van den Berg, M. Iversen, On the minimization of Dirichlet eigenvalues of the Laplace operator. J. Geom. Anal. **23**, 660–676 (2013)
32. B. Velichkov, Existence and regularity results for some shape optimization problems. Ph.D. thesis, http://cvgmt.sns.it

On a Classical Spectral Optimization Problem in Linear Elasticity

Davide Buoso and Pier Domenico Lamberti

Abstract We consider a classical shape optimization problem for the eigenvalues of elliptic operators with homogeneous boundary conditions on domains in the N-dimensional Euclidean space. We survey recent results concerning the analytic dependence of the elementary symmetric functions of the eigenvalues upon domain perturbation and the role of balls as critical points of such functions subject to volume constraint. Our discussion concerns Dirichlet and buckling-type problems for polyharmonic operators, the Neumann and the intermediate problems for the biharmonic operator, the Lamé and the Reissner–Mindlin systems.

Keywords Polyharmonic operators · Eigenvalues · Domain perturbation

Mathematics Subject Classification (2010) Primary: 35J40 · Secondary: 35J57 · 35B20 · 74K20

1 Introduction

Let Ω be a bounded domain (i.e., a bounded connected open set) in $\mathbb{R}^N$. As is well known the problem

$$\begin{cases} -\Delta u = \gamma u, & \text{in } \Omega, \\ u = 0, & \text{on } \Omega, \end{cases}$$

admits a divergent sequence of non-negative eigenvalues

D. Buoso (✉) · P.D. Lamberti
Dipartimento di Matematica, Università degli Studi di Padova,
Via Trieste, 63, 35126 Padova, Italy
e-mail: dbuoso@math.unipd.it

P.D. Lamberti
e-mail: lamberti@math.unipd.it

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_3

$$0 < \gamma_1[\Omega] < \gamma_2[\Omega] \leq \cdots \leq \gamma_j[\Omega] \leq \ldots,$$

where each eigenvalue is repeated as many times as its multiplicity (which is finite). A classical problem in shape optimization consists in minimizing the eigenvalues $\gamma_j[\Omega]$ under the assumption that the measure of Ω is fixed. With regard to this, the most famous result is probably the Rayleigh–Faber–Krahn inequality which reads

$$\gamma_1[\Omega^*] \leq \gamma_1[\Omega], \tag{1.1}$$

where Ω^* is a ball with the same measure of Ω. In other words, the ball minimizes the first eigenvalue of the Dirichlet Laplacian among all domains with prescribed measure. Note that the first eigenvalue has multiplicity one. This inequality has been generalized in several directions aiming at minimization or maximization results in the case of other boundary conditions (for example, Neumann, Robin, Steklov boundary conditions), other operators (for example, the biharmonic operator), more general eigenvalue-type problems (for example, the buckling problem for the biharmonic operator) and other eigenvalues $\gamma_j[\Omega]$ with $j \neq 1$. It is impossible to quote here all available results in this field and we refer to the monographs by Bucur and Buttazzo [2], Henrot [18] and Kesavan [21] for extensive discussions and references.

We note that very little is known in the case of polyharmonic operators and systems. We mention that in the case of the biharmonic operator with Dirichlet boundary conditions inequality (1.1) is known as The Rayleigh Conjecture and has been proved by Nadirashvili [27] for $N = 2$ and by Ashbaugh and Benguria [1] for $N = 2, 3$. We also quote the papers by Bucur, Ferrero and Gazzola [3, 4] concerning the biharmonic operator with Steklov boundary conditions and Chasman [12] for Neumann boundary conditions. See also the extensive monograph by Gazzola, Grunau and Sweers [15] for more information on polyharmonic operators. As for systems, we quote the papers by Kawohl and Sweers [19, 20] which contain interesting lower bounds for the first eigenvalue of the Lamé system.

It should be noted that understanding the behavior of higher eigenvalues is a difficult task even in the case of the Dirichlet Laplacian. A famous result by Buttazzo and Dalmaso [11] and its recent improvement by Mazzoleni and Pratelli [26] guarantee the existence of a minimizer for $\gamma_j[\Omega]$ in the class of quasiopen sets but no information on the shape of such minimizer is given. However, it is proved in Wolf and Keller [30] that the minimizers of higher eigenvalues in general are not balls and not even unions of balls. Moreover, the numerical approach by Oudet [28] allows to get an idea of the shape of the minimizers of lower eigenvalues which confirms the negative result in [30].

One of the problems arising in the study of higher eigenvalues is related to bifurcation phenomena associated with the variation of their multiplicity which leads to complications such as, for example, lack of differentiability of the eigenvalues with respect to domain perturbation. However, as it was pointed out in [23, 25] this problem does not affect the elementary symmetric functions of the eigenvalues which depend real-analytically on the domain. This suggests that the elementary symmetric functions of the eigenvalues might be natural objects in the optimization

of multiple eigenvalues. In fact, it turns out that balls are critical points with volume constraint for the elementary symmetric functions of the eigenvalues. This property was proved for the Dirichlet and Neumann Laplacians in [24] and later was proved for polyharmonic operators in [7, 8].

In this survey paper, we adopt this point of view and show that the analysis initiated in [22–25] can be extended to a large variety of problems arising in linear elasticity including Dirichlet and buckling-type eigenvalue problems for polyharmonic operators, biharmonic operator with Neumann and intermediate boundary conditions, Lamé and Reissner–Mindlin systems. Details and proofs can be found in [5–10].

Our aim is not only to collect results spread in different papers but also to present them in a unitary way. In particular, we provide a Hadamard-type formula for the shape derivatives of the eigenvalues of the biharmonic operator which is valid not only for Dirichlet boundary conditions (as in the classical case) but also for Neumann and intermediate boundary conditions. In the case of simple eigenvalues such formula reads

$$\frac{d\gamma_n[\phi_\epsilon(\Omega)]}{d\epsilon}_{|_{\epsilon=0}} = \int_{\partial\Omega}\left(|D^2u|^2 - 2\left(\frac{\partial^2 u}{\partial\nu^2}\right)^2 + 2\frac{\partial u}{\partial\nu}\left(\mathrm{div}_{\partial\Omega}[(D^2u)\nu] + \frac{\partial\Delta u}{\partial\nu}\right) - \gamma u^2\right)\psi\cdot n d\sigma, \tag{1.2}$$

where it is assumed that Ω is sufficiently smooth, u is an eigenfunction normalized in $L^2(\Omega)$ associated with a simple eigenvalue $\gamma_n[\Omega]$, and ϕ_ϵ are perturbations of the identity I of the type $\phi_\epsilon = I + \epsilon\psi$, $\epsilon \in \mathbb{R}$. See Theorems 3.1, 3.2 and Lemma 3.3 for the precise statements and for the case of multiple eigenvalues. Note that in the case of Dirichlet boundary conditions the previous formula gives exactly the celebrated Hadamard formula

$$\frac{d\gamma_n[\phi_\epsilon(\Omega)]}{d\epsilon}_{|_{\epsilon=0}} = -\int_{\partial\Omega}\left(\frac{\partial^2 u}{\partial\nu^2}\right)^2\psi\cdot n d\sigma, \tag{1.3}$$

discussed by Hadamard [17] in the study of a clamped plate (see also Grinfeld [16]).

This paper is organized as follows: in Sect. 2 we formulate the eigenvalue problems under consideration, in Sect. 3 we state the available analyticity results for the dependence of the eigenvalues upon domain perturbation, in Sect. 4 we show that balls are critical points for the elementary symmetric functions of the eigenvalues.

2 The Eigenvalue Problems

Let Ω be an open set in $\mathbb{R}^N$. We denote by $H^m(\Omega)$ the Sobolev space of real-valued functions in $L^2(\Omega)$ with weak derivatives up to order m in $L^2(\Omega)$ endowed with its

standard norm, and by $H_0^m(\Omega)$ the closure in $H^m(\Omega)$ of $C_c^\infty(\Omega)$. We consider the following eigenvalue problems on sufficiently regular open sets Ω.

Dirichlet and buckling problems for polyharmonic operators

For $m, n \in \mathbb{N}$, $0 \le m < n$, we consider the problem

$$\mathcal{P}_{nm}: \begin{cases} (-\Delta)^n u = \gamma(-\Delta)^m u, & \text{in } \Omega, \\ u = \frac{\partial u}{\partial \nu} = \cdots = \frac{\partial^{n-1} u}{\partial \nu^{n-1}} = 0, & \text{on } \partial\Omega, \end{cases} \tag{2.1}$$

where ν denotes the unit outer normal to $\partial\Omega$. The case $m = 0$ gives the classical eigenvalue problem for the polyharmonic operator $(-\Delta)^n$ with Dirichlet boundary conditions, while the case $m > 0$ represents a buckling-type problem. For $N = 2$, $\mathcal{P}_{10}$ arises for example in the study of a vibrating membrane stretched in a fixed frame, $\mathcal{P}_{20}$ corresponds to the case of a vibrating clamped plate and $\mathcal{P}_{21}$ is related to plate buckling. If Ω is a bounded open set of class C^1 then problem (2.1) has a sequence of eigenvalues $\gamma_j^{\mathcal{P}_{nm}}$ which can be described by the Min–Max Principle. Namely,

$$\gamma_j^{\mathcal{P}_{nm}} = \min_{\substack{E \subset H_0^n(\Omega) \\ \dim E = j}} \max_{\substack{u \in E \\ u \neq 0}} R_{nm}[u], \tag{2.2}$$

for all $j \in \mathbb{N}$, where $R_{nm}[u]$ is the Rayleigh quotient defined by

$$R_{nm}[u] = \begin{cases} \frac{\int_\Omega |\Delta^r u|^2 dx}{\int_\Omega |\Delta^s u|^2 dx}, & \text{if } n = 2r, \quad m = 2s, \\ \frac{\int_\Omega |\Delta^r u|^2 dx}{\int_\Omega |\nabla \Delta^s u|^2 dx}, & \text{if } n = 2r, \quad m = 2s+1, \\ \frac{\int_\Omega |\nabla \Delta^r u|^2 dx}{\int_\Omega |\Delta^s u|^2 dx}, & \text{if } n = 2r+1, \quad m = 2s, \\ \frac{\int_\Omega |\nabla \Delta^r u|^2 dx}{\int_\Omega |\nabla \Delta^s u|^2 dx}, & \text{if } n = 2r+1, \quad m = 2s+1. \end{cases}$$

Neumann and intermediate eigenvalue problems for the biharmonic operator

By Neumann eigenvalue problem for the biharmonic operator we mean the problem

$$\mathcal{N}: \begin{cases} \Delta^2 u = \gamma u, & \text{in } \Omega, \\ \frac{\partial^2 u}{\partial^2 \nu} = 0, & \text{on } \partial\Omega, \\ \operatorname{div}_{\partial\Omega}[(D^2 u)\nu] + \frac{\partial \Delta u}{\partial \nu} = 0, & \text{on } \partial\Omega, \end{cases} \tag{2.3}$$

where $D^2 u$ denotes the Hessian matrix of u, $\operatorname{div}_{\partial\Omega}$ denotes the tangential divergence operator on $\partial\Omega$. We recall that $\operatorname{div}_{\partial\Omega} f = \operatorname{div} f - [(\nabla f)\nu] \cdot \nu$, for any vector field f smooth enough defined in a neighborhood of $\partial\Omega$. Note that we need Ω to be at least of class C^2 for the classical formulation to make sense, since we need the normal ν to be differentiable, as can easily be seen from the boundary conditions; however, we shall interpret problem (2.3) in the following weak sense

$$\int_\Omega D^2u : D^2\varphi dx = \gamma \int_\Omega u\varphi dx, \quad \forall\, \varphi \in H^2(\Omega), \tag{2.4}$$

where $D^2u : D^2\varphi = \sum_{i,j=1}^N u_{x_ix_j}\varphi_{x_ix_j}$. It is well-known that if Ω is a bounded open set of class C^1 then problem (2.3) has a sequence of eigenvalues $\gamma_j^{\mathcal{N}}$ given by

$$\gamma_j^{\mathcal{N}} = \min_{\substack{E\subset H^2(\Omega)\\ \dim E=j}} \max_{\substack{u\in E\\ u\neq 0}} \frac{\int_\Omega |D^2u|^2dx}{\int_\Omega u^2dx}, \tag{2.5}$$

for all $j \in \mathbb{N}$, where $|D^2u|^2 = \sum_{i,j=1}^N u_{x_ix_j}^2$.

If in (2.4) the space $H^2(\Omega)$ is replaced by the space $H^2(\Omega) \cap H_0^1(\Omega)$ we obtain the weak formulation of the classical eigenvalue problem

$$\mathcal{I}: \begin{cases} \Delta^2 u = \gamma u, & \text{in } \Omega,\\ u = 0, & \text{on } \partial\Omega,\\ \Delta u - K\frac{\partial u}{\partial \nu} = 0, & \text{on } \partial\Omega, \end{cases} \tag{2.6}$$

where K denotes the mean curvature of $\partial\Omega$ (the sum of the principal curvatures). Since $H_0^2(\Omega) \subset H^2(\Omega) \cap H_0^1(\Omega) \subset H^2(\Omega)$ and the spaces $H_0^2(\Omega)$, $H^2(\Omega)$ are the natural spaces associated with the Dirichlet problem $\mathcal{P}_{20}$ and the Neumann problem $\mathcal{N}$ respectively, we refer to (2.6) as the eigenvalue problem for the biharmonic operator with intermediate boundary conditions. If Ω is of class C^1 then problem (2.6) has a sequence of eigenvalues $\gamma_j^{\mathcal{I}}$ given by

$$\gamma_j^{\mathcal{I}} = \min_{\substack{E\subset H^2(\Omega)\cap H_0^1(\Omega)\\ \dim E=j}} \max_{\substack{u\in E\\ u\neq 0}} \frac{\int_\Omega |D^2u|^2dx}{\int_\Omega u^2dx}, \tag{2.7}$$

for all $j \in \mathbb{N}$.

Eigenvalue problem for the Lamé system

The eigenvalue problem for the Lamé system reads

$$\mathcal{L}: \begin{cases} -\mu\Delta u - (\lambda+\mu)\nabla \mathrm{div} u = \gamma u, & \text{in } \Omega,\\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{2.8}$$

where the unknown u is a function taking values in $\mathbb{R}^N$ and $\lambda, \mu > 0$ are (the Lamé) constants . If Ω is of class C^1 then problem (2.8) has a sequence of eigenvalues $\gamma_j^{\mathcal{L}}$ given by

$$\gamma_j^{\mathcal{L}} = \min_{\substack{E\subset (H_0^1(\Omega))^N\\ \dim E=j}} \max_{\substack{u\in E\\ u\neq 0}} \frac{\int_\Omega \mu|\nabla u|^2 + (\lambda+\mu)\mathrm{div}^2 u dx}{\int_\Omega u^2dx}, \tag{2.9}$$

for all $j \in \mathbb{N}$.

Eigenvalue problem for the Reissner–Mindlin system
Finally, we consider the eigenvalue problem for the Reissner–Mindlin system

$$\mathcal{R}:\begin{cases} -\frac{\mu}{12}\Delta\beta - \frac{\mu+\lambda}{12}\nabla\mathrm{div}\beta - \mu\frac{\kappa}{t^2}(\nabla w-\beta) = \frac{t^2\gamma}{12}\beta, & \text{in } \Omega,\\ -\mu\frac{k}{t^2}(\Delta w - \mathrm{div}\beta) = \gamma w, & \text{in } \Omega,\\ \beta = 0,\ w = 0, & \text{on } \Omega, \end{cases} \tag{2.10}$$

where the unknown $(\beta, w) = (\beta_1, \ldots, \beta_N, w)$ is a function with values in $\mathbb{R}^{N+1}$ and $\lambda, \mu, \kappa, t > 0$ are constants. According to the Reissner–Mindlin model for moderately thin plates, for $N = 2$ system (2.10) describes the free vibration modes of an elastic clamped plate $\Omega \times (-t/2, t/2)$ with midplane Ω and thickness t. In that case λ and μ are the Lamé constants, κ is the correction factor, w the transverse displacement of the midplane, $\beta = (\beta_1, \beta_2)$ the corresponding rotation and $t^2\gamma$ the vibration frequency.

If Ω is of class C^1 then problem (2.10) has a sequence of eigenvalues $\gamma_j^{\mathcal{R}}$ given by

$$\gamma_j^{\mathcal{R}} = \min_{\substack{E\subset(H_0^1(\Omega))^{N+1}\\ \dim E=j}} \max_{\substack{(\beta,w)\in E\\ u\neq 0}} \frac{\int_\Omega \frac{\mu}{12}|\nabla\beta|^2\eta + \frac{\mu+\lambda}{12}\mathrm{div}^2\beta + \mu\frac{\kappa}{t^2}|\nabla w-\beta|^2 dx}{\int_\Omega w^2 + \frac{t^2}{12}|\beta|^2 dx}, \tag{2.11}$$

for all $j \in \mathbb{N}$.

3 Analyticity Results

Let Ω be a bounded open set in $\mathbb{R}^N$ of class C^1. In the sequel, we shall consider problems (2.1), (2.3), (2.6), (2.8), (2.10) on families of open sets parametrized by suitable diffeomorphisms ϕ defined on Ω. To do so, for $k \in \mathbb{N}$ we set

$$\mathcal{A}_\Omega^k = \left\{\phi \in C_b^k(\Omega;\mathbb{R}^N) : \inf_{\substack{x_1,x_2\in\Omega\\ x_1\neq x_2}} \frac{|\phi(x_1)-\phi(x_2)|}{|x_1-x_2|} > 0\right\},$$

where $C_b^k(\Omega;\mathbb{R}^N)$ denotes the space of all functions from Ω to $\mathbb{R}^N$ of class C^k, with bounded derivatives up to order k. Note that if $\phi \in \mathcal{A}_\Omega^k$ then ϕ is injective, Lipschitz continuous and $\inf_\Omega |\det\nabla\phi| > 0$. Moreover, $\phi(\Omega)$ is a bounded open set of class C^1 and the inverse map $\phi^{(-1)}$ belongs to $\mathcal{A}_{\phi(\Omega)}^k$. Thus it is natural to consider the above mentioned eigenvalue problems on $\phi(\Omega)$ and study the dependence of the corresponding eigenvalues $\gamma_j^{\mathcal{P}_{nm}}[\phi(\Omega)], \gamma_j^{\mathcal{N}}[\phi(\Omega)], \gamma_j^{\mathcal{T}}[\phi(\Omega)], \gamma_j^{\mathcal{L}}[\phi(\Omega)], \gamma_j^{\mathcal{R}}[\phi(\Omega)]$ on $\phi \in \mathcal{A}_\Omega^k$ for suitable values of k.

The choice of k depends on the problem. In the sequel it will be always understood that k is chosen as follows:

$$\begin{array}{ll} \text{Problem } \mathcal{P}_{nm}: & k=n, \\ \text{Probelms } \mathcal{N} \text{ and } \mathcal{I}: & k=2, \\ \text{Problems } \mathcal{L} \text{ and } \mathcal{R}: & k=1. \end{array} \tag{3.1}$$

Moreover, in order to shorten our notation, we shall write $\gamma_j[\phi]$ instead of $\gamma_j^{\mathcal{P}_{nm}}[\phi(\Omega)]$, $\gamma_j^{\mathcal{N}}[\phi(\Omega)]$, $\gamma_j^{\mathcal{I}}[\phi(\Omega)]$, $\gamma_j^{\mathcal{L}}[\phi(\Omega)]$, $\gamma_j^{\mathcal{R}}[\phi(\Omega)]$, with the understanding that our statements refer to any of the problems (2.1), (2.3), (2.6), (2.8), (2.10). We endow the space $C_b^k(\Omega\,;\mathbb{R}^N)$ with its usual norm defined by $\|f\|_{C_b^k(\Omega\,;\mathbb{R}^N)} = \sup_{|\alpha|\le k,\ x\in\Omega} |D^\alpha f(x)|$. We recall that $\mathcal{A}_\Omega^k$ is an open set in $C_b^k(\Omega\,;\mathbb{R}^N)$, see [23, Lemma 3.11]. Thus, it makes sense to study differentiability and analyticity properties of the maps $\phi \mapsto \gamma_j[\phi(\Omega)]$ defined for $\phi \in \mathcal{A}_\Omega^k$.

As in [23], we fix a finite set of indexes $F \subset \mathbb{N}$ and we consider those maps $\phi \in \mathcal{A}_\Omega^k$ for which the eigenvalues with indexes in F do not coincide with eigenvalues with indexes not in F; namely we set

$$\mathcal{A}_{F,\Omega}^k = \left\{\phi \in \mathcal{A}_\Omega^k :\ \gamma_j[\phi] \neq \gamma_l[\phi],\ \forall\, j \in F,\ l \in \mathbb{N}\setminus F\right\}.$$

It is also convenient to consider those maps $\phi \in \mathcal{A}_{F,\Omega}^k$ such that all the eigenvalues with index in F coincide and set

$$\Theta_{F,\Omega}^k = \left\{\phi \in \mathcal{A}_{F,\Omega}^k :\ \gamma_{j_1}[\phi] = \gamma_{j_2}[\phi],\ \ \forall\, j_1, j_2 \in F\right\}.$$

For $\phi \in \mathcal{A}_{F,\Omega}^k$, the elementary symmetric functions of the eigenvalues with index in F are defined by

$$\Gamma_{F,h}[\phi] = \sum_{\substack{j_1,\dots,j_h\in F \\ j_1<\cdots<j_h}} \gamma_{j_1}[\phi]\cdots\gamma_{j_h}[\phi], \quad h = 1,\dots,|F|. \tag{3.2}$$

In order to state Theorems 3.1 and 3.2, we need to define a quantity $M[u, v]$ where u, v are eigenfunctions associated with an eigenvalue γ on a smooth bounded open set Ω. For each problem, $M[u, v]$ is a real valued function defined on $\partial\Omega$ as follows:

- Problem $\mathcal{P}_{nm}$:

$$M[u, v] = \frac{\partial^n u}{\partial \nu^n}\frac{\partial^n v}{\partial \nu^n}; \tag{3.3}$$

- Problem $\mathcal{N}$:

$$M[u, v] = \gamma u v - D^2 u : D^2 v; \tag{3.4}$$

- Problem $\mathcal{I}$:

$$M[u,v] = D^2u : D^2v - 2\Delta_{\partial\Omega}\left(\frac{\partial u}{\partial \nu}\frac{\partial v}{\partial \nu}\right) - \left(\frac{\partial u}{\partial \nu}\frac{\partial^3 v}{\partial \nu^3} + \frac{\partial u}{\partial \nu}\frac{\partial^3 v}{\partial \nu^3}\right); \tag{3.5}$$

- Problem $\mathcal{L}$:

$$M[u,v] = \mu\frac{\partial u}{\partial \nu}\cdot\frac{\partial v}{\partial \nu} + (\mu+\lambda)\left(\frac{\partial u}{\partial \nu}\cdot\nu\right)\left(\frac{\partial v}{\partial \nu}\cdot\nu\right); \tag{3.6}$$

- Problem $\mathcal{R}$:

$$M[u,v] = \frac{\mu}{12}\frac{\partial \beta}{\partial \nu}\cdot\frac{\partial \theta}{\partial \nu} + \frac{\mu+\lambda}{12}\left(\frac{\partial \beta}{\partial \nu}\cdot\nu\right)\left(\frac{\partial \theta}{\partial \nu}\cdot\nu\right) + \frac{\kappa\mu}{t^2}\frac{\partial w}{\partial \nu}\frac{\partial u}{\partial \nu}; \tag{3.7}$$

where $u = (\beta, w)$ and $v = (\theta, u)$.

In (3.5), $\Delta_{\partial\Omega}$ denotes the tangential Laplacian on $\partial\Omega$. Recall that $\Delta_{\partial\Omega}u = \mathrm{div}_{\partial\Omega}\nabla_{\partial\Omega}u$ where $\nabla_{\partial\Omega}u = \nabla u - \frac{\partial u}{\partial \nu}\nu$ is the tangential gradient of u.

Moreover, formula (3.8) below is expressed in terms of a basis $\{u_l\}_F$ of the eigenspace associated with an eigenvalue γ on an open set $\tilde{\phi}(\Omega)$. It will be understood that such basis is orthonormal with respect to the appropriate L^2-scalar product, which is the standard scalar product in $L^2(\tilde{\phi}(\Omega))$ for problems (2.1) with $m = 0$, (2.3), (2.6), (2.8) and the scalar product defined by $\int_{\tilde{\phi}(\Omega)}(wv + \frac{t^2}{12}\beta\cdot\eta)dy$ for problem (2.10). Note that in the case of problem (2.1) with arbitrary m, we use the natural scalar product associated with right-hand side of the equation, i.e., the scalar product defined by $\int_{\tilde{\phi}(\Omega)}\Delta^{\frac{m}{2}}u\Delta^{\frac{m}{2}}vdy$ if m is even, and $\int_{\tilde{\phi}(\Omega)}\nabla\Delta^{\frac{m-1}{2}}u\nabla\Delta^{\frac{m-1}{2}}vdy$ if m is odd.

Then we have the following

Theorem 3.1 *Let Ω be a bounded open set in $\mathbb{R}^N$ of class C^1 and F be a finite set in $\mathbb{N}$. Let $k \in \mathbb{N}$ be as in (3.1). The set $\mathcal{A}^k_{F,\Omega}$ is open in $C^k_b(\Omega;\mathbb{R}^N)$ and the real-valued maps which take $\phi \in \mathcal{A}^k_{F,\Omega}$ to $\Gamma_{F,h}[\phi]$ are real-analytic on $\mathcal{A}^k_{F,\Omega}$ for all $h = 1,\dots,|F|$. Moreover, if $\tilde{\phi} \in \Theta^k_{F,\Omega}$ is such that the eigenvalues $\gamma_j[\tilde{\phi}]$ assume the common value $\gamma_F[\tilde{\phi}]$ for all $j \in F$, and $\tilde{\phi}(\Omega)$ is of class C^{2k} then the Frechét differential of the map $\Gamma_{F,h}$ at the point $\tilde{\phi}$ is delivered by the formula*

$$d|_{\phi=\tilde{\phi}}\Gamma_{F,h}[\psi] = -\gamma_F^{h-1}[\tilde{\phi}]\binom{|F|-1}{h-1}\sum_{l=1}^{|F|}\int_{\partial\tilde{\phi}(\Omega)} M[u_l,u_l]\zeta\cdot\nu d\sigma, \tag{3.8}$$

for all $\psi \in C^k_b(\Omega;\mathbb{R}^N)$, where $\{u_l\}_{l\in F}$ is an orthonormal basis of the eigenspace associated with $\gamma_F[\tilde{\phi}]$, and $\zeta = \psi\circ\tilde{\phi}^{(-1)}$.

The proof of this theorem can be done by adapting that of [23, Theorem 3.38] (see also [25, Theorem 2.5]). Namely, by pulling-back to Ω via ϕ the operator defined on

$\phi(\Omega)$, one reduces the problem to the study of a family of operators T_ϕ defined on the fixed domain Ω. Such operators turn out to be self-adjoint with respect to a scalar product also depending on ϕ, which is obtained by pulling-back the appropriate scalar product defined of $L^2(\phi(\Omega))$. Then it is possible to apply the abstract results of [23] in order to prove the real-analyticity of the symmetric functions of the eigenvalues. Formula (3.8) is also deduced by a general formula concerning the eigenvalues of self-adjoint operators proved in [23] combined with lengthy calculations which depend on the specific case under consideration. We refer to the papers indicated in the introduction for details.

If we consider domain perturbations depending real analytically on one scalar parameter, it is possible to describe all the eigenvalues splitting from a multiple eigenvalue of multiplicity m by means of m real-analytic functions. For the sake of completeness we state the following Rellich-Nagy-type theorem which can be proved by using the abstract results [23, Theorem 2.27, Corollary 2.28] which, in turn, are proved by an argument based on reduction to finite dimension.

Theorem 3.2 *Let Ω be a bounded open set in $\mathbb{R}^N$ of class C^1. Let $k \in \mathbb{N}$ be as in (3.1), $\tilde{\phi} \in \mathcal{A}^k_\Omega$ and $\{\phi_\epsilon\}_{\epsilon\in\mathbb{R}} \subset \mathcal{A}^k_\Omega$ be a family depending real-analytically on ϵ such that $\phi_0 = \tilde{\phi}$. Let $\tilde{\gamma}$ be an eigenvalue on $\tilde{\phi}(\Omega)$ of multiplicity m, namely $\tilde{\gamma} = \gamma_{n,t}[\tilde{\phi}] = \cdots = \gamma_{n+m-1,t}[\tilde{\phi}]$ for some $n \in \mathbb{N}$. Then there exists an open interval I containing zero and m real-analytic functions $g_1, \ldots, g_m$ from I to $\mathbb{R}$ such that $\{\gamma_{n,t}[\phi_\epsilon], \ldots, \gamma_{n+m-1,t}[\phi_\epsilon]\} = \{g_1(\epsilon), \ldots, g_m(\epsilon)\}$ for all $\epsilon \in I$. Moreover, if $\tilde{\phi}(\Omega)$ is an open set of class C^{2k} then the derivatives $g_1'(0), \ldots, g_m'(0)$ at zero of the functions $g_1, \ldots, g_m$ coincide with the eigenvalues of the matrix*

$$\left(-\int_{\tilde{\phi}(\Omega)} M[u_i, u_j]\zeta \cdot \nu d\sigma\right)_{i,j\in\{1,\ldots,m\}}$$

where u_i, $i = 1, \ldots, m$, is an orthonormal basis of the eigenspace associated with $\tilde{\gamma}$.

In the case of the biharmonic operator the quantities $M[u_i, u_j]$ can be represented by one single formula which is valid for the Dirichlet problem $\mathcal{P}_{20}$, the Neumann problem $\mathcal{N}$ and the intermediate problem $\mathcal{I}$. Indeed, we can prove the following.

Lemma 3.3 *Let u, v be eigenfunctions associated with the same eigenvalue γ of one of the problems $\mathcal{P}_{20}$, $\mathcal{N}$, $\mathcal{I}$ on a bounded open set Ω of class C^4. Then*

$$M[u, v] = 2\frac{\partial^2 u}{\partial \nu^2}\frac{\partial^2 v}{\partial \nu^2} - D^2u : D^2v + \gamma uv - \frac{\partial u}{\partial \nu}\left(\mathrm{div}_{\partial\Omega}[(D^2v)\nu] + \frac{\partial \Delta v}{\partial \nu}\right) - \frac{\partial v}{\partial \nu}\left(\mathrm{div}_{\partial\Omega}[(D^2u)\nu] + \frac{\partial \Delta u}{\partial \nu}\right). \quad (3.9)$$

In particular, in these cases formula (3.8) reads

$$d|_{\phi=\tilde{\phi}}\Gamma_{F,h}[\psi] = -\gamma_F^{h-1}[\tilde{\phi}]\binom{|F|-1}{h-1}\sum_{l=1}^{|F|}\int_{\partial\tilde{\phi}(\Omega)}\left(2\left(\frac{\partial^2 u_l}{\partial\nu^2}\right)^2 - |D^2u_l|^2 + \gamma u_l^2 - 2\frac{\partial u_l}{\partial\nu}\left(\mathrm{div}_{\partial\Omega}[(D^2u_l)\nu] + \frac{\partial\Delta u_l}{\partial\nu}\right)\right)\zeta\cdot\nu d\sigma. \quad (3.10)$$

Proof In the case of problem $\mathcal{P}_{20}$, taking into account the boundary conditions $u = v = 0$ on $\partial\Omega$ and $\nabla u = \nabla v = 0$ on $\partial\Omega$, it follows that $D^2u : D^2v = \frac{\partial^2 u}{\partial\nu^2}\frac{\partial^2 v}{\partial\nu^2}$ on $\partial\Omega$ hence the right-hand side of (3.9) equals the right-hand side of (3.3) with $n = 2$.

In the case of problem $\mathcal{N}$, functions u and v satisfy the boundary conditions in (2.3) hence we immediately conclude that the right-hand side of (3.9) equals the right-hand side of (3.4).

Finally, we consider the intermediate problem $\mathcal{I}$. In this case, several calculations are required. To begin with, we note that since $u = v = 0$ on $\partial\Omega$ we have

$$\begin{aligned}\Delta_{\partial\Omega}\left(\frac{\partial u}{\partial\nu}\frac{\partial v}{\partial\nu}\right) &= \Delta_{\partial\Omega}\left(\frac{\partial u}{\partial\nu}\right)\frac{\partial v}{\partial\nu} + 2\nabla_{\partial\Omega}\frac{\partial u}{\partial\nu}\nabla_{\partial\Omega}\frac{\partial v}{\partial\nu} + \frac{\partial u}{\partial\nu}\Delta_{\partial\Omega}\left(\frac{\partial v}{\partial\nu}\right)\\ &= \mathrm{div}_{\partial\Omega}[(D^2u)\nu]\frac{\partial v}{\partial\nu} + 2\nabla_{\partial\Omega}\frac{\partial u}{\partial\nu}\nabla_{\partial\Omega}\frac{\partial v}{\partial\nu} + \frac{\partial u}{\partial\nu}\mathrm{div}_{\partial\Omega}[(D^2v)\nu].\end{aligned} \quad (3.11)$$

On the other hand, we have

$$\begin{aligned}\Delta_{\partial\Omega}\left(\frac{\partial u}{\partial\nu}\frac{\partial v}{\partial\nu}\right) &= \Delta_{\partial\Omega}(\nabla u\cdot\nabla v) = \Delta(\nabla u\cdot\nabla v) - \frac{\partial^2(\nabla u\cdot\nabla v)}{\partial\nu^2}\\ &\quad - K\frac{\partial(\nabla u\cdot\nabla v)}{\partial\nu} = \nabla\Delta u\cdot\nabla v + \nabla\Delta v\cdot\nabla u + 2D^2u : D^2v\\ &\quad - 2[(D^2u)\nu)]\cdot[(D^2v)\nu] - \nabla u\frac{\partial^2\nabla v}{\partial\nu^2} - \nabla v\frac{\partial^2\nabla u}{\partial\nu^2} - K\nabla u\cdot\frac{\partial\nabla v}{\partial\nu}\\ &\quad - K\nabla v\cdot\frac{\partial\nabla u}{\partial\nu} = \frac{\partial\Delta u}{\partial\nu}\frac{\partial v}{\partial\nu} + \frac{\partial\Delta v}{\partial\nu}\frac{\partial u}{\partial\nu} + 2D^2u : D^2v\\ &\quad - 2\nabla_{\partial\Omega}\frac{\partial u}{\partial\nu}\nabla_{\partial\Omega}\frac{\partial v}{\partial\nu} - \frac{\partial u}{\partial\nu}\frac{\partial^3 v}{\partial\nu^3} - \frac{\partial v}{\partial\nu}\frac{\partial^3 u}{\partial\nu^3} - K\frac{\partial u}{\partial\nu}\frac{\partial^2 v}{\partial\nu^2} - K\frac{\partial v}{\partial\nu}\frac{\partial^2 u}{\partial\nu^2}.\end{aligned} \quad (3.12)$$

By taking into account that functions u and v satisfy the boundary condition $\frac{\partial^2 u}{\partial\nu^2} = \frac{\partial^2 v}{\partial\nu^2} = 0$ on $\partial\Omega$, and by summing the first and last terms in the respective equalities (3.11) and (3.12) we get

$$2\Delta_{\partial\Omega}\left(\frac{\partial u}{\partial \nu}\frac{\partial v}{\partial \nu}\right) = 2D^2u : D^2v + \frac{\partial u}{\partial \nu}\left(\mathrm{div}_{\partial\Omega}[(D^2v)\nu] + \frac{\partial \Delta v}{\partial \nu}\right) \times \frac{\partial v}{\partial \nu}\left(\mathrm{div}_{\partial\Omega}[(D^2u)\nu] + \frac{\partial \Delta u}{\partial \nu}\right) - \frac{\partial u}{\partial \nu}\frac{\partial^3 v}{\partial \nu^3} - \frac{\partial v}{\partial \nu}\frac{\partial^3 u}{\partial \nu^3}. \tag{3.13}$$

The previous equality shows that in the case of problem $\mathcal{I}$ the right-hand side of (3.9) equals the right-hand side of (3.5).

4 Isovolumetric Perturbations

Consider the following extremum problems for the symmetric functions of the eigenvalues

$$\min_{V[\phi]=\text{const}} \Gamma_{F,h}[\phi] \quad \text{or} \quad \max_{V[\phi]=\text{const}} \Gamma_{F,h}[\phi], \tag{4.1}$$

where $V[\phi]$ denotes the N-dimensional Lebesgue measure of $\phi(\Omega)$.

Note that if $\tilde{\phi} \in \mathcal{A}^k_\Omega$ is a minimizer or maximizer in (4.1) then $\tilde{\phi}$ is a critical domain transformation for the map $\phi \mapsto \Gamma_{F,h}[\phi]$ subject to volume constraint, i.e.,

$$\mathrm{Ker}\, \mathrm{d}|_{\phi=\tilde{\phi}} V \subset \mathrm{Ker}\, \mathrm{d}|_{\phi=\tilde{\phi}} \Lambda_{F,h}, \tag{4.2}$$

where V is the real valued function defined on $\mathcal{A}^k_\Omega$ which takes $\phi \in \mathcal{A}^k_\Omega$ to $V[\phi]$.

The following theorem provides a characterization of all critical domain transformations ϕ. See [24] for the case of the Dirichlet and Neumann Laplacians.

Theorem 4.1 *Let Ω be a bounded open set in $\mathbb{R}^N$ of class C^1. Let $k \in \mathbb{N}$ be as in (3.1). Let F be a finite subset of $\mathbb{N}$. Assume that $\tilde{\phi} \in \Theta^k_{F,\Omega}$ is such that $\tilde{\phi}(\Omega)$ is of class C^{2k} and that the eigenvalues $\gamma_j[\tilde{\phi}]$ have the common value $\gamma_F[\tilde{\phi}]$ for all $j \in F$. Let $\{u_l\}_{l\in F}$ be an orthornormal basis of the eigenspace corresponding to $\gamma_F[\tilde{\phi}]$. Then $\tilde{\phi}$ is a critical domain transformation for any of the functions $\Gamma_{F,h}$, $h = 1, \ldots, |F|$, with volume constraint if and only if there exists $C \in \mathbb{R}$ such that*

$$\sum_{l\in F} M[u_l, u_l] = C, \quad \text{on } \partial\tilde{\phi}(\Omega). \tag{4.3}$$

Formula (4.3) follows from an application of the Lagrange Multipliers Theorem (see e.g., Deimling [14, Sect. 26] for a formulation valid in the case of infinite dimensional spaces) and formula (3.8).

Finally, thanks to the rotation invariance of the operators related to the problems we have considered, it is possible to prove the following.

Theorem 4.2 *Let the same assumptions of Theorem 4.1 hold. If $\tilde{\phi}(\Omega)$ is a ball then condition (4.3) is satisfied.*

The proof of this theorem is based on the following main idea. First, we assume that $\tilde{\phi}(\Omega)$ is a ball with radius R centered at zero. In the case of polyharmonic operators, we have that by the rotation invariance of the Laplace operator, if $\{u_l\}_{l\in F}$ is an orthonormal basis of an eigenspace, then $\{u_l \circ A\}_{l\in F}$ is also an orthonormal basis of the same eigenspace for all $A \in O_N(\mathbb{R})$, where $O_N(\mathbb{R})$ denotes the group of orthogonal linear transformations in $\mathbb{R}^N$. Since both $\{u_l\}_{l\in F}$ and $\{u_l \circ A\}_{l\in F}$ are orthonormal bases of the same space, it follows that $\sum_{l=1}^{|F|} u_l^2 \circ A = \sum_{l=1}^{|F|} u_l^2$, for all $A \in O_N(\mathbb{R})$. Thus $\sum_{l=1}^{|F|} u_l^2$ is a radial function. Then the radiality of $\sum_{l=1}^{|F|} u_l^2$ combined with appropriate calculations and similar arguments as above, allows to conclude that (4.3) is satisfied. (Note that in the case of vector-valued functions, say in the case of the Lamé system for simplicity, one has clearly to rotate the vector itself by considering $A^t \cdot (u_l \circ A)$ where we identify A with its matrix.)

It would be interesting to describe the family of open sets $\tilde{\phi}(\Omega)$ for which condition (4.3) is satisfied. In the case of problem $\mathcal{P}_{10}$ a classical result by Serrin [29] guarantees that if condition (4.3) is satisfied for the first eigenfunction then $\tilde{\phi}(\Omega)$ is a ball. The same result has been proved by Dalmasso [13] in the case of problem $\mathcal{P}_{20}$ under the assumption that the first eigenfunction does not change sign; for problem $\mathcal{P}_{21}$ a different method by Weinberger and Willms leads to the same conclusion (see e.g., [18]).

Acknowledgments The authors acknowledge financial support from the research project 'Singular perturbation problems for differential operators', Progetto di Ateneo of the University of Padova. The authors are members of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

1. M.S. Ashbaugh, R.D. Benguria, On Rayleigh's conjecture for the clamped plate and its generalization to three dimensions. Duke Math. J. **78**(1), 1–17 (1995)
2. D. Bucur, G. Buttazzo, Variational methods in some shape optimization problems. *Appunti dei Corsi Tenuti da Docenti della Scuola*, [Notes of Courses Given by Teachers at the School] (Scuola Normale Superiore, Pisa, 2002), 217 pp
3. D. Bucur, A. Ferrero, F. Gazzola, On the first eigenvalue of a fourth order Steklov problem. Calc. Var. Part. Differ. Equ. **35**(1), 103–131 (2009)
4. D. Bucur, F. Gazzola, The first biharmonic Steklov eigenvalue: positivity preserving and shape optimization. Milan J. Math. **79**(1), 247–258 (2011)
5. D. Buoso, Shape differentiability of the eigenvalues of elliptic systems, submitted
6. D. Buoso, Ph.D. thesis, University of Padova, in preparation
7. D. Buoso, P.D. Lamberti, Eigenvalues of polyharmonic operators on variable domains. ESAIM: Control. Optim. Calc. Var. **19**(4), 1225–1235 (2013)
8. D. Buoso, P.D. Lamberti, Shape deformation for vibrating hinged plates. Math. Methods Appl. Sci. **37**(2), 237–244 (2014)

9. D. Buoso, P.D. Lamberti, Shape sensitivity analysis of the eigenvalues of the Reissner-Mindlin system. To appear on SIAM J. Math. Anal
10. D. Buoso, L. Provenzano, A few shape optimization results for a biharmonic Steklov problem, preprint
11. G. Buttazzo, G. Dal, Maso, An existence result for a class of shape optimization problems. Arch. Ration. Mech. Anal. **122**(2), 183–195 (1993)
12. L.M. Chasman, An isoperimetric inequality for fundamental tones of free plates. Commun. Math. Phys. **303**(2), 421–449 (2011)
13. R. Dalmasso, Un problème de symétrie pour une équation biharmonique. (French) [A problem of symmetry for a biharmonic equation] Ann. Fac. Sci. Toulouse Math. (5) 11, no. 3, 45–53 (1990)
14. K. Deimling, *Nonlinear Functional Analysis* (Springer, Berlin, 1985)
15. F. Gazzola, H.-C. Grunau, G. Sweers, *Polyharmonic Boundary Value Problems: Positivity Preserving and Nonlinear Higher Order Elliptic Equations in Bounded Domains*, Lecture Notes in Mathematics, vol. 1991 (Springer, Berlin, 2010)
16. P. Grinfeld, Hadamard's formula inside and out. J. Optim. Theory Appl. **146**(3), 654–690 (2010)
17. J. Hadamard, Mémoire sur le problème danalyse relatif à léquilibre des plaques elastiques encastrées. Oeuvres, tome **2** (1968)
18. A, Henrot, Extremum problems for eigenvalues of elliptic operators. Frontiers in Mathematics. Birkhäuser Verlag, Basel (2006)
19. B. Kawohl, Remarks on some old and current eigenvalue problems. *Partial Differential Equations of Elliptic Type (Cortona, 1992)*, Symposium Mathematics, vol. XXXV (Cambridge University Press, Cambridge, 1994), pp. 165–183
20. B. Kawohl, G. Sweers, Remarks on eigenvalues and eigenfunctions of a special elliptic system. Z. Angew. Math. Phys. **38**(5), 730–740 (1987)
21. S. Kesavan, *Symmetrization & Applications*. Series in Analysis (World Scientific Publishing Co. Pte. Ltd, Hackensack, 2006)
22. P.D. Lamberti, M. Lanza de Cristoforis, An analyticity result for the dependence of multiple eigenvalues and eigenspaces of the Laplace operator upon perturbation of the domain. Glasg. Math. J. **44**(1), 29–43 (2002)
23. P.D. Lamberti, M. Lanza de Cristoforis, A real analyticity result for symmetric functions of the eigenvalues of a domain dependent Dirichlet problem for the Laplace operator. J. Nonlinear Convex Anal. **5**(1), 19–42 (2004)
24. P.D. Lamberti, M. Lanza de Cristoforis, Critical points of the symmetric functions of the eigenvalues of the Laplace operator and overdetermined problems. J. Math. Soc. Jpn. **58**(1), 231–245 (2006)
25. P.D. Lamberti, M. Lanza de Cristoforis, A real analyticity result for symmetric functions of the eigenvalues of a domain-dependent Neumann problem for the Laplace operator. Mediterr. J. Math. **4**(4), 435–449 (2007)
26. D. Mazzoleni, A. Pratelli, Existence of minimizers for spectral problems. J. Math. Pures Appl. **100**(3), 433–453 (2013)
27. N.S. Nadirashvili, Rayleigh's conjecture on the principal frequency of the clamped plate. Arch. Ration. Mech. Anal. **129**(1), 1–10 (1995)
28. E. Oudet, Numerical minimization of eigenmodes of a membrane with respect to the domain. ESAIM Control Optim. Calc. Var. **10**(3), 315–330 (2004)
29. J. Serrin, A symmetry problem in potential theory. Arch. Ration. Mech. Anal. **43**, 304–318 (1971)
30. S.A. Wolf, J.B. Keller, Range of the first two eigenvalues of the Laplacian. Proc. R. Soc. Lond. Ser. A **447**(1930), 397–412 (1994)

Metric Spaces of Shapes and Geometries Constructed from Set Parametrized Functions

Michel C. Delfour

Abstract In modelling, optimization, control, or identification problems with respect to a family of subsets of a fixed *hold-all* in $\mathbb{R}^N$, the nice vector space structure of the calculus of variations and control theory is no longer available. The family of all subsets of the Euclidean space is a group for the algebraic *symmetric difference*. One way to construct a family of variable domains is to consider the images of a fixed subset of $\mathbb{R}^N$ by some family of transformations of $\mathbb{R}^N$. The group structure for the *composition* of transformations induces a group structure on the space of variable sets. Each variable set is *parametrized* by its associated transformation (homeomorphism or diffeomorphism). So it is topologically identical to the initial set. To get around this limitation, metric spaces of *set-parametrized functions* (*characteristic*, *distance*, *oriented distance*, *support function*) have been introduced (*Hausdorff metric* associated with the distance function and *Caccioppoli sets* with the characteristic function) but many more complete metric spaces can be constructed such as, for instance, the *sets of positive reach* or the *sets of bounded curvatures*. The paper surveys past and current constructions while introducing new metrics for families of sets or of embedded submanifolds with a prescribed smoothness yet allowing topological changes.

Keywords Shape · Geometry · Submanifold · Caccioppoli · Positive reach · Bounded curvatures · Group · Metric space · Optimization · Design · Identification · Control

Mathematics Subject Classification (2010) Primary: 49-2 · 54-02 · Secondary: 26-02 · 53-01

This research has been supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

M.C. Delfour (✉)
Centre de recherches mathématiques,
Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, QC H3C 3J7, Canada
e-mail: delfour@crm.umontreal.ca

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_4

1 Introduction

In this paper the intuitive terminology *geometry* and *shape* is used in a very broad sense. A *geometry* is a subset of a bigger but fixed set or *hold-all D* in the Euclidean space $\mathbb{R}^N$, $N \geq 1$ an integer, and a *shape* is usually associated with the structure of the boundary of the set which is not necessarily smooth.

In general, the sets under consideration will enjoy additional properties. For instance, they can be *manifolds* which are sets that locally look like some Euclidean space $\mathbb{R}^n$, $n \geq 1$. If properties such as volume and curvatures are required, a *smooth structure* can be added to the manifold structure to make sense of integration and differential calculus on that manifold (such as smooth, differential, Riemann manifolds). It provides a classification, a detailed analysis, and a differential calculus on such sets via local bases and Christoffel symbols.

In order to properly formulate optimization, design, identification, and control problems with respect to *geometrical sets*, a convenient representation of a set is necessary to build an analytical framework in which the full power of function analytic methods can be exploited. This is very much in the spirit of Analysis that builds more and more complex structures from simple ones.

The general objectives are

- to find analytical descriptions of geometries,
- to construct spaces of geometries,
- to relax geometrical properties: volume, perimeter, curvatures, etc.,
- to introduce topologies on spaces of geometries in order to make sense of continuity of shape functions and compact families,
- to characterize the tangent space to such spaces of geometries in order to build a semidifferential calculus for geometrical objective functions,

and many other related issues. This is essential to deal with the existence, the characterization and the approximation of optimal geometries. Those objectives have a strong intersection with Differential Geometry and Shape Sensitivity Analysis but also with Set-valued Analysis occurring in Optimization and a good dose of Functional Analysis, Metric Spaces and Group Theory.

In order to structure the discussion, we begin with the fundamental issue of the choice of a representation. The following two approaches are widely present in the literature and lead to the construction of complete metric spaces (some of which are infinite dimensional manifolds):

(i) to parametrize *geometries* by *functions* and
(ii) to parametrize *functions* by *geometries*.

Roughly speaking, the first approach amounts to consider images of a reference set by a family of homeomorphisms or diffeomorphisms. The *composition* of transformations induces a *group structure* on the family of sets and a *metric structure* by specifying a metric on the family of homeomorphisms or diffeomorphisms. When

the reference set is a smooth manifold,[1] the images inherit many of its properties. But, to its disadvantage, the topology of the images is the same as the one of the reference set. As a result, this approach does not allow topological changes. This includes the construction of (complete) *Courant metrics* by A.M. Micheletti [47] in 1972 extended in Delfour and Zolésio [25, Chap. 3] and constructions based on geodesics, around 1995, by A. Trouvé [63, 65] using diffeomorphisms generated by a family of velocity fields. In such constructions the key triangle inequality for the metric is obtained by introducing some form of *geodesic property*. There are also connections with the work of D. Mumford and his collaborators.

In this paper we privilege the second approach. It is in line with the *Hausdorff metric* [38] defined in term of the *distance function* in 1914 and the *Caccioppoli sets* [12] defined in term of the *characteristic function* in 1952. Both are set-parametrized functions. This allows arbitrary topological changes and the geometrical and smoothness properties of the associated set are related to the smoothness of the set-parametrized function in a neighborhood of the boundary of the set. This point of view is also present in *set-valued analysis* (cf., for instance, Aubin and Frankowska [5]) that deals with set-valued differential equations, subdifferentials and tangent cones in control and optimization problems.

Looking at geometries as sets is conceptually different from other approaches that look at geometries as *manifolds*, *parametric surfaces* (Reifenberg [58–61] around 1960), *varifolds* (Almgrem [2–4] around 1965), or *currents* (Federer–Fleming [33] around 1960) as was illustrated in the solution of the Plateau's problem of minimal surfaces. Yet, in the end, they share a large intersection.

The paper samples material from Delfour and Zolésio [25] with complementary and new results that sharpen earlier ones.

Section 2 recalls the group structure induced by the symmetric difference and the construction of the topological metric group of characteristic functions which are *functions parametrized by sets*.

Section 3 gives the simple example of the metric group obtained by a family of transformations of an hypograph as an example of *sets parametrized by functions*.

Section 4 discusses the issue of the analytical representation of geometries as illustrated by the examples of Sects. 2 and 3: parametrize sets by function or parametrize functions by sets.

Section 5 deals with the topological metric groups of characteristic functions in L^p-spaces for Lebesgue and Hausdorff measures. A short account of some compact families is provided.

Section 6 deals with families of distance functions of non-empty sets and the construction of the metrics of uniform convergence, Lipschitz convergence, and $W^{1,p}$-convergence. Unfortunately, unlike the family of characteristic functions, the Abelian group structure is lost. In order to recover it, the neutral element $\varnothing$ that is not compatible with the above metrics is required. To get around this, we construct new metrics which are equivalent to the previous ones on the family of distance functions

[1] In fact a closed set or a crack-free (see footnote 4 for the definition.) open set is sufficient.

of non-empty sets, but whose completion contains the distance function of the empty set. This is similar to the one-point compactification of $\mathbb{R}$.

Section 7 deals with metric spaces of oriented distance functions of sets with non-empty boundary. Of special interest is the subfamily of oriented distance functions of sets whose boundary has zero Lebesgue measure for which the volume functional is continuous in the $W^{1,p}$-topology. As in Sect. 6, the Abelian group structure is lost since the oriented distance function of the empty set $\varnothing$ is $+\infty$ and the one of $\mathbb{R}^N$ is $-\infty$. Again we construct new metrics which are equivalent to the previous ones on the family of oriented distance functions of sets with non-empty boundary, but whose completion contains the oriented distance function of the empty set and $\mathbb{R}^N$. This is similar to the two-point compactification of $\mathbb{R}$.

Section 8 is devoted to the boundary properties of a set via the oriented distance function and connections with *Caccioppoli sets and sets of bounded curvature*. It introduces new complete metrics that provide control over the smoothness of the boundary by using some *truncated* oriented distance function. This effectively controls the smoothness of the oriented distance function in a fixed *narrow band* around the boundary of each set.

Section 9 discusses *sets of positive reach* and embedded submanifolds and their relation to the smoothness of the squared distance function. New (complete) metrics are constructed on families of embedded submanifolds.

Section 10 briefly recalls the metric space associated with the support function of Convex Analysis.

2 Group, Symmetric Difference, and Characteristic Functions

2.1 Group Structure Induced by the Symmetric Difference

The first observation is that a space of geometries cannot be a vector space with respect to the field $\mathbb{R}$. Yet, a group, a vector space over the field $\mathbb{Z}_2$, and a Boolean ring structure can be introduced.

Given a nonempty hold-all D, consider the *power set*[2] of D

$$\mathcal{P}(D) \stackrel{\text{def}}{=} \{A : A \subset D\}\,.$$

Several algebraic operations such as the union, the intersection, and the complement with respect to D can be defined on $\mathcal{P}(D)$, but the following one is particularly interesting since it leads to a *group structure*. Denote by $\triangle$ the *symmetric difference* of two sets A and B in $\mathcal{P}(D)$ (Fig. 1):

[2]Also denoted 2^D which emphasizes the bijection between $\mathcal{P}(D)$ and the family of functions $\chi : D \to \{0, 1\}$, that is, from D to the two-element set $\{0, 1\}$.

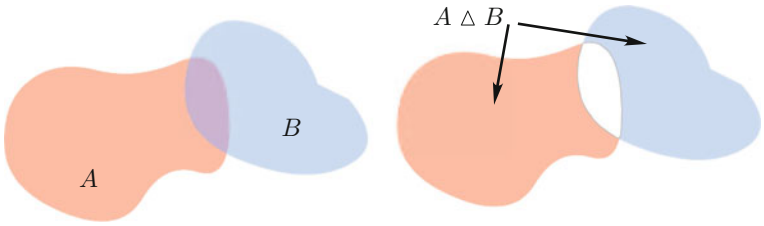


Fig. 1 Symmetric difference $A \triangle B$

$$A \triangle B \stackrel{\text{def}}{=} \left[A \cap \complement_D B\right] \cup \left[B \cap \complement_D A\right] = [A \cup B] \cap \complement_D [A \cap B],$$

where $\complement_D A = \{x \in D : x \notin A\}$. By definition, $\triangle$ is commutative and associative and $\varnothing$ is the *neutral element*

$$A \triangle B = B \triangle A, \quad (A \triangle B) \triangle C = A \triangle (B \triangle C), \quad A \triangle \varnothing = A.$$

An inverse B of A must verify

$$\begin{aligned} A \triangle B = \varnothing &\iff \left[A \cap \complement_D B\right] \cup \left[B \cap \complement_D A\right] = \varnothing \\ &\iff B \cap \complement_D A = \varnothing = A \cap \complement_D B \iff A = B. \end{aligned}$$

Therefore $A \triangle A = \varnothing$, every element A of $\mathcal{P}(D)$ is its own inverse, $-A = A$, and $(\mathcal{P}(D), \triangle)$ is an *Abelian group*. This yields a kind of *triangle inequality*:

$$(A \triangle B) \triangle (B \triangle C) = A \triangle C.$$

Because every element in this group is its own inverse, $(\mathcal{P}(D), \triangle)$ is in fact a *vector space* over the field $\mathbb{Z}_2$ with two elements. Finally, the *intersection* is distributive over the *symmetric difference*:

$$A \cap (B \triangle C) = (A \cap B) \triangle (A \cap C).$$

Hence the power set $\mathcal{P}(D)$ becomes a *ring* with *symmetric difference* as *addition* and *intersection* as *multiplication*. It is the prototypical example of a *Boolean ring*.

2.2 Metric Structures via Characteristic Functions

As a simple illustrative example, we put the elements of $\mathcal{P}(D)$ in correspondence with the *characteristic functions* which are *set-parametrized functions* and build a complete metric space structure on this set of functions. Associate with $A \in \mathcal{P}(D)$, the *characteristic function*

$$\chi_A(x) \stackrel{\text{def}}{=} \left\{ \begin{array}{ll} 1, & \text{if } x \in A \\ 0, & \text{if } x \in D\backslash A \end{array} \right\}, \qquad \boxed{\mathrm{X}(D) \stackrel{\text{def}}{=} \{\chi_A : A \subset D\}}$$
$$\mathcal{P}(D) \ni A \longleftrightarrow \chi_A \in \mathrm{X}(D).$$

The symmetric difference $\triangle$ on $\mathcal{P}(D)$ induces an Abelian group structure on $\mathrm{X}(D)$:

$$\boxed{(\chi_A \triangle \chi_B)(x) \stackrel{\text{def}}{=} |\chi_A(x) - \chi_B(x)| = \chi_{A\triangle B}(x), \quad x \in D.} \tag{2.1}$$

It is readily seen that $\mathrm{X}(D)$ is a closed subset of the Banach (vector) space of bounded functions on D endowed with the topology of uniform convergence

$$\mathcal{B}(D) \stackrel{\text{def}}{=} \{f : D \to \mathbb{R} : f \text{ bounded on } D\}, \qquad \boxed{\|f\|_{B(D)} \stackrel{\text{def}}{=} \sup_{x\in D} |f(x)|.}$$

Therefore, $\mathrm{X}(D)$ is a *complete metric space* for the metric

$$\boxed{\rho_D(A, B) \stackrel{\text{def}}{=} \|\chi_A - \chi_B\|_{\mathcal{B}(D)} = \|\chi_{A\triangle B}\|_{\mathcal{B}(D)}}$$

and $(\mathcal{P}(\mathcal{D}), \subset_{\mathcal{D}})$ is a *complete topological metric group*. This metric is both *left and right-invariant*

$$\rho_D(A \triangle C, B \triangle C) = \underset{=\rho_D(A\triangle B,\varnothing)}{\rho_D(A, B)} = \rho_D(C \triangle A, C \triangle B). \tag{2.2}$$

So, as in the case of vector spaces, it is sufficient to look at the neighbourhood of the neutral element $\varnothing$. Similarly, the tangent space in any point A will be the same as the tangent space in $\varnothing$. It turns out that this metric

$$\rho_D(A, B) \stackrel{\text{def}}{=} \|\chi_A - \chi_B\| = \|\chi_{A\triangle B}\| = \left\{ \begin{array}{ll} 0, & \text{if } A = B \\ 1, & \text{if } A \neq B \end{array} \right.$$

defines the *chaotic topology* on $\mathcal{P}(D)$.

3 Parametrize Geometries by Functions: Hypographs

The basic idea is to construct a family of variable sets from the images of a *reference set* by some group of homeomorphism or diffeomorphisms. As a simple example consider the hypograph (Fig. 2)

$$G_f \stackrel{\text{def}}{=} \left\{(x', x^N f(x')) : x' \in U,\ 0 \le x^N \le 1\right\} \tag{3.1}$$

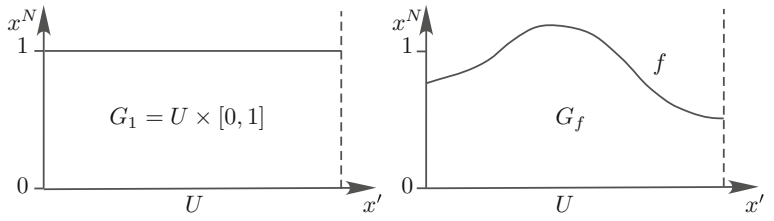


Fig. 2 Hypographs of the functions 1 and f on U

of a continuous strictly positive bounded function $f : U \to \mathbb{R}$ defined on a closed subset U of $\mathbb{R}^{N-1}$ and the following family of variable sets:

$$\mathcal{G}(U \times [0,1]) \stackrel{\text{def}}{=} \left\{ G_f : f \in C_+(U) \right\}, \tag{3.2}$$

where $C_+(U)$ is the set of continuous strictly positive bounded functions on U. Choose as *reference set* $G_1 = U \times [0,1]$. Associate with each f the transformation

$$(x', x^N) \mapsto T_f(x', x^N) \stackrel{\text{def}}{=} (x', x^N f(x')) : U \times \mathbb{R} \to U \times \mathbb{R}. \tag{3.3}$$

It is easy to verify that T_f is a bijection and that the composition

$$T_f \circ T_g = T_{fg} = T_g \circ T_f$$

is well-defined and corresponds to the pointwise product $(fg)(x') = f(x)g(x')$ of the functions f and g. The neutral element is T_1 for the function identically equal to 1, that is, $T_1 = I$. The inverse of T_f is $T_{f^{-1}}$ for $f^{-1}(x') = 1/f(x')$ the pointwise inverse of f. This defines an Abelian group which is isomorphic to the group $C_+(U)$ endowed with the pointwise product. Consider this group of homeomorphisms of the set $G_1 = U \times \mathbb{R}$

$$\mathcal{F} \stackrel{\text{def}}{=} \left\{ T_f : f \in C_+(U) \right\}. \tag{3.4}$$

With the above definitions, the set of variable sets will be the set of images of G_1 by the elements of the group $\mathcal{F}$:

$$\mathcal{G}(G_1) = \left\{ T_f(G_1) : f \in C_+(U) \right\}. \tag{3.5}$$

A natural candidate for a metric on the images $G_f = T_f(G_1)$ and $G_g = T_g(G_1)$ would be

$$\rho(T_f, T_g) \stackrel{\text{def}}{=} \sup_{x \in G_1} \|T_f(x) - T_g(x)\|_{\mathbb{R}^N} = \sup_{x' \in U} |f(x') - g(x')| \tag{3.6}$$

but the completion of $\mathcal{F}$ with respect to that metric yields degenerate T_f's. For instance, pick the sequence of constant functions $f_n(x) = 1/n$ on U. The sequence $\{f_n\} \subset C_+(U)$, converges to the function $f(x) = 0$ on U and $0 \notin C_+(U)$. Therefore, the sequence $\{T_{f_n}\}$ converges to $T_0 \notin \mathcal{F}$. To avoid this we could use the metric

$$\begin{aligned}\rho_0(T_f, T_g) &\stackrel{\text{def}}{=} \sup_{x\in G_1} \|T_f(x) - T_g(x)\|_{\mathbb{R}^N} + \sup_{x\in G_1} \|T_f^{-1}(x) - T_g^{-1}(x)\|_{\mathbb{R}^N} \\ &= \sup_{x'\in U} |f(x') - g(x')| + \sup_{x'\in U} |f^{-1}(x') - g^{-1}(x')|,\end{aligned} \tag{3.7}$$

but we loose the right-invariance. If we build in the right-invariance

$$\rho_1(T_f, T_g) \stackrel{\text{def}}{=} \rho_0(I, T_g \circ T_f^{-1}),$$

we loose the triangle inequality and only get a *semimetric*.[3]

To get around this difficulty, Micheletti [47] introduced the following general construction in 1972 that builds a *metric* from the function ρ_0 using the group structure. The first step is the construction of a distance from T_f to the neutral element $T_1 = I$ by introducing the *finite factorizations* of T_f in $\mathcal{F}$ of the form

$$T_f = T_{f_1} \circ \cdots \circ T_{f_k}, \quad f_i \in C_+(U).$$

and the function that minimizes the distance with respect to all *finite factorizations* of T_f in $\mathcal{F}$

$$\boxed{\rho(I, T_f) \stackrel{\text{def}}{=} \inf_{\substack{T_f = T_{f_1}\circ\cdots\circ T_{f_k} \\ f_i\in\mathcal{F},\, k\geq 1}} \sum_{i=1}^{k} \rho_0(I, T_{f_i}),} \tag{3.8}$$

Extend this function to all T_f and T_g in $\mathcal{F}$

$$\boxed{\rho(T_f, T_g) \stackrel{\text{def}}{=} \rho(I, T_g \circ T_f^{-1}) = \rho(I, T_{gf^{-1}}).} \tag{3.9}$$

From [25, Assumptions 2.1 and 2.2 and Theorems 2.1 and 2.2 in Chap. 3] the triangle inequality is verified and ρ is a right-invariant metric, that is, for all T_f, T_g and T_h in $\mathcal{F}$

$$\rho(T_f, T_g) = \rho(T_f \circ T_h, T_g \circ T_h).$$

---

[3] Given a space X, a function $d : X \times X \to \mathbb{R}$ is said to be a *semimetric* if

(i) $d(F, G) \geq 0$, for all F, G,
(ii) $d(F, G) = 0 \iff F = G$,
(iii) $d(F, G) = d(G, F)$, for all F, G.

This notion goes back to Fréchet and Menger.

In this definition, there is an implicit notion of *geodesic* when interpreting the factorization as a path in $\mathcal{F}$. This induces the following metric

$$d(G_f, G_g) \stackrel{\text{def}}{=} \rho(T_f, T_g), \tag{3.10}$$

on the group $\mathcal{G}(G_1)$ of images of G_1 by $\mathcal{F}$. So, in presence of a semimetric, there is an intimate relation between the existence of geodesics and the triangle inequality.

As for the notion of semidifferential, it requires the characterization of the Bouligand tangent cone in each point T_f of $\mathcal{F}$. It is obtained by taking the derivative at $t = 0$ of paths of the form $T_{f(1+tg)}$ at a given $f \in C_+(U)$ in the direction $g \in C^0(U)$, the space of bounded and uniformly continuous functions on U, and $t \neq 0$. For t sufficiently small $f(1 + tg) \in C_+(U)$ and it is natural to consider the following differential quotient

$$\frac{T_{f(1+tg)} \circ T_f^{-1} - I}{t} = \frac{T_{1+tg} - I}{t} \tag{3.11}$$

$$= \frac{T_{1+tg} - I}{t}(x', x^N) = (0, x^N\, g(x')). \tag{3.12}$$

So, the tangent space contains the Banach space $\{0\} \times C^0(U)$. It cannot be larger since for any $t \neq 0$ and any f

$$\frac{1}{t}(T_f - I)(x', x^N) = \left(0, \frac{x_N}{t}(f(x') - 1)\right) \in \{0\} \times C^0(U). \tag{3.13}$$

This metric group $\mathcal{F}$ is an example of an infinite dimensional differentiable manifold.

4 Analytical Representations of Geometries

This section discusses the issue of the analytical representation of geometries as illustrated by the examples of Sects. 2 and 4: parametrize sets by function or parametrize functions by sets.

4.1 Parametrize Geometries by Functions

The construction of Sect. 3 is generic and readily extends to more complex situations where ρ_0 is only a *semimetric*. Such metrics have been called *Courant metrics* by Micheletti [47].

In summary, the elements of the construction are (Figs. 3 and 4)

(i) $\mathcal{F}$ a group of transformations F of some subset D of $\mathbb{R}^N$

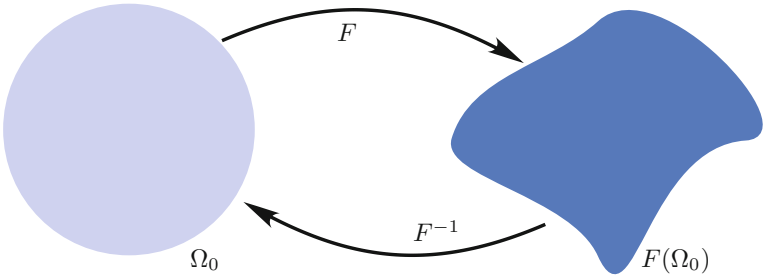


Fig. 3 Image of a fixed set Ω_0 by a bijection or transformation $F : \mathbb{R}^2 \to \mathbb{R}^2$

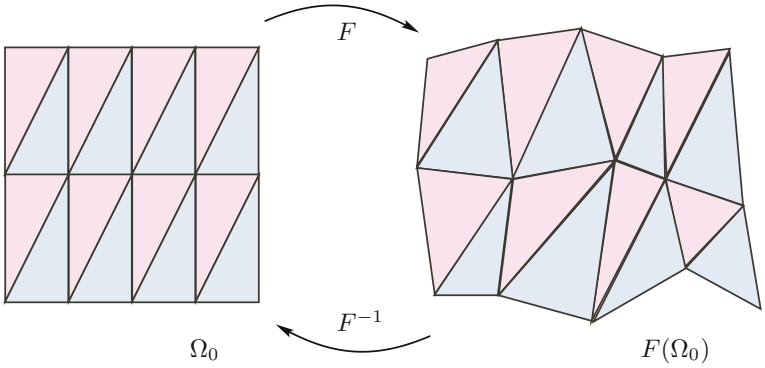


Fig. 4 Image of a fixed triangulation of a set by a bijection or transformation $F : \mathbb{R}^2 \to \mathbb{R}^2$

(ii) Ω_0, an open crack free[4] or closed subset of D
(iii) the variable domains are the images of Ω_0 by the elements $F \in \mathcal{F}$

[4] A set Ω is *crack-free* if $\overline{\complement\Omega} = \overline{\complement\overline{\Omega}}$. Clearly,

$$\overline{\complement\Omega} = \overline{\complement\overline{\Omega}} \iff \operatorname{int}\Omega = \operatorname{int}\overline{\Omega} \iff \partial\Omega = \partial\overline{\Omega}. \tag{4.1}$$

(cf. for instance, Delfour and Zolésio [25, Definition 7.1(ii), Chap. 8]). If, in addition, Ω is open, then Ω is crack-free if $\complement\Omega = \overline{\complement\overline{\Omega}}$ or, equivalently, if $\Omega = \operatorname{int}\overline{\Omega}$. In 1994 Henrot [39] introduced the terminology *Carathéodory set* for such a set that was later adopted by Tiba [62], and others such as [55], and [40]. However, this terminology does not seem to be standard. For instance, in the literature on polynomial approximations in the complex plane $\mathbb{C}$, a Carathéodory set is defined as follows.

Definition 4.1 (Dovgoshey [28] or Gaier [34]). A bounded subset Ω of $\mathbb{C}$ is said to be a *Carathéodory set* if the boundary of Ω coincides with the boundary of the unbounded component of the complement of $\overline{\Omega}$ (Figs. 5 and 6). A *Carathéodory domain* is a Carathéodory set if, in addition, Ω is simply connected. □

This definition excludes not only interior cracks but also bounded holes inside the set Ω as can be seen from the example of the annulus $\Omega = \{x \in \mathbb{R}^2 : 1 < |x| < 2\}$ in $\mathbb{R}^2$. So, it is more restrictive than $\partial\Omega = \partial\overline{\Omega}$. In order to avoid any ambiguity, we prefer the more intuitive and less ambiguous terminology *crack-free*.

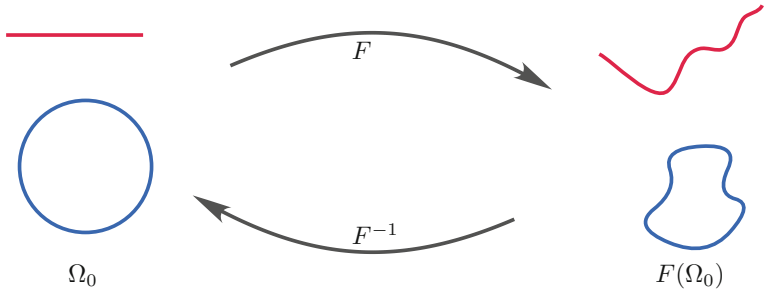


Fig. 5 Image of a *fixed line* or *circle* by a transformation $F : \mathbb{R}^2 \to \mathbb{R}^2$

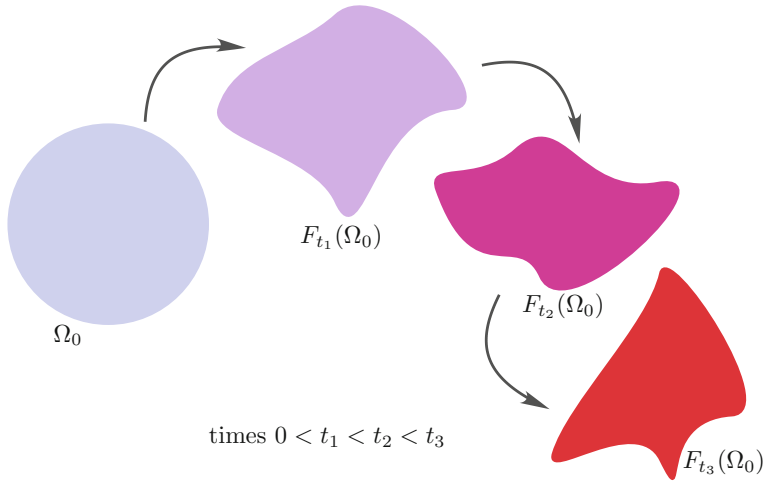


Fig. 6 Trajectory of a variable set

(iv) identify the images $F(\Omega_0)$ of Ω_0 with the quotient group $\mathcal{F}/\{F \in \mathcal{F} : F(\Omega_0) = \Omega_0\}$
(v) the *Courant metric* is a right-invariant metric on the quotient group.

The work of Micheletti [47] in 1972 was overlooked for decades. In that period, Murat-Simon [54] in 1976 constructed *pseudometrics* (that is, a semimetric in the terminology of footnote 3) and Azencott [6] and Trouvé [63–65] in 1995 constructed metrics using the idea of geodesics in the space of diffeomorphisms generated by a velocity field. A fairly complete account of Courant metrics, related constructions, and many examples can be found in Delfour and Zolesio [25, Chaps. 3 and 4]. It considerably expands the material in the 2001 edition of the book that emphasized and generalized the work of Micheletti.

For instance, in $\mathbb{R}^3$, Ω_0 can be

(a) a finite line (set of curves),
(b) a circle (set of closed curves),

(c) a sphere (set of closed surfaces with boundary), or
(d) an open ball (open sets whose boundary has no boundary).

The construction of Courant metrics on groups of transformations is to be compared with the constructions of Michor and Mumford [48–50], and Younes et al. [74].

4.2 Parametrize Functions by Geometries

In the example of Sect. 2 each set was identified with its characteristic function. This construction extends to measurable sets by introducing equivalence classes of sets and to many other set-parametrized functions such as, for instance, the distance function, the oriented distance function, and the support function.

► Identify a set with its *characteristic function*

$$\boxed{\Omega \longleftrightarrow \chi_\Omega} \quad \chi_\Omega \overset{\text{def}}{=} \left\{ \begin{array}{ll} 1, & \text{if } x \in \Omega \\ 0, & \text{if } x \notin \Omega \end{array} \right\}.$$

If we go to μ-measurable sets, then the correspondence will be with the equivalence class $[\Omega]_\mu$ of sets equal μ-almost everywhere to Ω. This is related to the measure theory of Lebesgue (1907) where the Lebesgue measure is a relaxation of the notion of volume, to the Caccioppoli sets where the norm of the distributional gradient of χ_Ω regarded as a measure yields a relaxation of the perimeter, and to geometric measure theory of Federer [32] (1969).

► Identify a set with its *distance function*

$$\boxed{[\Omega] \longleftrightarrow d_\Omega} \quad d_\Omega(x) \overset{\text{def}}{=} \left\{ \begin{array}{ll} \inf\limits_{y\in\Omega} \|y - x\|, & \text{if } \Omega \neq \varnothing \\ +\infty, & \text{if } \Omega = \varnothing \end{array} \right\}, \quad [\Omega] \overset{\text{def}}{=} \left\{ \Omega' : \overline{\Omega'} = \overline{\Omega} \right\}.$$

This is related to the Hausdorff (1914) metric that formalizes the early notion of *écart mutuel* of Pompéju (1905), to the *sets of positive reach* and *curvature measures* of Federer [31] (1959), and to set-valued analysis.

► Identify a set with its *oriented distance function*

$$\boxed{[\Omega] \longleftrightarrow b_\Omega} \quad b_\Omega(x) \overset{\text{def}}{=} d_\Omega(x) - d_{\complement\Omega}(x), \; [\Omega] \overset{\text{def}}{=} \left\{ \Omega' : \overline{\Omega'} = \overline{\Omega} \text{ and } \partial\Omega' = \partial\Omega \right\}.$$

This is a finer partition than the one with the distance function where both the closure of the set and its boundary are invariants. It is related to the geometry of and the calculus on hypersurfaces.

► Identify a set with its *support function*

$$\boxed{[\Omega] \longleftrightarrow \sigma_\Omega} \quad \sigma_\Omega(x) \overset{\text{def}}{=} \sup_{y\in\Omega} x \cdot y, \qquad [\Omega] \overset{\text{def}}{=} \left\{ \Omega' : \overline{\text{co}}\, \Omega' = \overline{\text{co}}\, \Omega \right\},$$

where $\overline{\text{co}}\, \Omega$ denotes the closed convex hull of Ω. This is related to convex analysis.

5 Metric Structures via Characteristic Functions of Measurable Sets

We have already seen in Sect. 2.2 that a (complete) metric structure can be associated with the Abelian group of characteristic functions of subsets of a set. This construction extends to measurable subsets of a set.

5.1 L^∞-Convergence of Measurable Sets

Given a *measure*[5] μ and a μ-measurable subset $D \subset \mathbb{R}^N$, let $L^\infty(D, \mu)$ denote the Banach space of equivalence classes of *μ-measurable functions bounded almost everywhere on D*. Denote by $[A]_\mu$ the equivalence class of μ-measurable subsets of D that are equal almost everywhere. Identify equivalence classes and characteristic functions

$$\boxed{\begin{array}{c}\left\{[A]_\mu : A \subset D\ \mu\text{-easurable}\right\} \supset [A]_\mu \\ \longleftrightarrow \\ \chi_A \in \mathrm{X}_\mu(D) \stackrel{\text{def}}{=} \{\chi_A : A \subset D\ \mu\text{-measurable}\}.\end{array}}$$

Since $A \triangle B$ is μ-measurable for A and B μ-measurable, we obtain a group in $L^\infty(D, \mu)$ which is closed for the metric

$$\boxed{\rho_D(A, B) \stackrel{\text{def}}{=} \|\chi_B - \chi_A\|_{L^\infty(D,\mu)}.}$$

Again we have an Abelian group, a left and right invariant metric, and completness of $(\mathrm{X}_\mu(D), \rho_D)$. This includes Lebesque and Hausdorff measures of all dimensions.

5.2 L^p-Convergence of Measurable Sets, $1 \leq p < \infty$

Given a μ-measurable subset $D \subset \mathbb{R}^N$, consider the Banach space of equivalence classes of μ-measurable p-integrable functions

$$L^p(D, \mu) \stackrel{\text{def}}{=} \left\{ f : D \to \mathbb{R} : \int_D |f|^p \, d\mu < \infty \right\}, \quad 1 \leq p < \infty.$$

[5] In the sense of Evans and Gariepy [30, Definition and Warning, p. 1].

The set

$$\mathrm{X}_\mu(D) \cap L^1(D,\mu) = \{\chi_A : A \subset D \text{ is } \mu\text{-measurable and } \mu(A) < \infty\} \tag{5.1}$$

is a subgroup of

$$\mathrm{X}_\mu(D) \overset{\text{def}}{=} \{\chi_A : A \subset D \text{ is } \mu\text{-measurable}\}.$$

in $L^p(D,\mu)$ with respect to $\triangle$.

Theorem 5.1 *Let $1 \le p < \infty$, let μ be a measure on $\mathbb{R}^N$, and let $\varnothing \neq D \subset \mathbb{R}^N$ be μ-measurable.*

(i) *$\mathrm{X}_\mu(D) \cap L^1(D,\mu)$ is closed in $L^p(D,\mu)$. The function*

$$\rho_{D,p}([A_2]_\mu, [A_1]_\mu) \overset{def}{=} \|\chi_{A_2} - \chi_{A_1}\|_{L^p(D,\mu)}$$

defines a complete metric structure *on the Abelian group $\mathrm{X}_\mu(D) \cap L^1(D,\mu)$ that makes it a topological group. If $\mu(D) < \infty$, then $\mathrm{X}_\mu(D) \cap L^1(D,\mu) = \mathrm{X}_\mu(D)$.*

(ii) *If, in addition, D is σ-finite with respect to μ for a family $\{D_k\}$ of μ-measurable subsets of D such that $\mu(D_k) < \infty$, for all $k \ge 1$, then $\mathrm{X}_\mu(D)$ is closed in $L^p_{loc}(D,\mu)$ and*

$$\rho_{D,p}([A_2]_\mu, [A_1]_\mu) \overset{def}{=} \sum_{k=1}^{\infty} \frac{1}{2^n} \frac{\|\chi_{A_2} - \chi_{A_1}\|_{L^p(D_k,\mu)}}{1 + \|\chi_{A_2} - \chi_{A_1}\|_{L^p(D_k,\mu)}}$$

defines a complete metric structure on the Abelian group $\mathrm{X}_\mu(D)$ that makes it a topological group. When μ is a Radon measure on $\mathbb{R}^N$, the assumption that D is σ-finite with respect to μ can be dropped.

An important property is that all the L^p-metrics, $1 \le p < \infty$, are equivalent. Part (i) covers sets of finite measure. Part (ii) handles sets with *infinite Lebesgue measure*, but not sets with *infinite Hausdorff measure*.

We can introduce families of sets in $\mathcal{P}(D)$ with more than one property and combine the associated metrics. For instance, the sum or the maximum of a finite number of metrics is a metric and other constructions that we shall see later.

5.3 Compact Families

Among the examples of compact families in L^1 are the celebrated *finite perimeter* or *Caccioppoli sets* (cf. [12] in 1952) as we shall see later in Sect. 8.1. They have been

used by E. De Giorgi [26][6] in 1953 and [27][7] in 1954 to solve *Plateau's problem of minimal surface* in the fifties.

Other more recent compact families have been constructed by adding the *uniform cone property* (D. Chenais [13] in 1975), the *uniform fat segment property* (a version on epigraphs was announced in a note by W. Liu et al. [44] in 2000 without proof followed by a paper by Tiba [62] in 2003; a geometrical (equivalent) version with proof was introduced in Delfour–Zolésio [24] in 2001 and expanded in [23] in 2007 and in [25] in 2011), and a bound on the *density perimeter* (Bucur–Zolésio [10] in 1996), or a bound on the *distribution curvature tensor* ([24] in 2001 and [25, Sect. 5, Chap. 7, pp. 354–358] in 2011). General *capacity conditions* have also been introduced by Bucur–Zolésio [9] in 1995.

It turns out that, for those families, we get much more than the L^p-convergence of characteristic functions. We get the stronger and richer $W^{1,p}$-convergence of the oriented the distance function of Sect. 7.2.3 (cf. see also Delfour–Zolésio [25, Sects. 13 and14, pp. 387–394]).

Other types of convergence can be found in the literature. In 1959, H. Federer [31, Sect. 4] introduced *sets of positive reach* associated with the square of distance function as we shall see in Sect. 9. He gives a compactness theorem (Theorem 9.1 in Sect. 9) with C^1-convergence of the square of the distance function for the family of subsets A of a bounded open hold all D that have a unique projection onto $\overline{A}$ from points in their tubular neighbourhood (cf. [31, Theorem 4.13], [25, Theorem 4.6, Chap. 6], and Theorem 9.2 in Sect. 9).

6 Metric Structures via Distance Functions

6.1 Pompéiu's Ecart Mutuel and Hausdorff Metric on Compact Sets

The *distance function* from a point x to a set A

$$\boxed{d_A(x) \overset{\text{def}}{=} \begin{cases} \inf\limits_{y\in A} |y-x|, & A \neq \varnothing, \\ +\infty, & A = \varnothing, \end{cases}} \tag{6.1}$$

is another example of a family of functions that is parametrized by sets. If $A \neq \varnothing$, this function is *Lipschitz continuous* with constant equal to 1

$$|d_A(y) - d_A(x)| \leq |y - x|.$$

[6] The first note published by De Giorgi describing his approach to Caccioppoli sets

[7] The first complete exposition by De Giorgi of the theory of Caccioppoli sets.

So, it is *uniformly continuous on* $\mathbb{R}^N$ and belongs to $W^{1,p}_{loc}(\mathbb{R}^N)$, $1 \le p \le \infty$. The *écart mutuel*

$$\boxed{\rho_H(A, B) \stackrel{\text{def}}{=} \max\left\{\sup_{x\in B} d_A(x), \sup_{y\in A} d_B(y)\right\},} \tag{6.2}$$

was introduced in the thesis of D. Pompéiu [57, Chap. III, pp. 281–282] in 1905. This notion was further studied as a metric on the family of *compact* subsets of $\mathbb{R}^N$ by F. Hausdorff in 1914. The compactness of the sets A and B is necessary to make this expression finite and to single out the unique closed representative in the associated equivalent classes of sets with the same distance function. Yet, a slightly different point of view makes it possible to relax the compactness assumption and consider stronger metrics.

6.2 Metric of Uniform Convergence

Given a hold-all $D \subset \mathbb{R}^N$ and a subset A of D, there are bijections between the distance function d_A, the equivalence class $[A]_d$, and the closure $\overline{A}$

$$d_A \longleftrightarrow [A]_d \stackrel{\text{def}}{=} \left\{B \subset \overline{D} : \overline{B} = \overline{A}\right\} \longleftrightarrow \overline{A}. \tag{6.3}$$

To define a metric on non-empty closed subsets that are not necessarily bounded is equivalent to define a metric on the family of distance functions

$$\boxed{C_d(D) \stackrel{\text{def}}{=} \left\{d_A : \forall A,\ \varnothing \neq A \subset \overline{D}\right\},} \tag{6.4}$$

where D is an arbitrary non-empty hold-all in $\mathbb{R}^N$. To work with open subsets Ω of an open hold-all D it is customary to use the distance to the complement $\complement\Omega$ and the family

$$\boxed{C^c_d(D) \stackrel{\text{def}}{=} \{d_{\complement\Omega} : \Omega \text{ open and } \Omega \subsetneq D\}} \tag{6.5}$$

of distance functions to the complement.

First consider the case of a bounded hold-all D. Denote by $C(\overline{D})$ the Banach space of uniformly continuous functions endowed with the norm

$$\|f\|_{C(D)} = \sup_{x\in D} |f(x)|. \tag{6.6}$$

Since the distance functions are uniformly continuous on $\mathbb{R}^N$, this induces a complete metric

$$\boxed{\rho([A]_d, [B]_d) \stackrel{\text{def}}{=} \|d_A - d_B\|_{C(D)} = \sup_{x \in D} |d_A(x) - d_B(x)|} \tag{6.7}$$

which coincides with the Hausdorff metric. Moreover, $C_d(D)$ is compact for that metric. So, it is a very weak topology. In particular, it does not preserve the volume.

When D is not bounded, use the Fréchet topology associated with compact subsets of $\overline{D}$ as we did for the characteristic function. Denoting this space by $C_{loc}(D)$, $C_d(D)$ is a complete metric space as a closed subspace of $C_{loc}(D)$. Therefore, the compactness assumption of the classical definition of the Hausdorff metric has been removed. Hence, $C_d(\mathbb{R}^N)$, the set of all non-empty closed subsets of $\mathbb{R}^N$, has a complete metric space structure.

6.3 Metric of Lipschitz Convergence

Adopting this point of view opens the door to other metrics by replacing the space of continuous functions $C_{loc}(D)$ by a function space compatible with the properties of the distance function. Indeed, since d_A is Lipschitz continuous on $\mathbb{R}^N$, it belongs to the Banach space $C^{0,1}(\mathbb{R}^N)$ of Lipschitz continuous functions on $\mathbb{R}^N$ with norm

$$\|f\| \stackrel{\text{def}}{=} |f(x_0)| + \sup_{\substack{x,y \in D \\ x \neq y}} \frac{|f(y) - f(x)|}{|y - x|} \tag{6.8}$$

for some fixed but arbitrary point $x_0 \in D$. This induces the following complete metric on $C_d(D)$

$$\boxed{\begin{aligned}\rho([A]_d, [B]_d) \stackrel{\text{def}}{=} & |d_B(x_0) - d_A(x_0)| \\ & + \sup_{\substack{x,y \in D \\ x \neq y}} \frac{|d_B(y) - d_A(y) - (d_B(x) - d_A(x))|}{|y - x|}.\end{aligned}} \tag{6.9}$$

With this metric $C_d(D)$ is no longer compact when D is bounded.

6.4 Metric of $W^{1,p}$-Convergence, $1 \leq p < \infty$

For D bounded open the family of distance functions $C_d(D)$ is a subset of $W^{1,p}(D)$, $1 \leq p < \infty$. This induces a new complete metric on $C_d(D)$

$$\boxed{\rho([A]_d, [B]_d) \stackrel{\text{def}}{=} \|d_A - d_B\|_{W^{1,p}(D)}.} \tag{6.10}$$

As for the previous Lipschitz convergence, $C_d(D)$ is no longer compact when D is bounded, but the volume of the closure of the subsets is a continuous function. Some additional conditions such as the *uniform cone* or *uniform fat segment* properties, the *density perimeter*, etc…are required to get compact families (cf. Sects. 5.3 and 7 on the oriented distance function).

When D is not bounded, use the Fréchet topology associated with compact subsets of $\overline{D}$ as we did for the characteristic function. The above $W^{1,p}(D)$-metrics are equivalent for all p, $1 \leq p < \infty$.

6.5 Metrics Compatible with the Abelian Group Structure

To our best knowledge, the following metric spaces are new. In contrast with the family $\mathrm{X}(D)$ of characteristic functions, the family $C_d(D)$ of distance functions excludes the empty set for which $d_\varnothing = +\infty$. As a result the Abelian group structure is lost. In order to recover the Abelian group structure $(\mathcal{P}(D), \triangle)$ on $C_d(D)$ we need to add the *neutral element* $d_\varnothing = +\infty$,

$$\overline{C}_d(D) \stackrel{\text{def}}{=} \left\{ d_A : A \subset \overline{D} \right\},$$

where the group operation on $\overline{C}_d(D)$ is defined as follows:

$$d_A \triangle d_B \stackrel{\text{def}}{=} d_{A \triangle B}$$
$$d_A \triangle d_\varnothing = d_A, \quad d_A \triangle d_B = d_B \triangle d_A, \quad d_A \triangle d_A = d_\varnothing.$$

But this is incompatible with the definition of the Hausdorff metric. To get around the presence of $+\infty$, introduce the following *equivalent metric* on $C_d(D)$

$$\rho(A, B) \stackrel{\text{def}}{=} \sup_{x \in D} \left| \frac{d_A(x)}{1 + d_A(x)} - \frac{d_B(x)}{1 + d_B(x)} \right| = \sup_{x \in D} \frac{|d_A(x) - d_B(x)|}{(1 + d_A(x))(1 + d_B(x))}$$
$$\rho(A, \varnothing) = \sup_{x \in D} \left\{ 1 - \frac{d_A(x)}{1 + d_A(x)} \right\} = \sup_{x \in D} \frac{1}{1 + d_A(x)}, \quad \rho(\varnothing, \varnothing) = 0.$$

The *completion of* $C_d(D)$ with respect to that metric is precisely $\overline{C}_d(D)$. The underlying construction is a classical compactification of $\mathbb{R}$ by adding points at infinity $+\infty$ or $-\infty$ (here we only add $+\infty$ since d_A is always positive). Other variations of that construction can be obtained by using a $W^{1,p}$-norm rather than the sup norm over D.

7 Metric Structures via Oriented Distance Functions

7.1 Oriented Distance Function: Definition and Some Properties

Given $A \subset \mathbb{R}^N$, the *oriented distance function*

$$\boxed{b_A(x) \stackrel{\text{def}}{=} d_A(x) - d_{\mathbb{R}^N \setminus A}} \tag{7.1}$$

is another example of a family of function that is parametrized by sets. It is finite if and only if the boundary $\partial A = \overline{A} \cap \overline{\mathbb{R}^N \setminus A}$ of A is not empty since

$$b_A(x) = \begin{cases} +\infty, & \text{if } A = \varnothing \\ \in \mathbb{R}, & \text{if } \partial A \neq \varnothing \\ -\infty, & \text{if } A = \mathbb{R}^N. \end{cases} \tag{7.2}$$

If $\partial A \neq \varnothing$, b_A is Lipschitz continuous with constant equal to 1

$$|b_A(y) - b_A(x)| \leq |y - x| .$$

So, it is uniformly continuous on $\mathbb{R}^N$ and a $W^{1.p}_{loc}(\mathbb{R}^N)$ function, $1 \leq p \leq \infty$.

As in the case of the distance function, we can define a metric on the family of oriented distance functions of subsets (not necessarily bounded) of $\overline{D}$, $\varnothing \neq D \subset \mathbb{R}^N$ open,

$$\boxed{C_b(D) \stackrel{\text{def}}{=} \left\{ b_A : \forall A,\ A \subset \overline{D},\ \varnothing \neq \partial A \right\}} \tag{7.3}$$

by using the bijection between b_A and the equivalence class

$$[A]_b \stackrel{\text{def}}{=} \left\{ B \subset \overline{D} : \overline{B} = \overline{A} \text{ and } \partial B = \partial A \right\} \longleftrightarrow b_A. \tag{7.4}$$

It is a finer partition than the one given by the distance function d_A. The closures $\overline{A}$ and $\overline{\complement A}$, the boundary ∂A, and the interiors

$$\text{int } A \stackrel{\text{def}}{=} \complement \left(\overline{\complement A} \right) \text{ and int } \complement A \stackrel{\text{def}}{=} \complement \overline{A} \tag{7.5}$$

are invariants in the equivalence class $[A]_b$, but, in general,

$$b_{\overline{A}} \lneqq b_A \lneqq b_{\text{int } A}. \tag{7.6}$$

For instance, consider the closed disk in $\mathbb{R}^2$ with an interior crack and hairs outside. The following conditions are necessary and sufficient

$$b_{\text{int } A} = b_A \iff \partial \,\text{int}\, A = \partial A \quad \text{and} \quad b_A = b_{\overline{A}} \iff \partial A = \partial \overline{A}. \tag{7.7}$$

Hence we get the following characterization: a set A, $\partial A \neq \varnothing$, is *crack free* (in the sense of the definition of the footnote 4) if and only if

$$b_A = b_{\overline{A}}.$$

At that level of generality, the boundary ∂A may have a non-zero "volume", that is, a non-zero N-dimensional Lebesgue measure $\mathrm{m}_N(\partial A) > 0$. Having a boundary with zero volume is closer to the intuitive notion of a geometric set. So, it will be natural to introduce the following terminology and notation.

Definition 7.1 (i) A set $A \subset \mathbb{R}^N$, $\partial A \neq \varnothing$, is said to have a *thin*[8] boundary if $\mathrm{m}_N(\partial A) = 0$.

(ii) Denote by

$$\boxed{C_b^0(D) \overset{\text{def}}{=} \left\{b_A : \forall A,\ A \subset \overline{D},\ \varnothing \neq \partial A \text{ and } \mathrm{m}_N(\partial A) = 0\right\}} \tag{7.8}$$

the subset of oriented distance functions with thin boundaries. (Cf. [25, Chap. 7, p. 349].) □

7.2 Some General Metrics on $C_b(D)$

In view of the fact that the function b_A has the same properties as the function d_A, we can define the same metrics on the family $C_b(D)$ as we did on $C_d(D)$.

7.2.1 Metric of Uniform Convergence

When D is bounded, we get a first complete metric

$$\boxed{\rho([A]_b, [B]_b) \overset{\text{def}}{=} \|b_A - b_B\|_{C(D)} = \sup_{x \in D} |b_A(x) - b_B(x)|} \tag{7.9}$$

similar to the one of Pompéiu–Hausdorff. This result is much more difficult to prove than its counterpart for the set of distance functions.

[8] This terminology should not be confused with the notion of *thin set* in the capacity sense.

When D is bounded $C_b(D)$ is also compact and when D is not bounded, use the Fréchet topology associated with compact subsets of $\overline{D}$ to get a complete metric.

7.2.2 Metric of Lipschitz Convergence

We now introduce a new metric. Since b_A is Lipschitz continuous on $\mathbb{R}^N$, it belongs to the Banach space $C^{0,1}(\mathbb{R}^N)$ of Lipschitz continuous functions on $\mathbb{R}^N$ with norm

$$\|f\|_{C^{0,1}} \stackrel{\text{def}}{=} |f(x_0)| + \sup_{\substack{x,y\in D\\ x\neq y}} \frac{|f(y)-f(x)|}{|y-x|} \tag{7.10}$$

for some fixed but arbitrary point $x_0 \in D$. This induces the following complete metric on $C_b(D)$

$$\boxed{\rho([A]_b,[B]_b) \stackrel{\text{def}}{=} |b_B(x_0)-b_A(x_0)| + \sup_{\substack{x,y\in D\\ x\neq y}} \frac{|b_B(y)-b_A(y)-(b_B(x)-b_A(x))|}{|y-x|}.}$$

With this metric $C_b(D)$ is no longer compact when D is bounded, but the volume is preserved. Some additional conditions such as the uniform cone or uniform fat segment properties are required to get compact families.

7.2.3 Metric of $W^{1,p}$-convergence, $1 \leq p < \infty$

For D bounded open the distance functions of $C_b(D)$ belong to $W^{1,p}(D)$, $1 \leq p < \infty$. This induces a new complete metric on $C_b(D)$ (cf. [25, Chap. 7, p. 352])

$$\boxed{\rho([A]_b,[B]_b) \stackrel{\text{def}}{=} \|b_A - b_B\|_{W^{1,p}(D)}.} \tag{7.11}$$

When D is not bounded, use the Fréchet topology associated with compact subsets of $\overline{D}$ as we did for the characteristic and distance functions. The above metrics are all equivalent for $1 \leq p < \infty$.

By definition, we can recover distance functions from the oriented distance function

$$d_A(x) = b_A^+(x) \stackrel{\text{def}}{=} \max\{b_A(x),0\}, \quad d_{\complement A}(x) = b_A^-(x) \stackrel{\text{def}}{=} \max\{-b_A(x),0\},$$
$$d_{\partial A}(x) = |b_A(x)| = d_A(x) + d_{\complement A}(x).$$

For simplicity, let D be bounded open. Then the corresponding mappings

$$b_A \mapsto (d_A, d_{\complement A}, d_{\partial A}) : W^{1,p}(D) \to W^{1,p}(D) \times W^{1,p}(D) \times W^{1,p}(D)$$

are Lipschitz continuous of constant one. This means the $W^{1,p}$-convergence for a sequence $\{b_{A_n}\}$ implies the $W^{1,p}$-convergence for the sequences $\{d_{A_n}\}$, $\{d_{\complement A_n}\}$, and $\{d_{\partial A_n}\}$. Moreover, since b_A is differentiable and $|\nabla b_A| \leq 1$ almost everywhere, we can also recover characteristic functions of the closure, interior and boundary of the set A: for almost all x

$$\begin{aligned}
\chi_{\overline{A}}(x) &= 1 - |\nabla d_A(x)| = 1 - \left|\nabla b_A^+(x)\right|, \quad \chi_{\text{int}\,\complement A} = \left|\nabla b_A^+(x)\right| \\
\chi_{\overline{\complement A}}(x) &= 1 - |\nabla d_{\complement A}(x)| = 1 - \left|\nabla b_A^-(x)\right|, \quad \chi_{\text{int}\,A} = \left|\nabla b_A^-(x)\right| \\
\chi_{\partial A}(x) &= 1 - |\nabla d_{\partial A}(x)| = 1 - |\nabla\,(d_A(x) + d_{\complement A}(x))| = 1 - |\nabla b_A(x)|\,.
\end{aligned} \tag{7.12}$$

In general A need not be Lebesgue measurable, but its closure, interior, and boundary are and we get the Lipschitz continuity of the mapping

$$b_A \mapsto (\chi_{\overline{A}}, \chi_{\text{int}\,A}, \chi_{\partial A}) : W^{1,p}(D) \to L^p(D) \times L^p(D) \times L^p(D). \tag{7.13}$$

7.2.4 $W^{1,p}$-metric on $C_b^0(D)$ and Volume Functional

In view of the continuity of the mapping (7.13), the subfamily $C_b^0(D)$ of subsets of D whose boundary has zero Lebesgue measure is closed in $W^{1,p}(D)$ and $C_b^0(D)$ is a complete metric space. The direct consequence of this is that any set A such that $\varnothing \neq \partial A$ and $\mathrm{m}_N(\partial A) = 0$ is Lebesgue measurable since the *outer measure* of $A \backslash \text{int}\, A \subset \partial A$ and $\overline{A} \backslash A \subset \partial A$ are zero. Hence

$$\mathrm{m}_N(A) = \mathrm{m}_N(\overline{A}) = \mathrm{m}_N(\text{int}\, A) \tag{7.14}$$

and the notion of volume is locally well-defined and is an invariant in the equivalence class $[A]_b$. Both int A and $\overline{A}$ belong to the equivalence class

$$[A]_\chi \overset{\text{def}}{=} \left\{B \subset \overline{D} : \chi_B = \chi_A \ \mathrm{m}_N\text{-a.e.}\right\}. \tag{7.15}$$

As a consequence, the volume functional is well-defined and continuous

$$b_A \mapsto \mathrm{m}_N(A) = \int_{\overline{D}} \left|\nabla b_A^-(x)\right| \, d\mathrm{m}_N : W^{1,p}(D) \to \mathbb{R} \tag{7.16}$$

is continuous. Observe that it would not have been possible to specify that $\mathrm{m}_N(\partial A) = 0$ at the level of the family $\mathrm{X}_{\mathrm{m}_N}(D)$ of m_N-measurable characteristic functions as defined by (5.1) in Sect. 5.2.

In view of the above properties, the $W^{1,p}$-topology on the family $C_b^0(D)$ of oriented distance functions plays a central role and it will be advantageous to prove compactness of subfamilies at the level.

7.3 Metrics Compatible with the Abelian Group Structure

To our best knowledge, the following metric spaces are new. Before leaving this section it is interesting to ask if there is some group structure on the families $C_b(D)$ and $C_b^0(D)$ as in the case of $\mathrm{X}_\mu(D)$. In order to induce the Abelian group structure $(\mathcal{P}(D), \triangle)$ on $C_b(D)$ we need to include the sets $\varnothing$ and possibly $\mathbb{R}^N$ when $D = \mathbb{R}^N$

$$\begin{gathered} b_A(x) \stackrel{\text{def}}{=} d_A(x) - d_{\mathbb{R}^N \setminus A}(x), \\ \Rightarrow b_\varnothing = d_\varnothing - d_{\mathbb{R}^N} = +\infty, \quad b_{\mathbb{R}^N} = d_{\mathbb{R}^N} - d_{\mathbb{R}^N \setminus \mathbb{R}^N} = -\infty \\ \overline{C}_b(D) \stackrel{\text{def}}{=} \left\{ b_A : A \subset \overline{D} \right\}. \end{gathered}$$

The group operation on $\overline{C}_b(D)$ can now be defined as follows:

$$\begin{gathered} b_A \triangle b_B \stackrel{\text{def}}{=} b_{A \triangle B} \\ b_A \triangle b_\varnothing = b_A, \quad b_A \triangle b_B = b_B \triangle b_A, \quad b_A \triangle b_A = b_\varnothing. \end{gathered}$$

To get around the $\pm\infty$, introduce the following *equivalent metric* on $C_b(D)$

$$\begin{gathered} \rho(A, B) \stackrel{\text{def}}{=} \sup_{x \in D} \left| \frac{b_A(x)}{1 + |b_A(x)|} - \frac{b_B(x)}{1 + |b_B(x)|} \right| \\ \rho(A, \varnothing) = \sup_{x \in D} \left| \frac{b_A(x)}{1 + |b_A(x)|} - 1 \right|, \quad \rho(A, \mathbb{R}^N) = \sup_{x \in \mathbb{R}^N} \left| \frac{b_A(x)}{1 + |b_A(x)|} + 1 \right|. \end{gathered}$$

The completion of $C_b(D)$ with respect to that metric is precisely $\overline{C}_b(D)$. Similar constructions can be done with the $W^{1,p}$-metric in place of the sup metric.

8 Boundary Properties of Sets in $C_b^0(D)$

8.1 Finite Perimeter or Caccioppoli Sets

Henri Lebesgue [41] was a pioneer in relaxing the notion of volume to the one of measure in 1904 and in showing that not all sets are measurable. In view of the correspondence between subsets of $\mathbb{R}^N$ and characteristic functions, a subset of $\mathbb{R}^N$ is measurable if and only if its characteristic function is a measurable function.

In the same spirit Renato Caccioppoli and Ennio de Giorgi used characteristic functions to extend the notion of *surface measure* of the boundary of a Lebesgue measurable set with smooth boundary to the case of a nonsmooth boundary. This measure was defined as a set function for the first time in 1927 by Caccioppoli [11]. In 1952 de Giorgi presented his first results developing the ideas of Caccioppoli on

the definition of the measure of boundaries of sets at the Salzburg Congress of the Austrian Mathematical Society independently proving some results of Caccioppoli. He published his first papers on the topic ([26] in 1953, [27] in 1954) and, after the death of Caccioppoli, started to call sets of finite perimeter *Caccioppoli sets*. In 1960 Herbert Federer and Wendell Fleming [33] and in 1969 Federer [32] introduced the notions of *normal and integral currents* and showed that Caccioppoli sets can be studied within the framework of the *theory of currents*. Yet, it is now customary to use the framework of functions of bounded variation.

We quote a few definitions and theorems from Giusti [35].

Definition 8.1 Given an open set D in $\mathbb{R}^N$ and a function $f \in L^1(D)$.

(i) The *total variation* of f in D

$$V(f, D) \stackrel{\text{def}}{=} \sup_{\substack{g=(g_1,\ldots,g_N)\in C_c^1(D;\mathbb{R}^N) \\ |g(x)|\le 1 \text{ in } D}} \int_D f \operatorname{div} g \, dx, \tag{8.1}$$

where $C_c^1(D; \mathbb{R}^N)$ is the space of continuously differentiable vector functions on D with compact support in D.[9]

(ii) The space of *functions of bounded variation*

$$\mathrm{BV}(D) \stackrel{\text{def}}{=} \left\{ f \in L^1(D) : V(f, D) < +\infty \right\}. \tag{8.3}$$

The space of *functions of locally bounded variation*

$$\mathrm{BV}_{loc}(D) \stackrel{\text{def}}{=} \{ f : D \to \mathbb{R} : \forall \Omega \subset D \text{ bounded open}, \ f \in \mathrm{BV}(\Omega) \}. \tag{8.4}$$

(iii) A Lebesgue measurable subset A of $\mathbb{R}^N$ is said to have *finite perimeter* with respect to an open set D if

$$\chi_A|_D \in \mathrm{BV}(D) \tag{8.5}$$

and we shall use the following notation for the *perimeter relative to* D

$$P_D(A) \stackrel{\text{def}}{=} V(\chi_A, D); \tag{8.6}$$

[9] Equivalently, given a function $f \in L^1(D)$, the linear function $\vec{\varphi} \mapsto -\int_D f \operatorname{div} \vec{\varphi} \, dx$ is the *distributional gradient* ∇f of f. The condition (8.1) means that this vector function is continuous on the space $\mathcal{B}^0(D)^N$ of bounded continuous vector functions on D. So it is an element of the dual of $\mathcal{B}^0(D)^N$. The space of continuous linear functions $L : \mathcal{B}(D) \to \mathbb{R}$ is denoted $M^1(D)$ with the usual norm

$$\|L\|_{M^1(D)} = \sup_{\substack{\varphi\in\mathcal{B}^0(D) \\ \|\varphi\|_{\mathcal{B}}(D)\le 1}} |L(\varphi)|, \quad \|\varphi\|_{\mathcal{B}}(D) \stackrel{\text{def}}{=} \sup_{x\in D} |\varphi(x)|. \tag{8.2}$$

So, the dual of $\mathcal{B}^0(D)^N$ is $M^1(D)^N$.

a Lebesgue measurable subset A of $\mathbb{R}^N$ is said to have *locally finite perimeter* with respect to D if for all bounded open subset Ω of D

$$\chi_A|_\Omega \in \mathrm{BV}(\Omega). \tag{8.7}$$

Note that the above definitions slightly differ from the standard ones where A is assumed to be a Borel subset of D. □

By definition, sets A with zero Lebesgue measure have zero perimeter even if their boundary has a strictly positive Lebesgue measure.

Example 8.1 Consider the set A of points with rational coordinates in the open unit disk $B_1(0)$ centered at the origin in $\mathbb{R}^2$. Then $\partial A = \overline{B_1(0)}$, $\mathrm{m}_2(A) = 0$, $\mathrm{m}_2(\partial A) = \pi > 0$, and ∂A has an interior $B_1(0)$ and a boundary $\partial(\partial A)$ equal to the unit circle. □

Submanifolds M of $\mathbb{R}^N$ such that $\overline{M} = \partial M$ and $\mathrm{m}_N(\partial M) = 0$ all have zero perimeters. This means that the notion of perimeter as an abstract boundary measure is only pertinent for sets A of non-zero Lebesgue measure. It is not a substitute for Hausdorff measures that work in all subdimensions. We shall come back to submanifolds later on.

Example 8.2 Similarly to Example 8.1, consider the set E of points with irrational coordinates in the open unit disk $B_1(0)$ centered at the origin in $\mathbb{R}^2$. Then $\partial E = \overline{B_1(0)}$, $\mathrm{m}_2(E) = \pi$, $\mathrm{m}_2(\partial E) = \pi > 0$, and ∂E has an interior $B_1(0)$ and a boundary $\partial(\partial E)$ equal to the unit circle. The sets E, $B_1(0)$, and $\overline{B_1(0)}$ belong to the same equivalence class in $L^1(D)$ but their boundaries differ: $\partial E = \overline{B_1(0)} \neq \partial B_1(0) = \partial\overline{B_1(0)}$. Yet, the boundary measure of E, $B_1(0)$, and $\overline{B_1(0)}$ relative to $\mathbb{R}^N$ are the same and their perimeters are all equal to 2π. Thus within the class it is more natural to work with the open set $B_1(0)$ rather than the set E since the boundary of $B_1(0)$ is smooth and the perimeter of $B_1(0)$ coincides withe its classical surface area. This motivates the introduction of a *nice representative* in the equivalence class of a set $\chi_E \in \mathrm{BV}(D)$

$$[E]_{\mathrm{BV}(D)} \stackrel{\mathrm{def}}{=} \{\chi_A : \chi_A = \chi_E \text{ in } \mathrm{BV}(D)\}$$

for which the relation between its boundary and the perimeter would be more intuitive. It is similar to the nice representative in the equivalence class of a function of $L^1(D)$ (functions equal almost everywhere) compared to the nice representative in the equivalence class of a function in the Sobolev space $W^{1,1}(D)$ (functions equal quasi-everywhere). □

We have the following characterization of a set with locally finite perimeter that emphasizes the fact that all the action takes place in arbitrary small neighbourhoods of boundary points.

Theorem 8.1 *A Lebesgue measurable subset A of* $\mathbb{R}^N$ *is* locally of finite perimeter *if and only if*

$$\forall x \in \partial A,\ \exists r > 0 \ \ such \ \ that \chi_A \in \mathrm{BV}(B_r(x)), \tag{8.8}$$

where $B_r(x)$ is the open ball in x of radius r.

The important compactness theorem associated with finite perimeter sets is the following (see, for instance, Giusti [35, Theorem 1.19]).

Theorem 8.2 *Let D be a bounded open subset which is sufficiently regular for the Rellich Theorem to hold. The sets of functions uniformly bounded in the* $\mathrm{BV}(D)$*-norm are relatively compact in $L^1(D)$-strong.*

Given a bounded open set D in $\mathbb{R}$, the family of Caccioppoli sets

$$\{\chi_A : A \subset D,\ \chi_A \in \mathrm{BV}(D)\}$$

endowed with the metric

$$\rho([A],[B]) \stackrel{\text{def}}{=} \|\chi_A - \chi_B\|_{\mathrm{BV}(D)}$$

is closed in $\mathrm{BV}(D)$. The family

$$\left\{\chi_A \in L^1(D, \mathrm{m}_N) : \chi_A \in \mathrm{BV}(D) \leq c\right\}$$

for some constant c and the Lebesgue measure m_N, is *sequentially compact* (and, hence, *compact*) in $L^1(D)$.

Unfortunately, this theorem does not preserve the zero m_N-measure of the boundary as can be seen from the following example which is an adaptation of Example 1.10 in [35, Example 1.10, p. 7].

Example 8.3 ([25, Example 6.2, p. 249]) Let $D = B(0,1)$ in $\mathbb{R}^2$ be the open ball in 0 of radius 1. For $i \geq 1$, let $\{x_i\}$ be an ordered sequence of all points in D with rational coordinates. Associate with each i the open ball

$$B_i = \{x \in D : |x - x_i| < \rho_i\}, \quad 0 < \rho_i \leq \min\{2^{-i}, 1 - |x_i|\}, \quad i \geq 1.$$

Define the new sequence of open subsets of D,

$$\Omega_n = \bigcup_{i=1}^{n} B_i, \quad n \geq 1,$$

and notice that

$$\forall n \geq 1, \quad \mathrm{m}_2(\partial\Omega_n) = 0, \quad P_D(\Omega_n) \leq 2\pi,$$

where $\partial\Omega_n$ is the boundary of Ω_n. Moreover, since the sequence of sets $\{\Omega_n\}$ is increasing,

$$\chi_{\Omega_n} \to \chi_\Omega \text{ in } L^1(D), \quad \Omega = \bigcup_{i=1}^{\infty} B_i, \quad P_D(\Omega) \leq \liminf_{n\to\infty} P_D(\Omega_n) \leq 2\pi.$$

Observe that $\overline{\Omega} = \overline{D}$ and that $\partial\Omega = \overline{\Omega} \cap \overline{\complement\Omega} \supset \overline{D} \cap \complement\Omega$ and

$$\mathrm{m}_2(\Omega_n) \leq \sum_{i=1}^{n} \mathrm{m}_2(B_i) \leq \sum_{i=1}^{n} \pi 2^{-2i} \leq \sum_{i=1}^{\infty} \pi 2^{-2i} = \frac{\pi}{3} \quad \Rightarrow \mathrm{m}_2(\Omega) \leq \frac{\pi}{3}.$$

Thus

$$\mathrm{m}_2(\partial\Omega) = \mathrm{m}_2(\overline{D} \cap \complement\Omega) \geq \mathrm{m}_2(\overline{D}) - \mathrm{m}_2(\Omega) \geq \pi - \frac{\pi}{3} = \frac{2\pi}{3},$$

since $\mathrm{m}_2(\overline{D}) = \mathrm{m}_2(D) = \pi$. For p, $1 \leq p < \infty$, the sequence of characteristic functions $\{\chi_{\Omega_n}\}$ converges to χ_Ω in $L^p(D)$-strong. However, for all n, $\mathrm{m}_2(\partial\Omega_n) = 0$, but $\mathrm{m}_2(\partial\Omega) > 0$. □

8.2 Preliminary Considerations

Having established that the Lebesgue measure of A is an invariant in the equivalence class $[A]_b$ of elements of $C_b^0(D)$, it is natural to turn to the properties of its boundary ∂A. Under which conditions can we say that ∂A is locally of finite Hausdorff measure in some dimension as, for instance, for *Caccioppoli sets*. Under which conditions can we make sense of its curvatures as, for instance, the *curvature measures* of H. Federer [31] for *sets of positive reach*.

For simplicity, let D be a bounded open Lipschitzian domain. Since for $b_A \in C_b^0(D)$, A is Lebesgue measurable, we can consider the space

$$C_b^0(\mathrm{BV}(D)) \stackrel{\text{def}}{=} \left\{ b_A : A \subset \overline{D},\ \partial A \neq \varnothing,\ \mathrm{m}_N(\partial A) = 0,\ \chi_A \in \mathrm{BV}(D) \right\} \tag{8.9}$$

which nicely combines the constraints on b_A and χ_A. In so doing we eliminate Caccioppoli sets for which $\mathrm{m}_N(\partial A) > 0$ (cf. [35, Example 1.10] and [25, Example 6.2, p. 249]). In view of the continuities (7.12), it is a complete metric space for the metric

$$\rho(A, B) \stackrel{\text{def}}{=} \|b_A - b_B\|_{W^{1,p}(D)} + \|\chi_A - \chi_B\|_{\mathrm{BV}(D)}, \quad 1 \leq p < +\infty. \tag{8.10}$$

In the above mentioned example the Caccioppoli set Ω with $\mathrm{m}_N(\partial\Omega) > 0$ is constructed from a sequence of sets Ω_n that belong to the space $C_b^0(\mathrm{BV}(D))$ and whose perimeter is uniformly bounded. So, the characteristic functions χ_{Ω_n} converge strongly to χ_Ω. However, the corresponding oriented distance functions b_{Ω_n} only weakly converge in $W^{1,p}(D)$ to some b_A which means that $\mathrm{m}_N(\partial A)$ may not be zero and that A and Ω may not be in the same equivalence class. Otherwise, we

would have had $m_N(\partial\Omega) = 0$. Therefore, the compactness theorem of De Giorgi does not extend to the simultaneous strong convergence of the oriented distance functions in $W^{1,p}(D)$. Somehow, we need an extra condition on a sequence $\{b_{A_n}\}$ in $C_b^0(\mathrm{BV}(D))$ (for instance, $\|\nabla\chi_{A_n}\|_{\mathrm{BV}(D)} \leq c$) to get the sequential compactness. To our best knowledge, this is an open question.

8.3 Sets of Bounded Curvature

We recall the definition of the family of sets with bounded or locally bounded curvature (cf. [25, Chap. 7, p. 381]). They include $C^{1,1}$-domains, convex sets, and the sets of positive reach of H. Federer [31]. They yield compactness theorems for subfamilies of $C_b(D)$ and $C_b^0(D)$ in the $W^{1,p}(D)$-metric.

Definition 8.2 (i) Given a bounded open nonempty holdall D in $\mathbb{R}^N$ and a subset A of $\overline{D}$, $\partial A \neq \varnothing$, its *boundary* ∂A is said to be of *bounded curvature* with respect to D if

$$\nabla b_A \in \mathrm{BV}(D)^N. \tag{8.11}$$

This family of sets will be denoted as follows

$$\mathrm{BC}_b(D) \overset{\text{def}}{=} \left\{b_A \in C_b(D) \,:\, \nabla b_A \in \mathrm{BV}(D)^N\right\}$$

and the subfamily of subsets whose boundary has zero Lebesgue measure

$$\mathrm{BC}_b^0(D) \overset{\text{def}}{=} \left\{b_A \in C_b^0(D) \,:\, \nabla b_A \in \mathrm{BV}(D)^N\right\}.$$

(ii) Given a subset A of $\mathbb{R}^N$, $\partial A \neq \varnothing$, its *boundary* ∂A is said to be of *locally bounded curvature* if ∇b_A belongs to $\mathrm{BV}_{\mathrm{loc}}(\mathbb{R}^N)^N$. The family of sets whose boundary is of locally bounded curvature will be denoted by

$$\mathrm{BC}_b \overset{\text{def}}{=} \left\{b_A \in C_b(\mathbb{R}^N) \,:\, \nabla b_A \in \mathrm{BV}_{\mathrm{loc}}(\mathbb{R}^N)^N\right\}$$

and the subfamily of subsets whose boundary has zero Lebesgue measure

$$\mathrm{BC}_b^0 \overset{\text{def}}{=} \left\{b_A \in C_b^0(\mathbb{R}^N) \,:\, \nabla b_A \in \mathrm{BV}_{\mathrm{loc}}(\mathbb{R}^N)^N\right\}.$$

□

A first observation is that at any point $x \notin \partial A$, there exists a ball $B_r(x) \subset \mathbb{R}^N \backslash \partial A$ where the components of ∇b_A are always of locally bounded variation. This means

that it is sufficient to impose that condition locally at each point of the boundary $\partial(\partial A)$ (recall that $\partial A = b_A^{-1}\{0\}$ may have a non-empty interior).

Theorem 8.3 ([25, Theorem 5.1, p. 354]) *Let A, $\partial A \neq \varnothing$, be a subset of $\mathbb{R}^N$. Then A is of locally bounded curvature if and only*

$$\forall x \in \partial(\partial A),\ \exists \rho > 0 \text{ such that } \nabla b_A \in \mathrm{BV}(B(x,\rho))^N, \tag{8.12}$$

where $B(x,\rho)$ is the open ball of radius $\rho > 0$ in x.

It turns out that the sets A such that $b_A \in \mathrm{BC}_b^0(D)$ are sets of finite perimeters and that those such that $b_A \in \mathrm{BC}_b^0$ are sets of locally finite perimeter. The next theorem nicely completes [25, Theorem 5.2, p. 354 and Theorems 11.1 and 11.2, p. 382] that only proved the result for $\chi_{\partial A}$ and $\mathrm{m}_N(\partial A)$ not necessarily equal to zero.

Theorem 8.4 *(i) Given an open set D and $A \subset D$ such that $\partial A \neq \varnothing$ and $b_A \in \mathrm{BC}_b^0(D)$, then $\nabla d_A \in \mathrm{BV}(D)^N$ and $\chi_A \in \mathrm{BV}(D)$.*
(ii) Given $b_A \in \mathrm{BC}_b^0$, then $\nabla b_A \in \mathrm{BV}_{loc}(\mathbb{R}^N)^N$ and $\chi_A \in \mathrm{BV}_{loc}(\mathbb{R}^N)$.

Proof (i) For an arbitrary $\varphi \in \mathcal{D}(D)$,

$$\int_D \partial_i d_A(x)\, \partial_j \varphi(x)\, dx = \int_{D\backslash\overline{A}} \partial_i d_A(x)\, \partial_j \varphi(x)\, dx = \int_{D\backslash\overline{A}} \partial_i b_A(x)\, \partial_j \varphi(x)\, dx,$$

since $\mathrm{m}_N(\partial A) = 0$, $d_A(x) = \max\{0, b_A(x)\}$, and

$$\nabla d_A = \nabla b_A \quad \text{a.e. on } D\backslash\overline{A}.$$

Hence,

$$V(\partial_i d_A, D) = V(\partial_i b_A, D\backslash\overline{A}) \leq V(\partial_i b_A, D) < \infty$$

since $\partial_i b_A \in \mathrm{BV}(D)$. This implies that $\nabla d_A \in \mathrm{BV}(D)^N$.

For ∇d_A in $\mathrm{BV}(D)^N$, there exists a sequence $\{u_k\}$ in $C^\infty(D)^N$ such that[10]

$$u_k \to \nabla d_A \text{ in } L^1(D)^N \quad \text{and} \quad \|Du_k\|_{M^1(D)} \to \|D^2 d_A\|_{M^1(D)}$$

as k goes to infinity, and since $|\nabla d_A(x)| \leq 1$, this sequence can be chosen in such a way that

$$\forall k \geq 1, \quad |u_k(x)| \leq 1.$$

This follows from the use of mollifiers (cf. [35, Theorem 1.17, p. 15]). Since $\mathrm{m}_N(\partial A) = 0$,

$$\chi_A = \chi_{\overline{A}} = 1 - |\nabla d_A| = 1 - |\nabla d_A|^2.$$

[10] See footnote 9 for the definition of $M^1(D)$ in the scalar case. Here we use the same notation for $N \times N$ matrices of such functions.

For all V in $\mathcal{D}(D)^N$

$$-\int_D \chi_A \operatorname{div} V\, dx = \int_D (|\nabla d_A|^2 - 1) \operatorname{div} V\, dx = \int_D |\nabla d_A|^2 \operatorname{div} V\, dx.$$

For each u_k

$$\int_D |u_k|^2 \operatorname{div} V\, dx = -2\int_D [{}^*Du_k]u_k \cdot V\, dx = -2\int_D u_k \cdot [Du_k]V\, dx,$$

where ${}^*Du_k$ is the transpose of the Jacobian matrix Du_k and

$$\begin{aligned}\left|\int_D |u_k|^2 \operatorname{div} V\, dx\right| &\le 2\int_D |u_k|\,|Du_k|\,|V|\, dx \\ &\le 2\,\|Du_k\|_{L^1}\|V\|_{C(D)} \le 2\,\|Du_k\|_{M^1}\|V\|_{C(D)}\end{aligned}$$

since for $W^{1,1}(D)$-functions $\|\nabla f\|_{L^1(D)^N} = \|\nabla f\|_{M^1(D)^N}$. Therefore, as k goes to infinity,

$$\left|\int_D \chi_A \operatorname{div} V\, dx\right| = \left|\int_D |\nabla d_A|^2 \operatorname{div} V\, dx\right| \le 2\|D^2 d_A\|_{M^1}\|V\|_{C(D)}.$$

Therefore, $\nabla\chi_A \in M^1(D)^N$.

(ii) It is sufficient to check the property at every point $x \in \partial A$ using $B_r(x), r > 0$, in place of D in part (i). □

Now, for sets with bounded curvature, we have both global and local compactness theorems. We add a part (ii) below that gives the convergence of the characteristic functions for sets whose boundary has zero measure. The condition is slightly stronger than the one in the compactness Theorem 8.2 of de Giorgi. Could it be used to prove existence of minimal surfaces?

Theorem 8.5 ([25, Theorem 11.1, p. 382]) *Let D be a nonempty bounded open Lipschitzian subset of $\mathbb{R}^N$.*

(i) Then, for any sequence $\{A_n\}$, $\partial A_n \neq \varnothing$, of subsets of $\overline{D}$ such that

$$\exists c > 0, \forall n \ge 1, \quad \|D^2 b_{A_n}\|_{M^1(D)} \le c, \tag{8.13}$$

there exist a subsequence $\{A_{n_k}\}$ and a set A, $\partial A \neq \varnothing$, such that $\nabla b_A \in \mathrm{BV}(D)^N$ and

$$b_{A_{n_k}} \to b_A \text{ in } W^{1,p}(D)\text{-strong and } \chi_{\operatorname{int} A_{n_k}} \to \chi_{\operatorname{int} A} \text{ in } L^p(D)\text{-strong}$$

for all p, $1 \le p < \infty$. Moreover, for all $\varphi \in \mathcal{D}^0(D)$,

$$\lim_{n\to\infty} \langle \partial_{ij} b_{A_{n_k}}, \varphi\rangle = \langle \partial_{ij} b_A, \varphi\rangle, \quad 1 \le i, j \le N, \quad \|D^2 b_A\|_{M^1(D)} \le c. \quad (8.14)$$

(ii) If, in addition, $\mathrm{m}_N(\partial A_n) = 0$ *for all n, then* χ_A *belongs to* $\mathrm{BV}(D)$ *and* $\chi_{A_{n_k}} \to \chi_A$ *in* $L^p(D)$*-strong.*

A more interesting version of the above theorem is obtained by localizing the condition (8.13) around the boundary of each set A.

Given $h > 0$ and a subset $A \subset \mathbb{R}^N$, introduce the *tubular neighbourhoods* of thickness $h > 0$ around A and around ∂A:

$$U_h(A) \stackrel{\text{def}}{=} \left\{x \in \mathbb{R}^N : d_A(x) < h\right\} \quad \Rightarrow U_h(\partial A) = \left\{x \in \mathbb{R}^N : |b_A(x)| < h\right\},$$

since $d_{\partial A}(x) = |b_A(x)|$.

Theorem 8.6 ([25, Theorem 11.2, p. 382]) *Let* $\varnothing \neq D \subset \mathbb{R}^N$ *be bounded open and* $\{A_n\}$, $\partial A_n \neq \varnothing$, *be a sequence of subsets of* $\overline{D}$. *Assume that there exist* $h > 0$ *and* $c > 0$ *such that*

$$\forall n, \quad \|D^2 b_{A_n}\|_{M^1(U_h(\partial A_n))} \le c. \quad (8.15)$$

(i) Then, there exist a subsequence $\{A_{n_k}\}$ *and* $A \subset \overline{D}$, $\varnothing \neq \partial A$, *such that* $\nabla b_A \in \mathrm{BV}_{loc}(\mathbb{R}^N)^N$ *and for all p,* $1 \le p < \infty$,

$$b_{A_{n_k}} \to b_A \text{ in } W^{1,p}(U_h(D))\text{-strong}. \quad (8.16)$$

Moreover, for all $\varphi \in \mathcal{D}^0(U_h(\partial A))$,

$$\lim_{k\to\infty} \langle \partial_{ij} b_{A_{n_k}}, \varphi\rangle = \langle \partial_{ij} b_A, \varphi\rangle, \quad 1 \le i, j \le N, \quad \|D^2 b_A\|_{M^1(U_h(\partial A))} \le c.$$

(ii) If, in addition, $\mathrm{m}_N(\partial A_n) = 0$ *for all n, then* χ_A *belongs to* $\mathrm{BV}_{loc}(\mathbb{R}^N)$ *and* $\chi_{A_{n_k}} \to \chi_A$ *in* $L^p_{loc}(\mathbb{R}^N)$*-strong.*

8.4 Metrics for Families of Sets with Smooth Boundaries

8.4.1 Oriented Distance Function and Differential Geometry

The connection between the smoothness of the boundary of a set and the smoothness of the oriented distance function in a neighbourhood of its boundary has a long history. However, the statements and the proofs have been and are not always accurate or complete (cf. [25, Theorem 8.1 and footnote 4, Chap. 7, p. 365]). It is true for sets of class C^k, $k \ge 2$, and the equivalence remains true down to $C^{1,1}$.

Theorem 8.7 (Local version) *Let $A \subset \mathbb{R}^N$, $\partial A \neq \varnothing$ and let $x \in \partial A$. A is locally a set of class $C^{1,1}$ at x if and only if*

$$\exists \rho > 0 \text{ such that } b_A \in C^{1,1}(\overline{B_\rho(x)}) \quad \text{and} \quad \mathrm{m}_N(\partial A \cap B_\rho(x)) = 0. \tag{8.17}$$

Moreover, on ∂A, the gradient ∇b_A coincides with the exterior unit normal n to ∂A, the projection *$p_{\partial A}$ onto ∂A is given by*

$$p_{\partial A}(x) \stackrel{def}{=} x - b_A(x)\, \nabla b_A(x),$$

$\nabla b_A = n \circ p_{\partial A}$ in $\overline{B_\rho(x)}$, and ∂A is locally a $C^{1,1}$-submanifold of dimension $N-1$ at x.

The *orthogonal projection* P onto the *tangent plane* $T_x \partial A$ is given by

$$P(x) \stackrel{\text{def}}{=} I - \nabla b_A(x)\, {}^*\nabla b_A(x),$$

where ${}^*V$ denotes the *transpose of a column vector* V in $\mathbb{R}^N$. Thus, $V {}^*V$ is a square matrix and ${}^*V\, V$ is the inner product of V with itself. In summary, at each point $x \in \partial A$

(1) $\nabla b_A(x)$ is the unit exterior normal to the set A,
(2) $P(x) \stackrel{\text{def}}{=} I - \nabla b_A(x)\, {}^*\nabla b_A(x)$ is the *first fundamental form* of ∂A,
(3) $D^2 b_A(x)$ is the *second fundamental form* of ∂A,
(4) $(D^2 b_A(x))^2$ is the *third fundamental form* of ∂A,
(...) …
(N + 1) $(D^2 b_A(x))^{N-1}$ is the *Nth fundamental form* of ∂A,

where $D^2 b_A(x)$ is the Hessian matrix of $b_A(x)$.

The *eigenvalues* of the matrix $D^2 b_A(x)$ are 0 and the $N-1$ principal curvatures of ∂A at the point x. Therefore, $\Delta b_A(x)$ is equal to the sum of the eigenvalues which is called the *mean curvature* by mathematicians while in other areas "mean" means the sum of the eigenvalues divided by $N-1$. With this definition of b_A, the curvatures of the boundary of the unit ball in $\mathbb{R}^N$ all are positive. Hence, there is an *implicit orientation* of the boundary ∂A which is an embedded submanifold of dimension $N-1$.

From this a *complete intrinsic tangential differential calculus* is available without *local covariant and contravariant bases and Christoffel symbols*. For instance, given a function $f : \partial A \to \mathbb{R}$, we consider its extention $f \circ p_{\partial A}$ in a neighbourhood $V(x)$ of a point $x \in \partial A$. The so-called *tangential gradient* $\nabla_{\partial A} f$ of f and the tangential *Jacobian matrix* $D_{\partial A} \vec{v}$ of a vector function $\vec{v} : \partial A \to \mathbb{R}^N$ are given by the usual gradient and Jacobian matrix of their extension, that is,

$$\nabla_{\partial A} f = \nabla (f \circ p_{\partial A})|_{\partial A} \text{ and } D_{\partial A} f = D(f \circ p_{\partial A})|_{\partial A} \text{ on } V(x) \cap \partial A. \tag{8.18}$$

The *tangential Laplacian* or *Laplace-Beltrami* $\Delta_{\partial A} f$ is given by

$$\Delta_{\partial A} f = \Delta(f \circ p_{\partial A})|_{\partial A} \text{ on } V(x) \cap \partial A. \tag{8.19}$$

Furthermore, since $\nabla b_A = \nabla b_A \circ p_{\partial A} = n \circ p_{\partial A}$, n the unit exterior normal, in the neighborhood $V(x)$, the *second fundamental form* of ∂A is given by

$$D_{\partial A} n = D_{\partial A}(\nabla b_A) = D(\nabla b_A \circ p_{\partial A})|_{\partial A} = D(\nabla b_A)|_{\partial A} = D^2 b_A|_{\partial A} \tag{8.20}$$

(cf. Delfour–Zolésio ([19, 25]) and Delfour ([14–16]) for a complete account of this approach with applications to the theory of thin and asymptotic shells).

8.4.2 Some Observations: Towards a Classification of Sets

There is an interesting pattern emerging from the previous section: smoothness of the boundary in terms of the smoothness of b_A in a neighbourhood of ∂A and the boundary ∂A has zero volume, that is, zero Lebesgue measure. Yet, the equivalence breaks down for sets of class $C^{1,1-\varepsilon}$, $0 \leq \varepsilon < 1$, as can be seen from the example of the hypergraph of a $C^{1,1-\varepsilon}$-function (cf. [25, Example 6.2, Chap. 6, p. 313]). The boundary is a $C^{1,1-\varepsilon}$ submanifold, but the distance function is not even C^1 in any neighbourhood of the point $(0, 0)$.

Example 8.4 Consider the two-dimensional domain Ω defined as the epigraph of the function f:

$$\Omega \stackrel{\text{def}}{=} \{(x, z) \,:\, f(x) > z,\ \forall x \in \mathbb{R}\}, \quad f(x) \stackrel{\text{def}}{=} |x|^{2-\frac{1}{n}}$$

for some arbitrary integer $n \geq 1$. Ω is a set of class $C^{1,1-1/n}$ and $\partial\Omega$ is a $C^{1,1-1/n}$-submanifold of dimension 1. In view of the presence of the absolute value in $(0, 0)$, it is the point where the smoothness of $\partial\Omega$ is minimum. We claim that

$$\mathrm{Sk}(\partial\Omega) = \mathrm{Sk}(\Omega) = \{(0, y) \,:\, y > 0\} \quad \text{and} \quad \partial\Omega \cap \overline{\mathrm{Sk}(\partial\Omega)} = (0, 0) \neq \varnothing.$$

Because the skeleton is a line touching $\partial\Omega$ in $(0, 0)$, $d^2_{\partial\Omega}$ and d^2_{Ω} cannot be C^1 in any neighborhood of $(0, 0)$. □

For sets such that $\partial A \neq \varnothing$, the smoothness of the boundary around a point $x \in \partial A$ is related to the smoothness of b_A in a neighbourhood of x:

$$\partial A \text{ locally } C^{1,1} \text{ at } x \iff$$
$$\exists \rho > 0 \text{ such that } b_A \in C^{1,1}(\overline{B_\rho(x)}) \text{ and } \mathrm{m}_N(\partial A \cap B_\rho(x)) = 0.$$

This extends the older result of Gilbart and Trudinger: for $k \geq 2$ and $x \in \partial A$

$$\partial A \text{ locally } C^k \text{ at } x \iff$$
$$\exists \rho > 0 \text{ such that } b_A \in C^k(\overline{B_\rho(x)}) \text{ and } \mathrm{m}_N(\partial A \cap B_\rho(x)) = 0.$$

The condition that ∂A has zero measure is often missing in the literature. So the first step in introducing a metric on sets with smooth boundaries is to restrict the family of oriented distance functions to

$$\boxed{C_b^0(D) \stackrel{\text{def}}{=} \{b_A \in C_b(D) : \mathrm{m}_N(\partial A) = 0\}}$$

which happens to be a closed subset of $C^{0,1}(D)$ and the metric topology associated with $W^{1,p}(D)$ since the other distance functions can be recovered from the map

$$b_A \mapsto (b_A^+, b_A^-, |b_A|) = (d_A, d_{\complement A}, d_{\partial A})$$

and the characteristic functions from the maps

$$b_A \mapsto b_A^- = d_{\complement A} \mapsto \chi_{\text{int } A} = |\nabla d_{\complement A}|, \tag{8.21}$$

$$b_A \mapsto b_A^+ = d_A \mapsto \chi_{\text{int } \complement A} = |\nabla d_A|, \tag{8.22}$$

$$b_A \mapsto \chi_{\partial A} = 1 - |\nabla b_A|. \tag{8.23}$$

One of the advantages of the function b_A is that the $W^{1,p}$-convergence of sequences will imply the L^p-convergence of the characteristic functions of int A, int $\complement A$, and ∂A, that is, continuity of the volume of those sets.

The second step is to enrich the $C^{0,1}$ and $W^{1,p}$ norms with some seminorms on the derivatives of b_A as in the construction of norms on Sobolev spaces and k-differentiable functions. Unfortunately, it is not that simple since the oriented distance function b_A of an open domain with a C^∞ boundary is not necessarily C^∞ in $\mathbb{R}^N$. It is C^∞ only in a neighbourhood of ∂A.

8.4.3 Smooth Truncation in a Narrow Band Around the Boundary

Let $k > 0$ be an arbitrary number and consider the *tubular neighbourhood* of ∂A:

$$U_k(\partial A) \stackrel{\text{def}}{=} \left\{x \in \mathbb{R}^N : |b_A(x)| < k\right\}.$$

Given a (small) number $h > 0$, let $\beta : \mathbb{R} \to 0, 1$ be an infinitely differentiable *cut-off function* such that

$$\beta(y) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } |y| \le h, \\ \in (0,1), & \text{if } h < |y| < 2h, \\ 0, & \text{if } |y| \ge 2h. \end{cases}$$

Associate with b_A the *truncated oriented distance function*

$$\boxed{b_A^h(x) \stackrel{\text{def}}{=} \beta(b_A(x))\, b_A(x)} \quad \Rightarrow b_A^h = b_A \text{ in } U_h(\partial A)$$

The function b_A^h coincides with b_A in the *narrow band* $U_h(\partial A)$ around ∂A and is zero outside of $U_{2h}(\partial A)$ where the skeleton might be present. The function b_A^h was introduced in [21] in 2004. Consider the families of sets

$$C_{b^h}^{1,1} \stackrel{\text{def}}{=} \left\{ b_A \in C_b^0(D) : b_A^h \in C^{1,1}(\overline{D}) \right\},$$
$$C_{b^h}^{k} \stackrel{\text{def}}{=} \left\{ b_A \in C_b^0(D) : b_A^h \in C^{k}(\overline{D}) \right\},\ k \ge 2,$$

and, for D bounded, the metrics

$$\rho_{C^{1,1}}(A, B) \stackrel{\text{def}}{=} \|b_A - b_B\|_{C^{0,1}(D)} + \|b_A^h - b_B^h\|_{C^{1,1}(D)}$$
$$\rho_{C^{k}}(A, B) \stackrel{\text{def}}{=} \|b_A - b_B\|_{C^{0,1}(D)} + \|b_A^h - b_B^h\|_{C^{k}(D)},\ k \ge 2.$$

Similarly, with Sobolev spaces

$$C_{b^h}^{W^{m,p}} \stackrel{\text{def}}{=} \left\{ b_A \in C_b^0(D) : b_A^h \in W^{m,p}(D) \right\},$$
$$\rho_{W^{m,p}}(A, B) \stackrel{\text{def}}{=} \|b_A - b_B\|_{W^{1,p}(D)} + \|b_A^h - b_B^h\|_{W^{m,p}(D)},\ m \ge 1,\ p \ge 1.$$

The choice of β is not unique. For instance, we could use two parameters $0 < h < k$ instead of $0 < h < 2h$. The spaces depend on β, but it seems to be the minimal price to pay to effectively control the smoothness of the boundary through a global smoothness condition on D. To our best knowledge, it is the first time that a metric is defined on families of sets with smooth boundaries without constraints on the topology of the sets such as, for instance, in the case of the images of a fixed set endowed with Courant metrics.

9 Sets of Positive Reach, Submanifolds, Squared Distance Functions

9.1 Sets of Positive Reach

Roughly speaking a subset A of $\mathbb{R}^N$ is of *positive reach* if there exists $h > 0$ such that the projection onto $\overline{A}$ of all points in the tubular neighbourhood $U_h(A)$ is unique. Such sets have been introduced by Federer [31] in 1959 in the context of curvature measures (cf. also [25, Sect. 6, Chap. 6]).

The projection $p_A(x)$ onto $\overline{A}$ is directly connected with the square of the distance function: the set $\Pi_A(x)$ of projections corresponds to the minimizers in the following problem

$$\inf_{x \in \overline{A}} \|a - x\| = d_A(x).$$

For instance, when A is convex, the projection is unique and denoted by $p_A(x)$ for each point x in $\mathbb{R}^N$. The function

$$x \mapsto f_A(x) \stackrel{\text{def}}{=} \frac{1}{2}\left(\|x\|^2 - d_A^2(x)\right) : \mathbb{R}^N \to \mathbb{R} \tag{9.1}$$

is convex and continuous. Its Hadamard semi-differential $d_H f_A(x; v)$ exists at each point $x \in \mathbb{R}^N$ and in all directions $v \in \mathbb{R}^N$ and

$$d_H f_A(x; v) = \max_{p \in \Pi_A(x)} p \cdot v \quad \Rightarrow d_H d_A^2(x; v) = 2 \min_{p \in \Pi_A(x)} (x - p) \cdot v. \tag{9.2}$$

The distributional gradients of f_A and d_A^2 belong to $\mathrm{BV}_{\mathrm{loc}}(\mathbb{R}^N)^N$ (cf [25, Chap. 6, Theorems 3.2(ii) and 3.3(ii)–(iv)]).

When $\Pi_A(x)$ is a singleton, $\Pi_A(x) = \{p_A(x)\}$, then

$$d_H f_A(x; v) = p_A(x) \cdot v \quad \Rightarrow \nabla f_A(x) = p_A(x) \tag{9.3}$$

and f_A and d_A^2 are both Fréchet différentiable in x. In particular,

$$\nabla d_A^2(x) = 2\,(x - p_A(x)) \quad \Rightarrow p_A(x) = x - \frac{1}{2}\nabla d_A^2(x) = \nabla f_A(x)$$

and $\nabla d_A(x)$ exists for $x \notin \partial\overline{A}$. From [25, Theorem 6.3, Chap. 6] we have the following equivalence: there exists $h > 0$ such that the projection of points in $U_h(A)$ onto $\overline{A}$ is unique if and only if for all k, $0 < k < h$, $d_A^2 \in C^{1,1}(\overline{U_k(A)})$. For convex sets $h = +\infty$ and $d_A^2 \in C^{1,1}_{\mathrm{loc}}(\mathbb{R}^N)$.

The following simple compactness theorem for sets of positive reach was given in H. Federer [31, Theorem 4.13, Remark 4.14] (cf. also, [24, Theorem 8.1, Sect. 8, Chap. 4, p. 194]).

Theorem 9.1 *Let D be a fixed bounded open subset of $\mathbb{R}^N$. Let $\{A_n\}$, $A_n \neq \varnothing$, be a sequence of subsets of $\overline{D}$. Assume that there exists $h > 0$ such that*

$$\forall n, \quad d_{A_n}^2 \in C^{1,1}(\overline{U_h(A_n)}). \tag{9.4}$$

Then there exist a subsequence $\{A_{n_k}\}$ and $A \subset \overline{D}$, $A \neq \varnothing$, such that $d_A^2 \in C^{1,1}(\overline{U_h(A)})$ and

$$d_{A_{n_k}}^2 \to d_A^2 \text{ in } C^1(\overline{U_h(A)}).$$

This is a compactness theorem similar to the compactness of $C_d(D)$ in $C^0(D)$ for some bounded open subset D of $\mathbb{R}^N$.

From this, we can introduce metric spaces of sets of positive reach.

Theorem 9.2 (cf. [25, Theorems 6.6 and 6.7, Sect. 6.3, Chap. 6]) *Given an open* (resp. *bounded open*) *holdall D in $\mathbb{R}^N$ and $h > 0$, the family*

$$C_{d,h}(D) \stackrel{def}{=} \left\{ d_A : \varnothing \neq \overline{A} \subset \overline{D},\ d_A \in C^{1,1}(\overline{U_h(A)}) \right\} \tag{9.5}$$

is closed in $C_{loc}(D)$ (resp. *compact in $C(\overline{D})$*).

9.2 C^k-Submanifolds, $k \geq 2$

We can eliminate all sets of positive reach such that $\partial\overline{A} = \varnothing$, since either $A = \varnothing$ and $d_A = +\infty$ or $A = \mathbb{R}^N$ and $d_A = 0$.

If $\partial\overline{A} \neq \varnothing$, we have two cases: either $\operatorname{int}\overline{A} \neq \varnothing$ or $\operatorname{int}\overline{A} = \varnothing$, that is, $\overline{A} = \partial\overline{A}$. The first case includes all convex sets with a non-empty interior, but, in general, it cannot occur under the smoother assumption $d_A^2 \in C^2(\overline{U_h(A)})$.

Theorem 9.3 *Let $A \subset \mathbb{R}^N$, $\partial\overline{A} \neq \varnothing$, and assume that*

$$\exists h > 0 \text{ such that } d_A^2 \in C^2(\overline{U_h(A)}). \tag{9.6}$$

Then $\partial\overline{A} = \overline{A}$, $\overline{\complement\overline{A}} = \mathbb{R}^N$, $\operatorname{int}\overline{A} = \varnothing$, and $\mathrm{m}_N(\partial\overline{A}) = 0$.

Proof By definition of f_A in terms of d_A^2, $f_A \in C^2(\overline{U_h(A)})$. Hence, the projection onto $\overline{A}$ is $p_A(x) = \nabla f_A(x) \in C^1(\overline{U_h(A)})$, and $Dp_A(x) = D^2 f_A(x) \in C^0(\overline{U_h(A)})$. As such $p_A \circ p_A = p_A$,

$$D^2 f_A \circ p_A \, D^2 f_A = D^2 f_A \quad \Rightarrow \; D^2 f_A \circ p_A \, D^2 f_A \circ p_A = D^2 f_A \circ p_A,$$

and $D^2 f_A \circ p_A$ is an *orthogonal projector* of $\mathbb{R}^N$ onto Im $D^2 f_A \circ p_A$. For $x \in \text{int } \overline{A}$, $\nabla d_A^2(x) = 0$ and $D^2 d_A^2(x) = 0$ which implies that $\nabla f_A(x) = x$ and $D^2 f_A(x) = I$. In particular, by continuity, $D^2 f_A(x) = I$ and $D^2 d_A^2(x) = 0$ for all $x \in \partial\overline{A}$.

Now,

$$d_A^2(x) = \|x - p_A(x)\|^2 = \left\| \frac{1}{2} \nabla d_A^2(x) \right\|^2 \quad \Rightarrow \; \nabla d_A^2(x) = \frac{1}{2} D^2 d_A^2(x) \, \nabla d_A^2(x).$$

For $x \in U_h(A) \backslash \overline{A}$, $d_A(x) > 0$, $\nabla d_A(x)$ exists, $\|\nabla d_A(x)\| = 1$, and

$$\nabla d_A(x) = \frac{\nabla d_A^2(x)}{2 d_A(x)} = \frac{1}{2} D^2 d_A^2(x) \, \frac{\nabla d_A^2(x)}{2 d_A(x)} = \frac{1}{2} D^2 d_A^2(x) \, \nabla d_A(x).$$

The ray $x_t = p_A(x) + t \, \nabla d_A(x)$, $0 < t \le d_A(x)$, belongs to $U_h(A) \backslash \overline{A}$ and $p_A(x_t) = p_A(x)$ and $\nabla d_A(x_t) = \nabla d_A(x)$ with $\|\nabla d_A(x)\| = 1$. As a result, by continuity of $D^2 d_A^2$,

$$0 \neq \nabla d_A(x) = \frac{1}{2} D^2 d_A^2(x_t) \, \nabla d_A(x) \to \frac{1}{2} D^2 d_A^2(p_A(x)) \, \nabla d_A(x) = 0$$

which yields a contradiction. Therefore, int $\overline{A} = \varnothing$.

If $\text{m}_N(\overline{A}) > 0$,

$$0 < \text{m}_N(\overline{A}) = \text{m}_N(\partial\overline{A}) + \text{m}_N(\text{int } \overline{A}) = \text{m}_N(\text{int } \overline{A}) \quad \Rightarrow \; \text{int } \overline{A} \neq \varnothing$$

which contradicts the first part of the proof. □

So the smoothness of the square of the distance function is more appropriate for submanifolds[11] of $\mathbb{R}^N$ of co-dimension greater or equal to one than for sets with a non-empty interior.[12]

Theorem 9.4 *Let $A \subset \mathbb{R}^N$ such that $\partial\overline{A} \neq \varnothing$ and let $k \ge 2$, be an integer. The following conditions are equivalent:*

(i) there exists $h > 0$ such that $d_A^2 \in C^k(\overline{U_h(A)})$;
(ii) $\overline{A}$ is a C^k-submanifold of $\mathbb{R}^N$ of dimension less or equal to $N - 1$.

[11] The notion of submanifold of $\mathbb{R}^n$ used here is the one of Berger and Gostiaux [8, Definition 2.1.1 and Theorem 2.1.2 in Chap. 2]. It coincides with the notion of *embedded submanifold* of J.M. Lee [42, p. 174, Chap. 8].

[12] In absence of a precise reference, Poly and Raby [56] (see also [25, Chap. 6, Sect. 6.2, p. 310–315]) gave a proof of the following *folk theorem* for $2 \le k \le +\infty$ and A closed:

$$A \text{ is locally a } C^k\text{-submanifold at } x \iff d_A^2 \in C^k(V(x)) \text{ in some neighbourhood of } x.$$

.

Under the above conditions, $\overline{A} = \partial\overline{A}$, $\operatorname{int}\overline{A} = \varnothing$, $\overline{\complement A} = \mathbb{R}^N$, $\mathrm{m}_N(\overline{A}) = 0$, *the tangent space at* $x \in \overline{A}$ *is* $\operatorname{Im} D^2 f_A(x)$, *and the dimension of* $\overline{A}$ *is equal to the dimension of* $\operatorname{Im} D^2 f_A(x)$.

Proof Cf. [25, Theorem 6.5, Sect. 6.2, Chap. 6]. □

The range between $C^{1,1}$ and C^2 still needs to be investigated. Another issue that we shall address in the next section is the fact that a local condition such as $d_A^2 \in C^k(\overline{U_h(A)})$ is not very convenient. This can be fixed by putting the condition on the function $d_A^h(x) = \beta(d_A(x))\, d_A(x)$ similar to the function b_A^h introduced in Sect. 8.4.3.

Note that we can replace A by ∂A and d_A by $d_{\partial A}$ in the above considerations and start with b_A since $d_{\partial A}^2 = b_A^2$.

Theorem 9.5 *Let* $A \subset \mathbb{R}^N$ *such that* $\partial A \neq \varnothing$ *and let* $k \geq 2$, *be an integer. The following conditions are equivalent:*

(i) there exists $h > 0$ *such that* $b_A^2 \in C^k(\overline{U_h(\partial A)})$;

(ii) ∂A *is a* C^k*-submanifold of* $\mathbb{R}^N$ *of dimension less or equal to* $N - 1$.

Under the above conditions, $\partial A = \partial(\partial A)$, $\operatorname{int}\partial A = \varnothing$, $\mathrm{m}_N(\partial A) = 0$, *the tangent space at* $x \in \partial A$ *is* $\operatorname{Im} D^2 f_{\partial A}(x)$, *and the dimension of* ∂A *is equal to the dimension of* $\operatorname{Im} D^2 f_{\partial A}(x)$.

9.3 Metric on Families of Submanifolds

An *(embedded) submanifold*[13] A of $\mathbb{R}^N$ has the property that $\overline{\complement A} = \mathbb{R}^N$ which implies that

$$b_A = d_A \iff d_{\complement A} = 0 \iff \overline{\complement A} = \mathbb{R}^N \iff \overline{A} = \partial A.$$

In addition, it is expected that $\mathrm{m}_N(A) = 0$. Note that the fact that $\overline{A} = \partial A$ does not imply that the Lebesgue measure of ∂A is zero.

We have the following unexpected and apparently new results that emphasizes the fundamental role of b_A over d_A and the natural transition from one to the other.

Theorem 9.6 *Let* $D \subset \mathbb{R}^N$ *be a bounded open hold-all in* $\mathbb{R}^N$.

(i) The family

$$S_d(D) \stackrel{def}{=} \left\{ d_A : \varnothing \neq A \subset \overline{D},\ \overline{\complement A} = \mathbb{R}^N \right\}$$

is a closed subset of $C_d(D)$ *with the* C^0*-norm.*

[13] Cf. Footnote 11 for the definition of an (embedded) submanifold.

(ii) The subfamily

$$S_d^0(D) \overset{def}{=} \left\{ d_A : \varnothing \neq A \subset \overline{D},\ \overline{\complement A} = \mathbb{R}^N \text{ and } \mathrm{m}_N(A) = 0 \right\}$$

is a closed subset of $C_d^0(D)$ with the $C^{0,1}$, or $W^{1,p}$-metrics.

Proof (i) Let $\{d_{A_n}\}$, $A_n \neq \varnothing$, be a Cauchy sequence in $S_d^0(D)$ in the C^0-metric. By definition, $\partial A_n = \overline{A_n} \neq \varnothing$. Since $b_{A_n} = d_{A_n}$, $\{b_{A_n}\}$ is also a Cauchy sequence in the C^0-metric. By completeness of $C_b(D)$, there exists $A \subset \overline{D}$ such that $\partial A \neq \varnothing$ and $b_{A_n} \to b_A$ in $C^0(D)$. But $0 = d_{\complement A_n} \to d_{\complement A}$ and $\overline{\complement A} = \mathbb{R}^N$ and $d_{A_n} \to d_A$ in $C^0(D)$.

(ii) Same proof as in (i) with the additional property that, by continuity, the measure $\mathrm{m}_N(A) = \mathrm{m}_N(\partial A) = 0$ of the boundary is preserved in the $C^{0,1}$ or $W^{1,p}$-metrics. □

As seen in Theorem 9.4, the smoothness of a set A such that $\overline{A} = \partial A$ is characterized by the smoothness of d_A^2 rather than the one of b_A.

To our best knowledge, the following complete metric spaces are new and provide a powerful framework for optimization problems where the smoothness of the boundary needs to be controlled without constraint on the topology of the sets. Consider the families (recall that $b_A = d_A$ so that $d_A^h = b_A^h$)

$$C_{d^h}^{1,1} \overset{\text{def}}{=} \left\{ d_A \in S_d^0(D) : (d_A^h)^2 \in C^{1,1}(\overline{D}) \right\},$$
$$C_{d^h}^{k} \overset{\text{def}}{=} \left\{ d_A \in S_b^0(D) : (d_A^h)^2 \in C^{k}(\overline{D}) \right\},\ k \geq 2,$$

and, for D bounded, the metrics

$$\rho_{C^{1,1}}(A, B) \overset{\text{def}}{=} \|d_A - d_B\|_{C^{0,1}(D)} + \|(d_A^h)^2 - (d_B^h)^2\|_{C^{1,1}(D)}$$
$$\rho_{C^{k}}(A, B) \overset{\text{def}}{=} \|d_A - d_B\|_{C^{0,1}(D)} + \|(d_A^h)^2 - (d_B^h)^2\|_{C^{k}(D)},\ k \geq 2.$$

Similarly, for D bounded open, with Sobolev spaces

$$S_{d^h}^{W^{m,p}} \overset{\text{def}}{=} \left\{ d_A \in S_b^0(D) : (d_A^h)^2 \in W^{m,p}(D) \right\},$$
$$\rho_{W^{m,p}}(A, B) \overset{\text{def}}{=} \|d_A - d_B\|_{W^{1,p}(D)} + \|(d_A^h)^2 - (d_B^h)^2\|_{W^{m,p}(D)},\ m \geq 1,\ p \geq 1.$$

They are all complete metric spaces. The smoothness of d_A^2 in a *narrow band* around A effectively controls the smoothness of $\partial A = \overline{A}$ without imposing a constraint on the topology as in the case of Courant metric spaces of images.

10 Metric Structure via the Support Function

To further illustrate the use of set parametrized functions, we give a last example for families of closed convex sets. Associate with a set $\varnothing \neq A \subset \mathbb{R}^N$, its *support function* and the following equivalence classes

$$\sigma_A(x) \stackrel{\text{def}}{=} \sup_{y \in A} x \cdot y, \quad [A] \stackrel{\text{def}}{=} \left\{ A' : \overline{\text{co}}\, A' = \overline{\text{co}}\, A \right\}.$$

Recall that

- σ_A is *convex, upper semicontinuous* and *positively homogeneous*.
- σ_A is *uniformly Lipschitz continuous* if and only if A is bounded.

Identify

$$\boxed{[A] \longleftrightarrow \sigma_A}$$

From [18], the family

$$C_\sigma(\mathbb{R}^N) \stackrel{\text{def}}{=} \left\{ \sigma_A; \varnothing \neq A \subset \mathbb{R}^N \text{ and bounded} \right\}$$

is a complete metric space for the metric

$$\rho(A, B) \stackrel{\text{def}}{=} \sup_{x \in B_1(0)} |\sigma_A(x) - \sigma_B(x)|, \quad B_1(0) \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^N : \|x\| < 1 \right\}.$$

If D is bounded, then $C_\sigma(D)$ is compact. This metric structure is particularly interesting for problems where the variable domains are closed convex sets. It is interesting that the family of closed convex subsets of a bounded hold-all D is compact in all the metric topologies associated with χ_A, d_A, b_A. So, it is sufficient to have the lower semicontinuity of the objective functional in any one of those metric space topologies to get the existence of minimizers.

11 Some Concluding Remarks

Several new constructions and metric spaces have been presented in this paper. For instance, the topological group structure of characteristic functions via the symmetric difference can now be preserved for the families of distance (oriented distance) functions of subsets that include the empty set and the whole space by introducing new equivalent metrics that contain the distance (oriented distance) functions of the empty set and the whole space in their completion. Another example are the new metrics that effectively controls the smoothness of variable sets in a narrow band of fixed thickness around the boundary of the sets without constraint on the topology of the sets. They complement the constructions and metric spaces in [25].

It is clear that important *patterns* are emerging. One may or may not want to use them in specific applications, but they exist and it is better to take advantage of them.

- It is possible to construct a *wide range of metrics* and complete metrics spaces. The choice is application dependent.
- The metric spaces associated with the oriented distance function are well adapted to deal with a wide range of families of open sets with a smooth boundary. For instance, the classical uniform cone property, the more recent uniform cusp property, the density perimeter, and the uniform bound on families of sets with bounded curvatures all lead to *optimal* compactness theorems with convergence of the oriented distance function in the $W^{1,p}$-strong metric. By optimal, we mean the strongest convergence. For instance, the uniform cone property was initially proved for the strong convergence of the characteristic function χ_A in L^p, but, in fact, we have the strong convergence of the oriented distance function b_A in $W^{1,p}$ and hence the convergence in all the other topologies related to χ_A, d_A, $d_{\complement A}$, and $d_{\partial A}$.
- In some cases the *group and the metric structures* are sufficiently compatible to introduce some form of *tangent space* (not necessarily a Hilbert or Banach space) and, possibly, *semidifferentials or differentials of functions*.
- Similar techniques can be developed for families of *submanifolds of* $\mathbb{R}^N$ of co-dimension greater or equal to one.

Finally, it is urgent to explore connections with the construction of metrics and metric structures introduced in the work of Gromov [37] and its pressing applications in areas such as in Image Processing by Memoli [45] and Memoli and Shapiro [46].

Acknowledgments This paper has focused on some selected aspects of this very important and fruitful field of activities. Unfortunately, it was not possible to comment on or detail all the interesting and challenging ongoing work taking place. Some papers that have not been cited are included in the list of references but it is very far from a complete list.

References

1. D. Adalsteinsson, J.A. Sethian, The fast construction of extension velocities in level set methods. J. Computat. Phys. **148**, 2–22 (1999)
2. F.J. Almgren Jr, *The Theory of Varifolds—a Variational Calculus in the Large for thek-dImensional Area Integrand*, Mimeographed notes (Princeton University Library, Princeton, 1965)
3. F.J. Almgren Jr, *Plateau's Problem* (W.A. Benjamin, New York, 1966)
4. F.J. Jr Almgren, Existence and regularity almost everywhere of solutions to elliptic variational problems among surfaces of varying topological type and singularity structure. Ann Math **87**(2), 321–391 (1968)
5. J.-P. Aubin, H. Frankowska, *Set-Valued Analysis* (Birkhäuser, Boston, 1990)
6. R. Azencott, Geodesics in diffeomorphisms groups: Deformation distance between shapes, in *International Conference on Stochastic Structures and Monte-Carlo Optimisation, Cortona, Italy* (1994)

7. J.C. Barolet, *Représentation et détection des images et des surfaces déformables, mémoire de maîtrise* Université de Montréal, Département de mathématiques et de statistique (2011)
8. M. Berger, B. Gostiaux, *Differential Geometry: Manifolds, Curves and Surfaces* (Springer, New York, 1988). (French trans: *Géométrie différentielle : variétés, courbes et surfaces*, 2nd edn. (Presses Universitaire de France, Paris, 1987), p. 1992
9. D. Bucur, J.-P. Zolésio, N-dimensional shape optimization under capacitary constraint. J. Differ. Equ. **123**(2), 504–522 (1995)
10. D. Bucur, J.-P. Zolésio, Free boundary problems and density perimeter. J. Differ. Equ. **126**(2), 224–243 (1996)
11. R. Caccioppoli, *Sulla quadratura delle superfici piane e curve*, Atti della Accademia Nazionale dei Lincei. Rendiconti. Classe di Scienze Fisiche, Matematiche e Naturali **VI** (in Italian) 6 (1927), pp. 142–146
12. R. Caccioppoli, *Misura e integrazione sugli insiemi dimensionalmente orientati*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Nat. **12**(8) (1952), 3–11; 137–146
13. D. Chenais, On the existence of a solution in a domain identification problem. J. Math. Anal. Appl. **52**, 189–219 (1975)
14. M.C. Delfour, Tangential differential calculus and functional analysis on a $C^{1,1}$ submanifold, in *Differential-Geometric Methods in the Control of Partial Differential Equations*, vol. 268, Contemporary Mathematics, ed. by R. Gulliver, W. Littman, R. Triggiani (AMS Publications, Providence, 2000), pp. 83–115
15. M.C. Delfour, Intrinsic differential geometric methods in the asymptotic analysis of linear thin shells, *Boundaries, Interfaces, and Transitions (Banff, AB, 1995)*, vol. 13, CRM Proceedings and Lecture Notes (American Mathematical Society, Providence, 1998), pp. 19–90
16. M.C. Delfour, Representation of hypersurfaces and minimal smoothness of the midsurface in the theory of shells. Control Cybern **37**(4), 879–911 (2008)
17. M.C. Delfour, J.-P. Zolésio, Structure of shape derivatives for nonsmooth domains. J. Funct. Anal. **104**(1), 1–33 (1992)
18. M.C. Delfour, J.-P. Zolésio, Shape analysis via oriented distance functions. J. Funct. Anal. **123**(1), 129–201 (1994)
19. M.C. Delfour, J.-P. Zolésio, Differential equations for linear shells: comparison between intrinsic and classical models, *Advances in Mathematical Sciences: CRM's 25 Years (Montreal, PQ, 1994)*, vol. 11, CRM Proceedings and Lecture Notes (American Mathematical Society, Providence, 1997), pp. 41–124
20. M.C. Delfour, J.-P. Zolésio, The new family of cracked sets and the image segmentation problem revisited. Commun. Inf. Syst. **4**(1), 29–52 (2004)
21. M.C. Delfour, J.-P. Zolésio, Oriented distance function and its evolution equation for initial sets with thin boundary. SIAM J. Control Optim. **42**(6), 2286–2304 (2004)
22. M.C. Delfour, J.-P. Zolésio, Evolution equations for shapes and geometries. J. Evol. Equ. **6**(3), 399–417 (2006)
23. M.C. Delfour, J.-P. Zolésio, Uniform fat segment and cusp properties for compactness in shape optimization. Appl. Math. Optim. **55**, 385–419 (2007)
24. M.C. Delfour, J.-P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, 1st edn., SIAM series on Advances in Design and Control (Society for Industrial and Applied Mathematics, Philadelphia, 2001)
25. M.C. Delfour, J.-P. Zolésio, *Shapes and Geometries: Metrics, Analysis, Differential Calculus, and Optimization*, 2nd edn., SIAM series on Advances in Design and Control (Society for Industrial and Applied Mathematics, Philadelphia, 2011)
26. E. De Giorgi, Ennio, *Definizione ed espressione analitica del perimetro di un insieme*. (Definition and analytical expression of the perimeter of a set), Atti della Accademia Nazionale dei Lincei. Rendiconti. Classe di Scienze Fisiche, Matematiche e Naturali, VIII (in Italian) **14** (1953) 390–393
27. E. De Giorgi, Ennio, *Su una teoria generale della misura (r-1)-dimensionale in uno spazio ad r dimensioni*, (On a general theory of (r − 1-dimensional space), Annali di Matematica Pura e Applicata, Serie IV, (in Italian) **36**(1) (1954), 191–213

28. O. Dovgoshey, Certain characterization of Carathéodory domains. Comput. Methods Funct. Theory **5**(2), 489–503 (2005)
29. P. Dupuis, U. Grenander, M. Miller, Variational problems on flows of diffeomorphisms for image matching. Quart. Appl. Math. **56**, 587–600 (1998)
30. L.C. Evans, R.F. Gariepy, Measure theory and fine properties of functions, *Studies in Advanced Mathematics* (CRC Press, Boca Raton, 1992)
31. H. Federer, Curvature measures. Trans. Amer. Math. Soc. **93**, 418–419 (1959)
32. H. Federer, *Geometric Measure Theory* (Springer, Berlin, 1969)
33. H. Federer, W.H. Fleming, Normal and integral currents. Ann. Math. **72**(2), 458–520 (1960)
34. D. Gaier, *Vorlesungen über Approximation im Komplexen* (Birkhäuser Verlag, Basel, 1980)
35. E. Giusti, *Minimal Surfaces and Functions of Bounded Variation* (Birkhäuser, Boston, 1984)
36. J. Glaunès, A. Trouvé, L. Younes, Diffeomorphic matching of distributions: a new approach for unlabelled point-sets and sub-manifolds matching, in *Proceedings CVPR '04, 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2004)
37. M. Gromov, *Metric Structures for Riemannian and non-Riemannian Spaces*, Progress in Mathematics (Birkhäuser Boston Inc., Boston, 1999)
38. F. Hausdorff, *Grundzüge der Meugenlehre*, Walter de Gruyter, Leipzig, 1927; initial edition Leipzig, 1914 (transl. from the German 3rd edition (1937) in English by J. R. Aumann et al., *Set theory*, Chelsea Pub. Co., New York, 1957)
39. A. Henrot, How to prove symmetry in shape optimization problems? Control Cybern. **25**(5), 1001–1013 (1996). Shape optimization and scientific computations
40. A. Henrot, M. Pierre, About critical points of the energy in an electromagnetic shaping problem, *Boundary Control and Boundary Variation*, vol. 178, Lecture Notes in Control and Information Science (Springer, Berlin, 1992), pp. 238–252
41. H. Lebesgue, *Leçons sur l'intégration et la recherche des fonctions primitives* (Gauthier-Villars, Paris, 1904)
42. J.M. Lee, *Introduction to Smooth Manifolds* (Springer, New York, 2003)
43. H. Lebesgue, Sur l'intégration des fonctions discontinues. Ann. Sci. École Norm. Sup. **27**(3), 361–450 (1910)
44. W. Liu, P. Neittaanmäki, D. Tiba, Sur les problèmes d'optimisation structurelle. C. R. Acad. Sci. Paris Sér. I Math. **331**, 101 (2000)
45. F. Mémoli, On the use of Gromov-Hausdorff distances for shape comparison, in *Proceedings of PBG 2007* Prague, Czech Republic (2007)
46. F. Mémoli, G. Sapiro, A theoretical and computational framework for isometry invariant recognition of point cloud data. Found. Comput. Math. **5**(3), 313–347 (2005)
47. A.M. Micheletti, Metrica per famiglie di domini limitati e proprietà generiche degli autovalori. Ann. Scuola Norm. Sup. Pisa **26**(3), 683–694 (1972)
48. P.W. Michor, D. Mumford, Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms. Doc. Math. **10**, 217–245 (2005)
49. P.W. Michor, D. Mumford, Riemannian geometries on spaces of plane curves. J. Eur. Math. Soc. (JEMS) **8**(1), 1–48 (2006)
50. P.W. Michor, D. Mumford, An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach. Appl. Comput. Harmon. Anal. **23**(1), 74–113 (2007)
51. M.I. Miller, A. Trouvé, L. Younes, Computing large deformations via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. **61**(2), 139–157 (2005)
52. M.I. Miller, A. Trouvé, L. Younes, On the metrics and Euler-Lagrange equations of computational anatomy. Annu. Rev. Biomed. Eng. **4**, 375–405 (2002)
53. M.I. Miller, L. Younes, Group action, diffeomorphism and matching: a general framework. Int. J. Comput. Vis. **41**, 61–84 (2001)
54. F. Murat, J. Simon, *Sur le contrôle par un domaine géométrique, Rapport 76015* (Université Pierre et Marie Curie, Paris, 1976)
55. P. Neittaanmäki, J. Sprekels, D. Tiba, *Optimization of Elliptic Systems. Theory and applications*, Springer Monographs in Mathematics (Springer, New York, 2006)
56. J.-B. Poly, G. Raby, Fonction distance et singularités. Bull. Sci. Math. **108**(2), 187–195 (1984)

57. D. Pompéiu, Sur la continuité des variables complexes. Ann. Fac. Sci. de Toulouse Sci. Math. Sci. Phys. **7**(2), 265–315 (1905)
58. E.R. Reifenberg, Parametric surfaces. IV. The generalised Plateau problem. J. Lond. Math. Soc. **27**, 448–456 (1952)
59. E.R. Reifenberg, Parametric surfaces. V. Area. II. Proc. Lond. Math. Soc. **5**(3), 342–357 (1955)
60. E.R. Reifenberg, Solution of the Plateau problem for m-dimensional surfaces of varying topological type. Acta Math. **104**, 1–92 (1960)
61. E.R. Reifenberg, Solution of the Plateau problem for m-dimensional surfaces of varying topological type. Bull. Amer. Math. Soc. **66**, 312–313 (1960)
62. D. Tiba, A property of Sobolev spaces and existence in optimal design. Appl. Math. Optim. **47**, 45–58 (2003)
63. A. Trouvé, Action de groupe de dimension infinie et reconnaissance de formes. C. R. Acad. Sci. Paris Sér. I Math. **321**(8), 1031–1034 (1995)
64. A. Trouvé, *An approach of pattern recognition through infinite dimensional group actions* (Rapport de recherche du LMENS, France, 1995)
65. A. Trouvé, An infinite dimensional group approach for physics based models in patterns recognition, unpublished paper, November 1995
66. A. Trouvé, Diffeomorphisms groups and pattern matching in image analysis. Int. J. Comput. Vis. **28**(3), 213–221 (1998)
67. A. Trouvé, L. Younes, On a class of diffeomorphic matching problems in one dimension. SIAM J. Control Optim. **39**(4), 1112–1135 (2000)
68. L. Younes, A distance for elastic matching in object recognition. C. R. Acad. Sci. Paris Sér. I Math. **322**(2), 197–202 (1996)
69. L. Younes, Computable elastic distances between shapes. SIAM J. Appl. Math. **58**(2), 565–586 (1998)
70. L. Younes, Optimal matching between shapes via elastic deformations. Image Vis. Comput. **17**, 381–389 (1999)
71. L. Younes, *Invariance, déformations et reconnaissance de formes. Mathématiques & Applications (Berlin) [Mathematics & Applications]*, vol. 44 (Springer, Berlin, 2004)
72. L. Younes, Jacobi fields in groups of diffeomorphisms and applications. Quart. Appl. Math. **65**(1), 113–134 (2007)
73. L. Younes, *Shape and Diffeomorphisms*, vol. 171, Applied Mathematical Sciences (Springer, Berlin, 2010)
74. L. Younes, P. W. Michor, I. Shah, and D. Mumford, *A metric on shape space with explicit geodesics*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. **19**(1) (2008), 25–57

A Phase Field Approach for Shape and Topology Optimization in Stokes Flow

Harald Garcke and Claudia Hecht

Abstract A new formulation for shape and topology optimization in a Stokes flow is introduced. The investigated problem minimizes the total potential power of the flow. By combining a porous medium and a phase field approach we obtain a well-posed problem in a diffuse interface setting that can be reformulated into a problem without state equations. We can derive a sharp interface problem with zero permeability outside the fluid region as a Γ-limit of this porous medium—phase field problem.

Keywords Shape and topology optimization · Phase field method · Diffuse interfaces · Stokes flow · Fictitious domain · Γ-convergence

AMS Subject Classification 35R35 · 35Q35 · 49Q10 · 49Q20 · 76D07

1 Introduction

By shape optimization in fluids one generally refers to the problem of finding a shape of a fluid region, or of an obstacle inside a fluid, respectively, such that a certain objective functional is minimal. Often, one does not want to prescribe the topology of this region in advance, as one may not know how many connected components or holes of the shape are optimal for instance. There are several well-developed approaches for shape and topology optimization when it comes to finding the optimal configuration in a mixture of several conducting or elastic materials, see [4]. But even though there are numerous applications in the field of shape optimization in fluids, such as optimizing airplanes and cars, biomechanical design or several problems in the machine industry, the mathematical theory is not yet so elaborated than in other areas of shape and topology design. In industry, like in aerospace

H. Garcke (✉) · C. Hecht
Fakultät für Mathematik, Universität Regensburg, 93040 Regensburg, Germany
e-mail: harald.garcke@ur.de

C. Hecht
e-mail: claudia.hecht@ur.de

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_5

engineering, practical methods are quite sophisticated and one can find many mathematical contributions to those numerical methods, for instance in the field of shape sensitivity analysis. However, even basic mathematical questions like the existence of a minimizer remain open. In general, shape optimization problems are known to be not well-posed, see for instance [5], and hence several ideas have been developed in different areas to overcome this issue. Important contributions for this can be found in the field of finding optimal material configurations. Mentionable are certainly the ideas of using a perimeter penalization in optimal shape design and considering this problem in the framework of Caccioppoli sets, see [2], and of introducing a so-called ersatz material approach, see [8]. The latter replaces the void regions by a fictitious material which may be very weak for instance, see [1]. A comparable idea in a fluid dynamical setting has been proposed by [7], where the non-fluid region is replaced by a porous medium with small permeability. In this work, we extend this porous medium approach by including an additional perimeter penalization in order to arrive in a problem that can be generalised to nonlinear state equations and a large class of objective functionals. Anyhow, in this work we introduce this formulation by means of the well-known problem of minimizing the total potential power in a Stokes flow. This yields in particular a special structure of the problem where the state equations can be dropped. This is the best understood setting in shape optimization problems, see also comparable settings in material design [2, 8]. The design variable in the porous medium approach does not only take two discrete values for material and fluid, but can also have values in between and hence we obtain a diffuse interface. Consequently, also the perimeter functional is replaced by a functional, here the Ginzburg-Landau energy, corresponding to the perimeter on the diffuse interface level. The resulting porous medium—phase field problem will be introduced and discussed in more detail in Sect. 2 and can be roughly outlined as

$$\begin{aligned} &\min_{(\varphi,\boldsymbol{u})} \int_\Omega \frac{1}{2}\alpha_\epsilon(\varphi)\,|\boldsymbol{u}|^2 + \frac{\mu}{2}\,|\nabla \boldsymbol{u}|^2 - \boldsymbol{f}\cdot\boldsymbol{u} + \frac{\gamma\epsilon}{2}\,|\nabla\varphi|^2 + \frac{\gamma}{\epsilon}\psi(\varphi)\,\mathrm{d}x \\ &\text{subject to } \int_\Omega \alpha_\epsilon(\varphi)\,\boldsymbol{u}\cdot\boldsymbol{v} + \mu\nabla\boldsymbol{u}\cdot\nabla\boldsymbol{v}\,\mathrm{d}x = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x \quad \forall \boldsymbol{v}. \end{aligned}$$

This problem is formulated in more detail in (7) and we will show that it admits a minimizer. Additionally, we prove that the objective functional Γ-converges as the interfacial thickness tends to zero to a perimeter penalized sharp interface shape optimization problem where in particular the permeability of the non-fluid region is zero. The sharp interface problem, which is described in more detail in (11) and (12) in Sect. 3, is in a simplified form given as

$$\begin{aligned} &\min_{(\varphi,\boldsymbol{u})} \int_{\{\varphi=1\}} \frac{\mu}{2}\,|\nabla\boldsymbol{u}|^2 - \boldsymbol{f}\cdot\boldsymbol{u}\,\mathrm{d}x + \gamma c_0 P_\Omega(\{\varphi=1\}) \\ &\text{subject to } \int_{\{\varphi=1\}} \mu\nabla\boldsymbol{u}\cdot\nabla\boldsymbol{v}\,\mathrm{d}x = \int_{\{\varphi=1\}} \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x \quad \forall \boldsymbol{v}. \end{aligned}$$

2 Porous Medium—Phase Field Formulation

The investigated problem in this work is to minimize a certain objective functional, depending on the velocity of some fluid, by optimizing the shape, geometry and topology of the region which is filled with this fluid. This region can be chosen in a large class of admissible shapes but has to stay inside a given, fixed holdall container $\Omega \subset \mathbb{R}^d$ which is chosen such that

(A1) $\Omega \subseteq \mathbb{R}^d$, $d \in \{2, 3\}$, is a bounded Lipschitz domain with outer unit normal $\boldsymbol{n}$.

The velocity $\boldsymbol{u}$ and the pressure p of the fluid, whose viscosity is denoted by $\mu > 0$, are described by the Stokes equations

$$-\mu \Delta \boldsymbol{u} + \nabla p = \boldsymbol{f}, \qquad \operatorname{div} \boldsymbol{u} = 0 \tag{1}$$

inside the fluid region. We use Dirichlet boundary conditions on the boundary of Ω, hence we may prescribe some in-or outflow region on the boundary. Additionally, we allow here external forces $\boldsymbol{f}$ to act on the whole of Ω.

(A2) Let $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ denote the applied body force and $\boldsymbol{g} \in \boldsymbol{H}^{\frac{1}{2}}(\partial\Omega)$ the given boundary function such that $\int_{\partial\Omega} \boldsymbol{g} \cdot \boldsymbol{n} \, \mathrm{d}x = 0$.

We remark, that throughout this work $\mathbb{R}^d$-valued functions of function spaces of $\mathbb{R}^d$-valued functions are denoted by boldface letters.

The design variable describing the regions filled with fluid and the ones not filled with fluid is in general denoted by φ and is chosen in $H^1(\Omega)$. As already indicated in the introduction, we do not only allow φ to take the values that correspond to fluid regions (which means $\varphi = 1$) and non-fluid regions (hence $\varphi = -1$), but also values in between (i.e. $|\varphi| < 1$) and so we arrive in a diffuse interface setting. Additionally, we want to include a volume constraint on the design variable, so we only optimize over all $\varphi \in H^1(\Omega)$ fulfilling $\int_\Omega \varphi \, \mathrm{d}x \leq \beta |\Omega|$. The constant $\beta \in (-1, 1)$ is fixed but arbitrary and can be chosen dependent on the application. Including this constraint yields an additional upper bound on the amount of fluid that can be used during the optimization process. Hence, the admissible shapes in the optimization problem are described by all design functions inside

$$\Phi_{ad} := \left\{ \varphi \in H^1(\Omega) \mid |\varphi| \leq 1 \text{ a.e. in } \Omega, \int_\Omega \varphi \, \mathrm{d}x \leq \beta \, |\Omega| \right\}. \tag{2}$$

It is a known fact, that shape optimization problems lack in general existence of a minimizer, compare for instance the discussions in [15]. One approach to overcome this problem is the so called perimeter penalization, where a multiple of the perimeter of the fluid region is added to the objective functional. This excludes oscillations and microscopic perforations, see for instance [5], and hence realizes simultaneously certain engineering constraints. As we work in a diffuse interface setting, i.e. the

design variable does not only take discrete values, we do not add a multiple of perimeter functional to the objective functional but merely a multiple of the Ginzburg-Landau energy, namely

$$\gamma \int_\Omega \frac{\epsilon}{2} |\nabla \varphi|^2 + \frac{1}{\epsilon} \psi(\varphi) \, \mathrm{d}x,$$

since this energy is known to be a diffuse interface approximation of a multiple of the perimeter functional, see for instance [16]. Here, $\gamma > 0$ is an arbitrary constant which can be considered as a weighting parameter for the perimeter penalization and $\psi : \mathbb{R} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is a potential with two global minima at ± 1. In this work we choose a double obstacle potential, hence

$$\psi(\varphi) := \begin{cases} \frac{1}{2}\left(1 - \varphi^2\right), & \text{if } |\varphi| \leq 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

This gives rise to a so-called phase field problem where the phase field variable is given by the design function φ and the phase field parameter $\epsilon > 0$ describes the interface thickness. To be precise, the thickness of the interface is proportional to the small parameter $\epsilon > 0$.

Similar to [7], we replace the region outside the fluid by a porous medium with small permeability $(\overline{\alpha}_\epsilon)^{-1} > 0$. Thus we couple the permeability to the phase field parameter $\epsilon > 0$. The velocity $\boldsymbol{u}$ of the fluid in this porous medium is then, due to Darcy's law, described by

$$\overline{\alpha}_\epsilon \boldsymbol{u} - \mu \Delta \boldsymbol{u} + \nabla p = \boldsymbol{f}, \qquad \operatorname{div} \boldsymbol{u} = 0, \tag{3}$$

where p denotes the corresponding pressure. In the interfacial region we interpolate between the equations of flow through porous medium (3) and the Stokes equations (1) by using an interpolation function $\alpha_\epsilon : [-1, 1] \to [0, \overline{\alpha}_\epsilon]$ fulfilling the following assumptions:

(A3) Let $\alpha_\epsilon : [-1, 1] \to [0, \overline{\alpha}_\epsilon]$ be decreasing, surjective and continuous for every $\epsilon > 0$.
It is required that $\overline{\alpha}_\epsilon > 0$ is chosen such that $\lim_{\epsilon \searrow 0} \overline{\alpha}_\epsilon = +\infty$ and α_ϵ converges pointwise to some function $\alpha_0 : [-1, 1] \to [0, +\infty]$. Additionally, we impose $\alpha_\delta(x) \geq \alpha_\epsilon(x)$ if $\delta \leq \epsilon$ for all $x \in [-1, 1]$, $\lim_{\epsilon \searrow 0} \alpha_\epsilon(0) < \infty$ and a growth condition of the form $\overline{\alpha}_\epsilon = \mathrm{o}\left(\epsilon^{-\frac{2}{3}}\right)$.

Remark 1 For space dimension $d = 2$ we can even choose $\overline{\alpha}_\epsilon = \mathrm{o}\left(\epsilon^{-\kappa}\right)$ for any $\kappa \in (0, 1)$, compare also the proof of Theorem 2.

The complete state equations for our problem can be written in its strong form as

$$\alpha_\epsilon(\varphi)\boldsymbol{u} - \mu\Delta\boldsymbol{u} + \nabla p = \boldsymbol{f} \qquad \text{in } \Omega, \tag{4a}$$

$$\operatorname{div}\boldsymbol{u} = 0 \qquad \text{in } \Omega, \tag{4b}$$

$$\boldsymbol{u} = \boldsymbol{g} \qquad \text{on } \partial\Omega. \tag{4c}$$

The weak formulation of this system is given as follows: find $\boldsymbol{u} \in \boldsymbol{U} := \{\boldsymbol{v} \in \boldsymbol{H}^1(\Omega) \mid \operatorname{div}\boldsymbol{v} = 0, \boldsymbol{v}|_{\partial\Omega} = \boldsymbol{g}\}$ such that

$$\int_\Omega \alpha_\epsilon(\varphi)\,\boldsymbol{u}\cdot\boldsymbol{v} + \mu\nabla\boldsymbol{u}\cdot\nabla\boldsymbol{v}\,\mathrm{d}x = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x \quad \forall\boldsymbol{v}\in\boldsymbol{V} \tag{5}$$

with $\boldsymbol{V} := \{\boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega) \mid \operatorname{div}\boldsymbol{v} = 0\}$.

As mentioned above, our goal is to minimize the total potential power

$$\int_\Omega \frac{1}{2}\alpha_\epsilon(\varphi)\,|\boldsymbol{u}|^2 + \frac{\mu}{2}\,|\nabla\boldsymbol{u}|^2 - \boldsymbol{f}\cdot\boldsymbol{u}\,\mathrm{d}x \tag{6}$$

of the fluid. The first term in (6) can also be considered as a penalization term ensuring that $|\boldsymbol{u}|$ is small enough outside the fluid region (i.e. $\varphi = -1$), and vanishes in the limit $\epsilon \searrow 0$. In the sharp interface problem (hence "$\epsilon = 0$") the fact of $\boldsymbol{u}$ vanishing outside the fluid region is essential, compare Sect. 3.

We finally arrive in a porous medium—phase field formulation of the shape optimization problem:

$$\begin{aligned}\min_{(\varphi,\boldsymbol{u})} J_\epsilon(\varphi,\boldsymbol{u}) := &\int_\Omega \frac{1}{2}\alpha_\epsilon(\varphi)\,|\boldsymbol{u}|^2\,\mathrm{d}x + \int_\Omega \frac{\mu}{2}\,|\nabla\boldsymbol{u}|^2 - \boldsymbol{f}\cdot\boldsymbol{u}\,\mathrm{d}x\\ &+ \gamma\int_\Omega \frac{\epsilon}{2}\,|\nabla\varphi|^2 + \frac{1}{\epsilon}\psi(\varphi)\,\mathrm{d}x\\ \text{subject to} &(\varphi,\boldsymbol{u}) \in \Phi_{ad}\times\boldsymbol{U} \text{ and (5)}.\end{aligned} \tag{7}$$

We notice, that for fixed $\varphi \in \Phi_{ad}$ the weak formulated state equations (5) correspond exactly to the necessary and sufficient first order optimality conditions of the convex optimization problem

$$\min_{\boldsymbol{u}\in\boldsymbol{U}} J_\epsilon(\varphi,\boldsymbol{u}).$$

Therefore, the optimization problem (7) is in this case equivalent to

$$\begin{aligned}\min_{(\varphi,\boldsymbol{u})\in\Phi_{ad}\times\boldsymbol{U}} J_\epsilon(\varphi,\boldsymbol{u}) := &\int_\Omega \frac{1}{2}\alpha_\epsilon(\varphi)\,|\boldsymbol{u}|^2\,\mathrm{d}x + \int_\Omega \frac{\mu}{2}\,|\nabla\boldsymbol{u}|^2 - \boldsymbol{f}\cdot\boldsymbol{u}\,\mathrm{d}x\\ &+ \gamma\int_\Omega \frac{\epsilon}{2}\,|\nabla\varphi|^2 + \frac{1}{\epsilon}\psi(\varphi)\,\mathrm{d}x.\end{aligned} \tag{8}$$

In this formulation, no explicit state equations as constraint are necessary any more.

One major advantage of this porous medium—phase field formulation for shape optimization problems in fluids is the existence of a minimizer, as the following theorems shows:

Theorem 1 *For every* $\epsilon > 0$ *there exists a minimizer* $(\varphi_\epsilon, \boldsymbol{u}_\epsilon) \in \Phi_{ad} \times \boldsymbol{U}$ *of the optimization problem* (8).

Proof This can be established quite easily by using the direct method in the calculus of variations. For details we refer to [14]. □

Remark 2 We introduced the porous medium—phase field approach for the problem of minimizing the total potential power in a Stokes flow here. But this approach can also be applied to a larger class of objective functionals and also to different state equations like the stationary Navier-Stokes equations, see [14]. We could also include a term in the objective functional including the pressure of the fluid.

3 Sharp Interface Problem

The optimization problem (8) introduced in the previous section depends on the phase field parameter $\epsilon > 0$, which describes both the interfacial thickness and the permeability of the porous medium outside the fluid region. The natural question arising is what happens if ϵ tends to zero. We expect to arrive in a perimeter penalized sharp interface problem, whose solutions can be considered as so-called black-and-white solutions (see for instance [13]), which means that there exists only pure fluid regions and pure non-fluid regions with zero permeability. And actually, it can be verified in the framework of Γ-convergence that problem (8) has a sharp interface analogue. For a detailed introduction to the notion of Γ-convergence and its applications we refer here for instance to [11].

The resulting problem in the limit will be a shape optimization problem formulated in the setting of Caccioppoli sets. In order to formulate this problem in the right manner we briefly introduce some notation. However, for a detailed introduction into the theory of Caccioppoli sets and functions of bounded variation we refer here to [3, 12]. We call a function $\varphi \in L^1(\Omega)$ a function of bounded variation if its distributional derivative is a vector-valued finite Radon measure. The space of a functions of bounded variation in Ω is denoted by $BV(\Omega)$, and by $BV(\Omega, \{\pm 1\})$ we denote functions in $BV(\Omega)$ having only the values ± 1 a.e. in Ω. We then call a measurable set $E \subset \Omega$ Caccioppoli set, if $\chi_E \in BV(\Omega)$. For any Caccioppoli set E, one can hence define the total variation $|\mathrm{D}\chi_E|(\Omega)$ of $\mathrm{D}\chi_E$, as $\mathrm{D}\chi_E$ is a finite measure. This value is then called the perimeter of E in Ω and is denoted by $P_\Omega(E) := |\mathrm{D}\chi_E|(\Omega)$.

An important point in the formulation of the sharp interface problem is that the velocity $\boldsymbol{u}$ of the fluid is still defined on the whole of Ω, even though we have black-and-white solutions and there are only certain regions inside of Ω filled with fluid.

This is done by defining $\boldsymbol{u}$ to be zero if no fluid is present, which is the case if $\varphi = -1$. And hence the velocity is here an element in $\boldsymbol{U}^{\varphi} := \{\boldsymbol{u} \in \boldsymbol{U} \mid \boldsymbol{u} = \boldsymbol{0}$ a.e. in $\{\varphi = -1\}\}$ if $\varphi \in L^1(\Omega)$. And correspondingly, we introduce the space $\boldsymbol{V}^{\varphi} := \{\boldsymbol{u} \in \boldsymbol{V} \mid \boldsymbol{u} = \boldsymbol{0}$ a.e. in $\{\varphi = -1\}\}$.

The design space for the sharp interface problem is given as

$$\Phi_{ad}^0 := \left\{ \varphi \in BV(\Omega, \{\pm 1\}) \mid \int_\Omega \varphi \, \mathrm{d}x \leq \beta |\Omega|, \ \boldsymbol{U}^{\varphi} \neq \emptyset \right\}.$$

The constraint $\boldsymbol{U}^{\varphi} \neq \emptyset$ is a necessary condition in order to obtain at least one admissible velocity field for this case, since the two conditions of $\boldsymbol{u} = \boldsymbol{0}$ if $\varphi = -1$ and $\boldsymbol{u}|_{\partial\Omega} = \boldsymbol{g}$ may be conflicting.

We extend J_ϵ to the whole space $L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ by defining $J_\epsilon : L^1(\Omega) \times \boldsymbol{H}^1(\Omega) \to \overline{\mathbb{R}}$ as

$$J_\epsilon(\varphi, \boldsymbol{u}) := \begin{cases} \int_\Omega \frac{1}{2}\alpha_\epsilon(\varphi) |\boldsymbol{u}|^2 \, \mathrm{d}x + \int_\Omega \frac{\mu}{2} |\nabla \boldsymbol{u}|^2 - \boldsymbol{f} \cdot \boldsymbol{u} \, \mathrm{d}x + \\ + \gamma \int_\Omega \frac{\epsilon}{2} |\nabla \varphi|^2 + \frac{1}{\epsilon}\psi(\varphi) \, \mathrm{d}x, & \text{if } \varphi \in \Phi_{ad}, \boldsymbol{u} \in \boldsymbol{U}, \\ +\infty, & \text{else.} \end{cases} \tag{9}$$

We will show in Sect. 4 that the Γ-limit of $(J_\epsilon)_{\epsilon>0}$ for $\epsilon \searrow 0$ is given by $J_0 : L^1(\Omega) \times \boldsymbol{H}^1(\Omega) \to \overline{\mathbb{R}}$, where

$$J_0(\varphi, \boldsymbol{u}) := \begin{cases} \int_\Omega \frac{\mu}{2} |\nabla \boldsymbol{u}|^2 - \boldsymbol{f} \cdot \boldsymbol{u} \, \mathrm{d}x + c_0 \gamma P_\Omega(\{\varphi = 1\}), & \text{if } \varphi \in \Phi_{ad}^0, \boldsymbol{u} \in \boldsymbol{U}^{\varphi}, \\ +\infty, & \text{else.} \end{cases}$$

The constant $c_0 := \int_{-1}^{1} \sqrt{2\psi(s)} \, \mathrm{d}s = \frac{\pi}{2}$ arises due to technical reasons, compare Sect. 4.

We find as in the previous section, that the optimization problem

$$\min_{(\varphi, \boldsymbol{u}) \in L^1(\Omega) \times \boldsymbol{H}^1(\Omega)} J_0(\varphi, \boldsymbol{u}) \tag{10}$$

is equivalent to the optimization problem with state constraints given by

$$\min_{(\varphi, \boldsymbol{u})} J_0(\varphi, \boldsymbol{u}) := \int_\Omega \frac{\mu}{2} |\nabla \boldsymbol{u}|^2 - \boldsymbol{f} \cdot \boldsymbol{u} \, \mathrm{d}x + \gamma c_0 P_\Omega(\{\varphi = 1\}) \tag{11}$$

subject to $\varphi \in \Phi_{ad}^0$, $\boldsymbol{u} \in \boldsymbol{U}^{\varphi}$ and

$$-\mu \Delta \boldsymbol{u} + \nabla p = \boldsymbol{f} \qquad \text{in } \{\varphi = 1\}, \tag{12a}$$

$$\operatorname{div} \boldsymbol{u} = 0 \qquad \text{in } \Omega, \tag{12b}$$

$$\boldsymbol{u} = \boldsymbol{g} \qquad \text{on } \partial\Omega. \tag{12c}$$

The strong formulation (12) of the state equations are to be understood in the following weak sense: find $\boldsymbol{u} \in \boldsymbol{U}^{\varphi}$ such that

$$\int_\Omega \mu \nabla \boldsymbol{u} \cdot \nabla \boldsymbol{v} \, \mathrm{d}x = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} \, \mathrm{d}x \quad \forall \boldsymbol{v} \in \boldsymbol{V}^{\varphi}.$$

The shape optimization problem (10) allows in particular every Caccioppoli set as admissible shape, which yields that no geometric properties are prescribed. Additionally, no boundary regularity of the shapes is necessary and even the topology can change during the optimization process. Hence this yields a very large class of possible solutions, in contrast to existing formulations in shape optimization, see for instance [9, 10, 18]. Additionally, we will see in the next section, that there exists a minimizer for J_0, compare Remark 3, which is, as already mentioned above, not a trivial fact in shape optimization problems.

4 Sharp Interface Limit

Let X denote the topological space $L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ equipped with the strong $L^1(\Omega)$ and weak $\boldsymbol{H}^1(\Omega)$ topology. In this section we will show the already announced result that $(J_\epsilon)_{\epsilon>0}$ converges in the sense of Γ-convergence to J_0 as $\epsilon \searrow 0$ in X, hence in $L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ with respect to the topology of X. One important ingredient here is the special structure of the objective functional, hence that no state equations are necessary to be stated explicitly. The proof is based on the result of [16] which ensures that the Ginzburg-Landau energy Γ-converges in $L^1(\Omega)$ to a multiple of the perimeter functional as the phase field parameter ϵ tends to zero. We directly state the main result:

Theorem 2 *The functionals $(J_\epsilon)_{\epsilon>0}$ converge in the sense of Γ-convergence in X to J_0 as $\epsilon \searrow 0$.*

A direct and important consequence of this theorem is given by the following corollary:

Corollary 1 *Let $(\varphi_\epsilon, \boldsymbol{u}_\epsilon)$ be a minimizer of J_ϵ for every $\epsilon > 0$, whose existence is guaranteed by Theorem 1. Then there exists a subsequence, which will be denoted by the same, such that $(\varphi_\epsilon, \boldsymbol{u}_\epsilon)_{\epsilon>0}$ converges (strongly) in $L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ to some limit $(\varphi_0, \boldsymbol{u}_0)$. Additionally, $(\varphi_0, \boldsymbol{u}_0)$ is a minimizer of J_0 and $\lim_{\epsilon \searrow 0} J_\epsilon(\varphi_\epsilon, \boldsymbol{u}_\epsilon) = J_0(\varphi_0, \boldsymbol{u}_0)$.*

Remark 3 Corollary 1 ensures in particular the existence of a minimizer of J_0 and hence also the existence of a minimizer of the constrained optimization problem (11) and (12).

We start by proving the Γ-convergence result of Theorem 2:

Proof of Theorem 2 We use the sequential characterization of the Γ-limit, see [11]. Hence we have to prove two properties in order to deduce the theorem. First we show that for every $(\varphi, \boldsymbol{u}) \in L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$, there exists a sequence $(\varphi_\epsilon, \boldsymbol{u}_\epsilon)_{\epsilon>0} \subset L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ converging to $(\varphi, \boldsymbol{u})$ in X such that

$$\limsup_{\epsilon \searrow 0} J_\epsilon(\varphi_\epsilon, \boldsymbol{u}_\epsilon) \leq J_0(\varphi, \boldsymbol{u}).$$

This sequence is often called recovery sequence. The second step is to show that J_0 provides a lower bound, i.e. we have to show that for every sequence $(\varphi_\epsilon, \boldsymbol{u}_\epsilon)_{\epsilon>0} \subset L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ converging to some element $(\varphi, \boldsymbol{u})$ in X it holds

$$J_0(\varphi, \boldsymbol{u}) \leq \liminf_{\epsilon \searrow 0} J_\epsilon(\varphi_\epsilon, \boldsymbol{u}_\epsilon). \tag{13}$$

For this purpose, we start with a preparatory observation. Let $(\varphi_\epsilon)_{\epsilon>0}$ be any sequence converging pointwise almost everywhere in Ω to some $\varphi \in L^1(\Omega)$. As it holds $\alpha_\delta \leq \alpha_\epsilon$ for all $\epsilon \leq \delta$ we obtain for fixed $\delta > 0$ that

$$\alpha_\delta(\varphi(x)) = \lim_{\epsilon \searrow 0} \alpha_\delta(\varphi_\epsilon(x)) \leq \liminf_{\epsilon \searrow 0} \alpha_\epsilon(\varphi_\epsilon(x))$$

and thus, as $\delta \searrow 0$,

$$\alpha_0(\varphi(x)) = \lim_{\delta \searrow 0} (\alpha_\delta(\varphi(x))) \leq \liminf_{\epsilon \searrow 0} \alpha_\epsilon(\varphi_\epsilon(x))$$

for almost every $x \in \Omega$. On the other hand we have, as $\alpha_\epsilon \leq \alpha_0$, also

$$\limsup_{\epsilon \searrow 0} \alpha_\epsilon(\varphi_\epsilon(x)) \leq \limsup_{\epsilon \searrow 0} \alpha_0(\varphi_\epsilon(x)) = \alpha_0(\varphi(x)).$$

Altogether we thus find

$$\lim_{\epsilon \searrow 0} \alpha_\epsilon(\varphi_\epsilon(x)) = \alpha_0(\varphi(x)) \quad \text{for a.e. } x \in \Omega. \tag{14}$$

We next construct the recovery sequence and choose some $(\varphi, \boldsymbol{u}) \in \Phi^0_{ad} \times \boldsymbol{U}^\varphi$ with $J_0(\varphi, \boldsymbol{u}) < \infty$. To this end, we use the construction of [16], see also [6, 17], which ensures the existence of a sequence $(\varphi_\epsilon)_{\epsilon>0}$ converging strongly in $L^1(\Omega)$ to φ such that

$$\int_\Omega \varphi_\epsilon \, \mathrm{d}x \leq \int_\Omega \varphi \, \mathrm{d}x \leq \beta|\Omega| \quad \forall \epsilon \ll 1$$

and

$$\limsup_{\epsilon \searrow 0} \int_\Omega \frac{\epsilon}{2} |\nabla \varphi_\epsilon|^2 + \frac{1}{\epsilon} \psi(\varphi_\epsilon) \, \mathrm{d}x \leq c_0 P_\Omega(\{\varphi = 1\}).$$

The construction yields additionally the convergence rate

$$\|\varphi_\epsilon - \varphi\|_{L^1(\Omega)} = \mathcal{O}(\epsilon). \tag{15}$$

For details on this construction, in particular on the convergence rate, we refer also to [14]. From $\boldsymbol{u}|_{\{\varphi=-1\}} = \mathbf{0}$ and (14) we find $\lim_{\epsilon \searrow 0} \alpha_\epsilon(\varphi_\epsilon(x))|\boldsymbol{u}|^2(x) = 0$ for almost every $x \in \Omega$. This gives us in view of Lebesgue's dominated convergence theorem and by using the pointwise estimate

$$\alpha_\epsilon(\varphi_\epsilon)|\boldsymbol{u}|^2 \leq \alpha_\epsilon(0)|\boldsymbol{u}|^2 \leq \alpha_0(0)|\boldsymbol{u}|^2 \quad \text{a.e. in } \{\varphi_\epsilon \geq 0\}$$

that

$$\lim_{\epsilon \searrow 0} \int_{\{\varphi_\epsilon \geq 0\}} \alpha_\epsilon(\varphi_\epsilon)|\boldsymbol{u}|^2 \, \mathrm{d}x = 0.$$

We can use the pointwise estimates $|\varphi_\epsilon| \leq 1$, $|\varphi| \leq 1$ and the inclusion $\{\boldsymbol{u} \neq \mathbf{0}\} \subset \{\varphi = 1\}$ to obtain

$$\begin{aligned} \int_{\{\varphi_\epsilon \leq 0\}} \alpha_\epsilon(\varphi_\epsilon)|\boldsymbol{u}|^2 \, \mathrm{d}x &\leq \overline{\alpha}_\epsilon \int_\Omega \chi_{\{\varphi_\epsilon \leq 0, \varphi = 1\}} \underbrace{|\varphi_\epsilon - \varphi|}_{\geq 1} |\boldsymbol{u}|^2 \, \mathrm{d}x \\ &\leq C \overline{\alpha}_\epsilon \|\varphi - \varphi_\epsilon\|_{L^1(\Omega)}^{\frac{2}{3}} \|\boldsymbol{v}\|_{\boldsymbol{L}^6(\Omega)}^2. \end{aligned}$$

Combining the convergence rate (15) and $\overline{\alpha}_\epsilon = \mathrm{o}\left(\epsilon^{-\frac{2}{3}}\right)$, see Assumption (A3), we hence deduce $\lim_{\epsilon \searrow 0} \int_{\{\varphi_\epsilon \leq 0\}} \alpha_\epsilon(\varphi_\epsilon)|\boldsymbol{u}|^2 \, \mathrm{d}x = 0$ and so we end up with

$$\lim_{\epsilon \searrow 0} \int_\Omega \alpha_\epsilon(\varphi_\epsilon)|\boldsymbol{u}|^2 \, \mathrm{d}x = 0.$$

Altogether, this yields

$$\limsup_{\epsilon \searrow 0} J_\epsilon(\varphi_\epsilon, \boldsymbol{u}) \leq J_0(\varphi, \boldsymbol{u})$$

and finishes the first step in this proof.

It remains to show that J_0 is a lower bound on $(J_\epsilon)_{\epsilon>0}$ as described above. For this purpose, we choose an arbitrary sequence $(\varphi_\epsilon, \boldsymbol{u}_\epsilon)_{\epsilon>0} \subset L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ converging to some element $(\varphi, \boldsymbol{u})$ in X. Without loss of generality we assume $\liminf_{\epsilon \searrow 0} J_\epsilon(\varphi_\epsilon, \boldsymbol{u}_\epsilon) < \infty$, otherwise (13) is trivial. We use again the results

of [16] to observe that for an arbitrary sequence $(\varphi_\epsilon, \boldsymbol{u}_\epsilon)_{\epsilon>0} \subset L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ converging to some element $(\varphi, \boldsymbol{u})$ in X it holds

$$c_0 P_\Omega(\{\varphi = 1\}) \leq \liminf_{\epsilon \searrow 0} \int_\Omega \frac{\epsilon}{2} |\nabla \varphi_\epsilon|^2 + \frac{1}{\epsilon} \psi(\varphi_\epsilon) \, \mathrm{d}x.$$

Besides, we obtain with the pointwise considerations (14) and Fatou's lemma

$$\begin{aligned} \int_\Omega \alpha_0(\varphi) |\boldsymbol{u}|^2 \, \mathrm{d}x &= \int_\Omega \liminf_{\epsilon \searrow 0} (\alpha_\epsilon(\varphi_\epsilon)) \left(\liminf_{\epsilon \searrow 0} |\boldsymbol{u}_\epsilon|^2 \right) \mathrm{d}x \\ &\leq \int_\Omega \liminf_{\epsilon \searrow 0} \left(\alpha_\epsilon(\varphi_\epsilon) |\boldsymbol{u}_\epsilon|^2 \right) \mathrm{d}x \leq \liminf_{\epsilon \searrow 0} \int_\Omega \alpha_\epsilon(\varphi_\epsilon) |\boldsymbol{u}_\epsilon|^2 \, \mathrm{d}x. \end{aligned}$$

This yields in particular $\boldsymbol{u} = \boldsymbol{0}$ a.e. in $\{\varphi = -1\}$ and hence $\boldsymbol{u} \in \boldsymbol{U}^\varphi$. Additionally,

$$\boldsymbol{H}^1(\Omega) \ni \boldsymbol{u} \mapsto \int_\Omega \frac{\mu}{2} |\nabla \boldsymbol{u}|^2 - \boldsymbol{f} \cdot \boldsymbol{u} \, \mathrm{d}x$$

is a continuous, convex and thus weakly lower semicontinuous functional. And hence we obtain

$$J_0(\varphi, \boldsymbol{u}) \leq \liminf_{\epsilon \searrow 0} J_\epsilon(\varphi_\epsilon, \boldsymbol{u}_\epsilon)$$

and can hence finish the proof. For some additional technical details and generalizations we refer to [14]. □

Proof of Corollary 1 Similar as in the proof of Theorem 2 we construct for some arbitrary element $(\varphi, \boldsymbol{u}) \in L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ with $J_0(\varphi, \boldsymbol{u}) < \infty$ a sequence $(\widetilde{\varphi}_\epsilon, \widetilde{\boldsymbol{u}}_\epsilon)_{\epsilon>0} \subset L^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ such that $\limsup_{\epsilon \searrow 0} J_\epsilon(\widetilde{\varphi}_\epsilon, \widetilde{\boldsymbol{u}}_\epsilon) \leq J_0(\varphi, \boldsymbol{u})$. Using the minimizing property of $(\varphi_\epsilon, \boldsymbol{u}_\epsilon)$ for each $\epsilon > 0$ this implies that there is some constant $C > 0$ such that

$$J_\epsilon(\varphi_\epsilon, \boldsymbol{u}_\epsilon) \leq J_\epsilon(\widetilde{\varphi}_\epsilon, \widetilde{\boldsymbol{u}}_\epsilon) < C \quad \forall \epsilon \ll 1. \tag{16}$$

Therefrom, we find directly that $\int_\Omega \left(\frac{\epsilon}{2} |\nabla \varphi_\epsilon|^2 + \frac{1}{\epsilon} \psi(\varphi_\epsilon) \right) \mathrm{d}x \leq C$. As in [16] we can hence estimate $\int_\Omega |\nabla \phi(\varphi_\epsilon)| \, \mathrm{d}x$ with $\phi(t) = \int_0^t \sqrt{2\psi(s)} \, \mathrm{d}s$ and with the help of the compactness argument in [16, Proposition 3] we get thus the existence of a subsequence of $(\varphi_\epsilon)_{\epsilon>0}$, which we will denote by the same, converging in $L^1(\Omega)$ to some limit element φ_0. Additionally, we obtain thanks to (16) a subsequence of $(\boldsymbol{u}_\epsilon)_{\epsilon>0}$, which is again denoted by the same, that converges weakly in $\boldsymbol{H}^1(\Omega)$ to some limit element $\boldsymbol{u}_0$. This gives us in view of standard results for Γ-convergence, see [11], and the Γ-convergence result of Theorem 2 that the limit point $(\varphi_0, \boldsymbol{u}_0)$ is a minimizer of J_0 and

$$\lim_{\epsilon \searrow 0} J_\epsilon(\varphi_\epsilon, \boldsymbol{u}_\epsilon) = J_0(\varphi_0, \boldsymbol{u}_0). \tag{17}$$

Finally we combine

$$\begin{aligned} 0 &\leq \liminf_{\epsilon \searrow 0} \int_\Omega \alpha_\epsilon(\varphi_\epsilon) |\boldsymbol{u}_\epsilon|^2 \, \mathrm{d}x, \quad c_0 P_\Omega(\{\varphi_0 = 1\}) \\ &\leq \liminf_{\epsilon \searrow 0} \int_\Omega \frac{\epsilon}{2} |\nabla \varphi_\epsilon|^2 + \frac{1}{\epsilon} \psi(\varphi_\epsilon) \, \mathrm{d}x, \end{aligned}$$

see [16], and

$$\int_\Omega \frac{\mu}{2} |\nabla \boldsymbol{u}_0|^2 - \boldsymbol{f} \cdot \boldsymbol{u}_0 \, \mathrm{d}x \leq \liminf_{\epsilon \searrow 0} \left(\int_\Omega \frac{\mu}{2} |\nabla \boldsymbol{u}_\epsilon|^2 - \boldsymbol{f} \cdot \boldsymbol{u}_\epsilon \, \mathrm{d}x \right)$$

to deduce from (17) that

$$\int_\Omega \frac{\mu}{2} |\nabla \boldsymbol{u}_0|^2 - \boldsymbol{f} \cdot \boldsymbol{u}_0 \, \mathrm{d}x = \lim_{\epsilon \searrow 0} \left(\int_\Omega \frac{\mu}{2} |\nabla \boldsymbol{u}_\epsilon|^2 - \boldsymbol{f} \cdot \boldsymbol{u}_\epsilon \, \mathrm{d}x \right).$$

And hence we can conclude the strong convergence of $(\boldsymbol{u}_\epsilon)_{\epsilon > 0}$ to $\boldsymbol{u}_0$ in $\boldsymbol{H}^1(\Omega)$. For more details we refer to [14]. □

References

1. G. Allaire, F. Jouve, A level-set method for vibration and multiple loads structural optimization. Comput. Methods Appl. Mech. Eng. **194**(30), 3269–3290 (2005)
2. L. Ambrosio, G. Buttazzo, An optimal design problem with perimeter penalization. Calc. Var. Partial Differ. Equ. **1**(1), 55–69 (1993)
3. L. Ambrosio, N. Fusco, D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems* (Clarendon Press, Oxford, 2000)
4. M.P. Bendsøe, *Topology Optimization: Theory, Methods and Applications* (Springer, Berlin, 2003)
5. M.P. Bendsøe, R.B. Haber, C.S. Jog, A new approach to variable-topology shape design using a constraint on perimeter. Struct. Multidiscip. Optim. **11**(1–2), 1–12 (1996)
6. J.F. Blowey, C.M. Elliott, The Cahn-Hilliard gradient theory for phase separation with non-smooth free energy part I: mathematical analysis. Eur. J. Appl. Math. **2**(8), 233–280 (1991)
7. T. Borrvall, J. Petersson, Topology optimization of fluids in Stokes flow. Int. J. Numer. Methods Fluids **41**(1), 77–107 (2003)
8. B. Bourdin, A. Chambolle, Design-dependent loads in topology optimization. ESAIM Control Optim. Calc. Var. **9**(8), 19–48 (2003)
9. C. Brandenburg, F. Lindemann, M. Ulbrich, S. Ulbrich, A continuous adjoint approach to shape optimization for Navier Stokes flow, in *Optimal Control of Coupled Systems of Partial Differential Equations*, ed. by K. Kunisch, J. Sprekels, G. Leugering, F. Tröltzsch. International Series of Numerical Mathematics, vol. 158 (Birkhauser, New York, 2009), pp. 35–56
10. D. Bucur, J.P. Zolésio, N-dimensional shape optimization under capacitary constraint. J. Differ. Equ. **123**(2), 504–522 (1995)
11. G. Dal Maso, *An Introduction to* Γ*-Convergence*. Progress in Nonlinear Differential Equations and Their Applications. (Birkhäuser, 1993)
12. L.C. Evans, R.F. Gariepy, *Measure Theory and Fine Properties of Functions*. Mathematical Chemistry Series (CRC Press INC, Boca Raton, 1992)

13. A. Evgrafov, The limits of porous materials in the topology optimization of Stokes flows. Appl. Math. Optim. **52**(3), 263–277 (2005)
14. C. Hecht, Shape and topology optimization in fluids using a phase field approach and an application in structural optimization. Dissertation, University of Regensburg (2014)
15. B. Kawohl, A. Cellina, A. Ornelas, *Optimal Shape Design: Lectures Given at the Joint C.I.M./C.I.M.E. Summer School Held in Troia (Portugal), 1–6 June 1998*. Lecture Notes in Mathematics/C.I.M.E. Foundation (Subseries. Springer, 2000)
16. L. Modica, The gradient theory of phase transitions and the minimal interface criterion. Arch. Ration. Mech. Anal. **98**(2), 123–142 (1987)
17. P. Sternberg, The effect of a singular perturbation on nonconvex variational problems. Arch. Ration. Mech. Anal. **101**(3), 209–260 (1988)
18. V. Sˇverák, On optimal design. J. Math. Pures Appl. **72**, 537–551 (1993)

An Overview on the Cheeger Problem

Gian Paolo Leonardi

Keywords Cheeger sets · Prescribed mean curvature · Isoperimetric

Mathematics Subject Classification (2010) 49Q10 · 53A10 · 35P15

1 Introduction

Let us fix a bounded open set $\Omega \subset \mathbb{R}^n$, with $n \geq 2$. Given a Borel set $F \subset \mathbb{R}^n$ we denote by $|F|$ its Lebesgue measure (from now on, the *volume* of F) and by $P(F)$ its *perimeter* (see Sect. 3 for the definition of the perimeter functional). Then we define the *Cheeger constant of* Ω as

$$h(\Omega) := \inf \left\{ \frac{P(F)}{|F|} \ : \ F \subset \Omega, \ |F| > 0 \right\}. \tag{1}$$

Any set $E \subset \Omega$ such that $\frac{P(E)}{|E|} = h(\Omega)$ is called a *Cheeger set* of Ω. We shall generically refer to the *Cheeger problem*, as far as the computation or estimation of $h(\Omega)$, or the characterization of Cheeger sets of Ω, are concerned. As we will see later on, the Cheeger problem is deeply connected to other variational problems, ranging from eigenvalue estimates to capillarity models, and even to image segmentation techniques.

The purpose of this note is twofold. First, in order to provide some motivations to the reader, we shall briefly review three relevant problems that show a close

G.P. Leonardi (✉)
Dipartimento di Scienze Fisiche, Informatiche e Matematiche,
Università di Modena e Reggio Emilia, Via Campi 213/b, 41100 Modena, Italy
e-mail: gianpaolo.leonardi@unimore.it

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_6

connection to the Cheeger problem. Second, after some essential definitions and basic results we give an account of known facts about the Cheeger problem, as well as of some more recent results obtained by A. Pratelli and the author in [27]. Some key examples are presented in the final section. We also address the interested reader to [5–7, 10, 11, 20, 21, 33], where further applications, developments and extensions of the Cheeger problem are considered.

2 Some Motivations

In this section we synthetically describe three variational problems that are closely connected with the Cheeger problem.

2.1 Estimating the Smallest Eigenvalue of the Laplacian

The historical motivation of the Cheeger problem is an isoperimetric-type inequality that was first proved by J. Cheeger in [13] in the context of compact, n-dimensional Riemannian manifolds without boundary. As a consequence, one obtains the validity of a Poincaré inequality with optimal constant uniformly bounded from below by a geometric constant. Let $\lambda_2(M)$ be the least non-zero eigenvalue of the Laplace-Beltrami operator on M, then Cheeger proved that

$$\lambda_2(M) \geq \inf_{A \subset\subset M} \frac{P(A)^2}{4 \min\{V(A),\, V(M \setminus A)\}^2}, \tag{2}$$

where $V(A)$ and $P(A)$ denote, respectively, the Riemannian volume and perimeter of A. Here we skip the discussion of the problem on Riemannian manifolds and consider the analogous problem for the p-Laplacian ($1 \leq p < \infty$) with Dirichlet boundary conditions, with M replaced by a bounded open set $\Omega \subset \mathbb{R}^n$. To be more specific, we assume that Ω coincides with its essential interior, i.e., that it contains all points $x \in \mathbb{R}^n$ for which there exists $r > 0$ such that $|B(x, r) \setminus \Omega| = 0$. Under this assumption, all (slightly) different definitions of the Cheeger constant, that have been proposed or considered in previous works, actually agree. We thus exclude from our analysis domains (like, for instance, a planar open disc minus a diameter) which from the point of view of the Lebesgue measure (and of the perimeter) are not distinguishable from their essential interiors. By approximation (see Theorem 3.7) it will then be possible to deduce estimates that are valid for more general domains (and for the more "classical" definition of Cheeger constant, i.e. the minimization of the ratio $\frac{P(F)}{|F|}$ among relatively compact subdomains $F \subset\subset \Omega$).

Let $\lambda_p(\Omega)$ denote the smallest "eigenvalue" of the p-Laplacian with Dirichlet boundary conditions, for $1 \leq p < \infty$:

$$\lambda_p(\Omega) := \inf_{u \in W_0^{1,p}(\Omega)} \frac{\|\nabla u\|_p^p}{\|u\|_p^p}.$$

Arguing as in Cheeger's paper (see [22, 26]) one can easily show that

$$\lambda_p(\Omega) \geq \frac{h(\Omega)^p}{p^p}, \tag{3}$$

where $h(\Omega)$ is defined in (1). The proof of (3) goes as follows: take $u \in W_0^{1,p}(\Omega)$ with a positive Sobolev norm, and set $q = \frac{p}{p-1}$. By noting that $p/q = p - 1$ and thanks to Hölder's inequality, one finds

$$\frac{\int |\nabla u|^p}{\int |u|^p} \geq \frac{\left(\int |u|^{p-1} |\nabla u|\right)^p}{\left(\int |u|^p\right)^p} = \frac{\left(\int |\nabla |u|^p|\right)^p}{p^p \left(\int |u|^p\right)^p}. \tag{4}$$

Setting $f = |u|^p$, by coarea formula (see [4]) one gets

$$\begin{aligned} \int |\nabla f| &= \int_0^{+\infty} \frac{P(\{f > t\})}{|\{f > t\}|} \cdot |\{f > t\}|\, dt \geq h(\Omega) \cdot \int_0^{+\infty} |\{f > t\}|\, dt \\ &= h(\Omega) \cdot \int f, \end{aligned} \tag{5}$$

then by (5) one deduces that

$$\frac{\left(\int |\nabla |u|^p|\right)}{\left(\int |u|^p\right)} = \frac{\int |\nabla f|}{\int f} \geq h(\Omega).$$

Therefore, (3) follows from this last inequality combined with (4).

Remark 2.1 We note that, as $p \to 1$, the left-hand side of (3) tends to $\lambda_1(\Omega)$ while the right-hand side tends to $h(\Omega)$. Moreover, we have

$$\lambda_1(\Omega) = h(\Omega), \tag{6}$$

which means that (3) becomes sharp as $p \to 1$. Proving (6) amounts to show that $\lambda_1(\Omega) \leq h(\Omega)$, as the other inequality directly follows from (3). To this aim, one can exploit (5) on a function $f \in W_0^{1,1}(\Omega)$ that suitably approximates the characteristic function of a set of finite perimeter $F \subset \Omega$, for which $\frac{P(F)}{|F|} \simeq h(\Omega)$. To this aim, it is not restrictive to assume that F is relatively compact in Ω and that ∂F is smooth, hence f can be defined as a standard regularization of χ_F, in such a way that $0 \leq f \leq 1$,

$|F| \simeq \int f$ and $P(F) \simeq \int |\nabla f|$. In conclusion one obtains

$$\int |\nabla f| \simeq \frac{P(F)}{|F|} \int f \simeq h(\Omega) \int f,$$

which implies (6).

2.2 Existence of Graphs with Prescribed Mean Curvature

Let $\Omega \subset \mathbb{R}^n$ be a bounded open domain with Lipschitz boundary. The prescribed mean curvature equation is a nonlinear elliptic partial differential equation of the form

$$\operatorname{div}\left(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\right) = H(x), \tag{7}$$

where H is a given function on Ω. For the moment we do not specify any further property of H and u. It is well-known that the left-hand side of (7) represents the scalar mean curvature of the graph $t = u(x)$, up to a division by $n-1$. The prescribed mean curvature equation arises as the Euler–Lagrange equation of the functional

$$\begin{aligned}\mathcal{J}[u] = &\int_\Omega \sqrt{1+|\nabla u(x)|^2}\, dx + \int_\Omega H(x)u(x)\, dx \\ &+ \int_{\partial\Omega} |u(y) - \varphi(y)|\, d\mathcal{H}^{n-1}(y)\end{aligned} \tag{8}$$

where $\mathcal{H}^{n-1}$ denotes the Hausdorff $(n-1)$-dimensional measure in $\mathbb{R}^n$ and $\varphi \in L^1(\partial\Omega)$ is a boundary datum. The minimization of (8) corresponds to the physical problem of finding the stable equilibrium configurations for a fluid-gas interface in a cylindrical tube of cross-section Ω, subject to surface tension, bulk forces, and boundary conditions. In [19] (see also [18]) some conditions for the existence and uniqueness of solutions to (7) (even without specifying boundary conditions) are found. In particular, we have the following result.

Theorem 2.2 ([19]) *Let Ω be a bounded Lipschitz domain, and let $H \in Lip(\Omega)$. Then, the Eq. (7) admits at least a solution $u \in C^2(\Omega)$ if and only if*

$$\left|\int_A H(x)\, dx\right| < P(A) \tag{9}$$

for all $A \subset \Omega$ with $0 < |A| < |\Omega|$. If, in addition, $\left|\int_\Omega H(x)\right| = P(\Omega)$, then the solution u to (7) is unique up to additive constants and has a "vertical contact" at $\partial\Omega$.

The proof of Theorem 2.2 uses a straightforward application of the divergence theorem for the "only if" part, while becomes more technical in the "if" part. The case $|\int_\Omega H(x)| < P(\Omega)$ is easier and can be handled by showing the existence of smooth minimizers of the functional $\mathcal{J}[u]$ defined in (8). The critical case $|\int_\Omega H(x)| = P(\Omega)$ is more subtle and requires the notion of generalized solution of (7) in the sense of Miranda [28]. We refer to [19] for more details.

In order to better exploit the link between the existence of solutions to (7) and the Cheeger problem, we focus on the case $H(x) = H$ constant. We also assume without loss of generality that $H \geq 0$. Then, Theorem 2.2 implies that a solution $u \in C^2(\Omega)$ to the constant mean curvature equation

$$\operatorname{div}\left(\frac{\nabla u(x)}{\sqrt{1+|\nabla u(x)|^2}}\right) = H \tag{10}$$

exists if and only if $H \leq h(\Omega)$ and no proper subset A of Ω is Cheeger in Ω. In this sense, the Cheeger constant provides a threshold for the prescribed mean curvature, in order that a solution to (10) may exist. A particularly interesting situation occurs in the limit case $H = h(\Omega)$ and when Ω is *uniquely self-Cheeger*, in the sense that Ω is Cheeger in itself and no other proper subset of Ω is Cheeger in Ω. Indeed, in this case one gains not only existence but also *uniqueness* (up to a vertical translation) of the solution to (10). This much more rigid situation corresponds to the case of a graph with constant mean curvature $H = h(\Omega)$, that meets the boundary of the cylinder $\Omega \times \mathbb{R}$ in a tangential way (thus, the gradient $\nabla u(x)$ blows up as x tends to $\partial\Omega$) and whose geometrical shape is, therefore, uniquely determined up to a translation. The physical interest for these optimal shapes becomes immediately apparent: indeed, they represent the equilibrium configurations of the capillary free-surfaces formed by perfectly wetting fluids inside a cylindrical container of cross-section Ω under zero gravity conditions.

2.3 Stable Shapes for Total Variation Minimization

In [31] (see also the analysis performed in [12]) a variational method, now called *ROF model*, was proposed for the regularization of noisy images. Let $g \in L^2(\mathbb{R}^2)$ be a given *image* to be regularized. The idea is to preserve the essential contours and textures of the objects depicted in the image, while removing noise. To this aim, one can solve the following variational problem:

$$\min_{u \in L^2(\mathbb{R}^n) \cap BV(\mathbb{R}^n)} \int_{\mathbb{R}^n} |Du| + \frac{1}{2\lambda} \int_{\mathbb{R}^n} |u-g|^2, \tag{11}$$

where $|Du|$ is the total variation measure associated with the distributional gradient of u, and λ is a positive parameter. We notice that the functional defined in (11)

is strictly convex, and it is not difficult to prove existence (and uniqueness!) of a solution. One could then be tempted to write the following Euler–Lagrange equation associated with (11):

$$\lambda \operatorname{div}\left(\frac{Du}{|Du|}\right) = u - g. \tag{12}$$

However this is far from being correct, since one expects that some "staircasing effects" occurs in the solution, and therefore that its gradient vanishes on regions of positive Lebesgue measure. The correct way of writing the Euler–Lagrange equation can thus be found by means of convex analysis. We recall that the total variation of Du is the convex functional defined by

$$|Du|(\mathbb{R}^n) = \int_{\mathbb{R}^n} |Du| := \sup\left\{ \int u \operatorname{div} \xi : \ \xi \in C_c^1(\mathbb{R}^n; \mathbb{R}^n), \ |g| \le 1 \right\}.$$

We shall also set $J[u] = |Du|(\mathbb{R}^n)$. Being $J[u]$ convex, we can consider its subdifferential at $u \in L^2(\mathbb{R}^n)$:

$$\partial J[u] = \{v \in L^2(\mathbb{R}^n) : \ J[u + w] \ge J[u] + \langle v, w\rangle \text{ for all } w \in L^2(\mathbb{R}^n)\}.$$

Then, the Euler–Lagrange relation derived from the minimality of u with respect to problem (11) is $0 \in \partial J[u] + \frac{u-g}{\lambda}$ or, equivalently,

$$\frac{g-u}{\lambda} \in \partial J[u]. \tag{13}$$

It is possible to show that the subdifferential $\partial J[u]$ consists of the divergences of vector fields that "calibrate" the distributional gradient Du. More precisely, one can rewrite the Euler–Lagrange inclusion (13) in the following, equivalent form: *there exists a vector field* $\xi_u \in L^\infty(\mathbb{R}^n)$ *such that* $|\xi_u| \le 1$, $\operatorname{div} \xi_u \in L^2(\mathbb{R}^n)$, $Du = \xi_u |Du|$ *and*

$$\operatorname{div} \xi_u = \frac{u-g}{\lambda}. \tag{14}$$

Let us assume from now on that $g = \chi_\Omega$ is the characteristic function of some bounded Lipschitz domain Ω. The goal is to characterize the domains Ω for which the solution u of (11) with $g = \chi_\Omega$ is a "scaled copy of g", i.e. of the form $u = \mu\chi_\Omega$, with $\mu \ge 0$. This means that the regularization produced by the ROF model (11) determines, in this case, a change of the contrast, but not of the shape of the initial image $g = \chi_\Omega$.

Following [2], we say that a Lipschitz domain Ω is *calibrable* if $P(\Omega) < \infty$ and if there exists a vector field $\xi \in L^\infty(\mathbb{R}^n; \mathbb{R}^n)$ such that $|\xi| \le 1$, $\xi = \nu_\Omega$ $\mathcal{H}^{n-1}$-almost everywhere on $\partial\Omega$, and

$$-\operatorname{div} \xi = \frac{P(\Omega)}{|\Omega|}\chi_\Omega$$

in the distributional sense. This notion of calibrability is already present in the context of existence and uniqueness problems for graphs with prescribed mean curvature (see [19]). By using (14) one can derive the following result (see [2, 3]).

Theorem 2.3 *The function* $u_\lambda = \left(1 - \frac{P(\Omega)}{|\Omega|}\lambda\right)^+ \chi_\Omega$ *is the unique minimizer of* (11) *with* $g = \chi_\Omega$ *if and only if* Ω *is calibrable.*

Proof First we consider the case $\lambda < \frac{|\Omega|}{P(\Omega)}$, so that

$$u_\lambda = \left(1 - \frac{P(\Omega)}{|\Omega|}\lambda\right)\chi_\Omega.$$

Then we can easily check that u_λ is the unique minimizer of (11) if and only if there exists a vector field $\xi \in K$ satisfying $D\chi_\Omega = \xi|D\chi_\Omega|$ and such that (14) holds for $g = \chi_\Omega$, which means that

$$-\operatorname{div}\xi = \frac{P(\Omega)}{|\Omega|}\chi_\Omega,$$

that is, Ω is calibrable. Concerning the case $\lambda \geq \frac{|\Omega|}{P(\Omega)}$, we observe that (14) is satisfied when $u = 0$, $g = \chi_\Omega$ and

$$\xi_u = \xi_0 = \frac{|\Omega|}{\lambda P(\Omega)}\xi,$$

where ξ denotes a calibrating vector field for Ω. Note that $|\xi_0| \leq 1$ in this case, thus $\operatorname{div}\xi_0 \in \partial J[0]$. □

One can appreciate the close connection between ROF minimization and the Cheeger problem, through this notion of calibrability. To clarify this point, let us first recall the notion of *mean-convexity*. We say that an open set $\Omega \subset \mathbb{R}^n$ with finite perimeter is mean-convex if for any Borel set $F \subset \mathbb{R}^n$ such that $\Omega \subset F$ we have $P(\Omega) \leq P(F)$. In other words, Ω minimizes the perimeter with respect to outer variations. Since the orthogonal projection onto a convex set is a 1-Lipschitz map, by the area formula one can easily infer that (bounded) convex sets are also mean-convex (the converse being not true in general). The following proposition holds.

Proposition 2.4 *Let* Ω *be a Lipschitz domain.*

(i) if Ω *is calibrable, then it is also mean-convex and self-Cheeger;*
(ii) if Ω *is convex and self-Cheeger, then it is calibrable.*

The proof of claim (ii) of Proposition 2.4 can be found in [2]. Here we only describe how to prove (i). Let $A \subset \Omega$ be a relatively compact subdomain with smooth boundary, then by the divergence theorem applied to the calibrating vector field ξ we get

$$\frac{P(\Omega)}{|\Omega|}|A| = -\int_A \operatorname{div}\xi = -\int_{\partial A} \xi \cdot \nu_A \leq P(A),$$

whence $\frac{P(\Omega)}{|\Omega|} \leq \frac{P(A)}{|A|}$. Since we can fix a sequence of relatively compact subdomains Ω_h converging to Ω both in measure and in perimeter, we also conclude that $h(\Omega) = \frac{P(\Omega)}{|\Omega|}$, that is, Ω is self-Cheeger. To prove that Ω minimizes the perimeter with respect to outer variations, we fix a bounded open set F with Lipschitz boundary and strictly containing Ω, then we apply the divergence theorem to the calibrating vector field ξ on $F \setminus \overline{\Omega}$:

$$\begin{aligned} 0 &= \int_{F\setminus\overline{\Omega}} \operatorname{div} \xi = \int_{\partial F} \xi \cdot \nu_F \, d\mathcal{H}^{n-1} - \int_{\partial\Omega} \xi \cdot \nu_F \, d\mathcal{H}^{n-1} \\ &\leq P(F) - P(\Omega), \end{aligned}$$

which implies that Ω is mean-convex.

3 Some General Results on the Cheeger Problem

After recalling some basic facts about sets of finite perimeter, we shall present some general (and mostly known) results on the Cheeger problem for domains in $\mathbb{R}^n$. We shall also recall some more specific results valid for planar domains, that will be needed in Sect. 4.

For a given $x \in \mathbb{R}^n$ and $r > 0$, we set $B_r(x) = \{y \in \mathbb{R}^n : \ |y - x| < r\}$, where $|v|$ is the Euclidean norm of the vector $v \in \mathbb{R}^n$. Given $A \subset \mathbb{R}^n$ we denote by χ_A its characteristic function. With a slight abuse of notation, we write $|A|$ for the Lebesgue (outer) measure of A. We then set $\omega_n = |B_1(0)|$. We define the *perimeter* of a Borel set E as

$$P(E) = \sup \left\{ \int_E \operatorname{div} g : \ g \in C^1_c(\mathbb{R}^n; \mathbb{R}^n), \ |g| \leq 1 \right\}.$$

When $P(E) < +\infty$, we say that E has finite perimeter (in $\mathbb{R}^n$). In this case, $P(E)$ coincides with the total variation of the distributional gradient of the characteristic function of E:

$$P(E) = |D\chi_E|(\mathbb{R}^n),$$

which more generally allows us to define the *relative perimeter*

$$P(E; A) := |D\chi_E|(A)$$

for any Borel set $A \subset \mathbb{R}^n$. By Radon-Nikodym Theorem we can find a Borel $\mathbb{R}^n$-valued function ν_E such that $|\nu_E| = 1$ $|D\chi_E|$-almost everywhere and

$$D\chi_E = -\nu_E |D\chi_E|.$$

One can interpret ν_E as a *generalized exterior normal* to the boundary of E. In order to clarify this concept, we recall the definition of *reduced boundary* $\partial^* E$. We say that $x \in \partial^* E$ if $0 < |E \cap B(x,r)| < \omega_n r^n$ for all $r > 0$ and

$$\exists\, \nu_E(x) := -\lim_{r\to 0^+} \frac{D\chi_E(B_r(x))}{|D\chi_E|(B_r(x))}, \qquad |\nu_E(x)| = 1.$$

Then we quote a classical result by De Giorgi [14].

Theorem 3.1 (De Giorgi) *Let E be a set of finite perimeter, then*

(i) *$\partial^* E$ is countably $\mathcal{H}^{n-1}$-rectifiable in the sense of Federer [17];*
(ii) *for any $x \in \partial^* E$, $\chi_{t(E-x)} \to \chi_{H_{\nu_E(x)}}$ in $L^1_{loc}(\mathbb{R}^n)$ as $t \to +\infty$, where H_ν denotes the half-space through 0 whose exterior normal is ν;*
(iii) *for any Borel set A, $P(E;A) = \mathcal{H}^{n-1}(A \cap \partial^* E)$;*
(iv) *$\int_E \operatorname{div} g = \int_{\partial^* E} g \cdot \nu_E \, d\mathcal{H}^{n-1}$ for any $g \in C^1_c(\mathbb{R}^n;\mathbb{R}^n)$.*

The perimeter functional extends the usual notion of $(n-1)$-dimensional measure of the boundary of a set, in the sense that, for instance, $P(E) = \mathcal{H}^{n-1}(\partial E)$ for any bounded set E with Lipschitz boundary. The advantage of using the perimeter functional instead of the Hausdorff measure in geometric variational problems is mainly due to the lower-semicontinuity and compactness properties stated in the following proposition (see, e.g., [4]).

Proposition 3.2 (Lower-semicontinuity and compactness) *Let $\Omega \subset \mathbb{R}^n$ be an open set and let $(E_j)_j$ be a sequence of Borel sets. We have the following well-known properties:*

(i) *if E is a Borel set, such that $\chi_{E_j} \to \chi_E$ in $L^1_{loc}(\Omega)$, then $P(E;\Omega) \le \liminf_j P(E_j;\Omega)$;*
(ii) *if there exists a constant $C > 0$ such that $P(E_j;\Omega) \le C$ for all j, then there exists a subsequence E_{j_k} and a Borel set E such that $\chi_{E_{j_k}} \to \chi_E$ in $L^1_{loc}(\Omega)$.*

Other useful properties of the perimeter (invariance by isometries and scaling property, isoperimetric inequality, lattice property) are collected in the next proposition.

Proposition 3.3 *Given two Borel sets $E, F \subset \mathbb{R}^n$ of finite perimeter, $\lambda > 0$ and an isometry $T : \mathbb{R}^n \to \mathbb{R}^n$, we have*

$$P(\lambda T(E)) = \lambda^{n-1} P(E)\,, \tag{15}$$

$$P(E) \ge n\omega_n^{\frac{1}{n}} |E|^{\frac{n-1}{n}}\,, \tag{16}$$

$$P(E\cup F) + P(E\cap F) \le P(E) + P(F). \tag{17}$$

We point out that sets of finite perimeter can be extremely weird. For instance, let G be the countable union of open balls of radius 2^{-i} centered at q_i, $i \in \mathbb{N}$, where $(q_i)_{i\in\mathbb{N}}$ is any enumeration of all points with rational coordinates in $\mathbb{R}^2$. By (17) and lower-semicontinuity of the perimeter, G has finite perimeter. However, its

topological boundary has a positive Lebesgue measure (thus in particular its $(n-1)$-dimensional Hausdorff measure is $+\infty$). While generic sets of finite perimeter may thus be very irregular, a regularity theory is available in particular for *minimizers* of the perimeter subject to a volume constraint (see [35]).

Theorem 3.4 (Regularity of perimeter minimizers with volume constraint) *Let Ω be a fixed open domain, and assume E is a Borel set satisfying the following property: $P(E;\Omega) < +\infty$ and for all Borel F such that $E \Delta F \subset\subset \Omega$ and $|F \cap \Omega| = |E \cap \Omega|$, it holds*

$$P(F;\Omega) \leq P(E;\Omega).$$

Then, $\partial^* E \cap \Omega$ is an analytic surface with constant mean curvature, and the singular set $(\partial E \setminus \partial^* E) \cap \Omega$ is a closed set with Hausdorff dimension at most $n-8$.

We now focus on the Cheeger problem, and in doing so we first present some general properties of the Cheeger constant $h(\Omega)$ and of Cheeger sets inside Ω, valid for any dimension $n \geq 2$ (see [22, 23, 34]).

Proposition 3.5 *Let $\Omega, \widetilde{\Omega} \subset \mathbb{R}^n$ be bounded, open sets. Then the following properties hold.*

(i) *If $\Omega \subset \widetilde{\Omega}$ then $h(\Omega) \geq h(\widetilde{\Omega})$.*
(ii) *For any $\lambda > 0$ and any isometry $T : \mathbb{R}^n \to \mathbb{R}^n$, one has $h(\lambda T(\Omega)) = \frac{1}{\lambda} h(\Omega)$.*
(iii) *There exists a (possibly non-unique) Cheeger set $E \subset \Omega$, i.e. such that $\frac{P(E)}{|E|} = h(\Omega)$.*
(iv) *If E is Cheeger in Ω, then E minimizes the relative perimeter among sets with the same volume; consequently, $\partial E \cap \Omega$ has the regularity stated in Theorem 3.4, and in particular $\partial^* E \cap \Omega$ is a hypersurface of constant mean curvature equal to $\frac{h(\Omega)}{n-1}$.*
(v) *If E is Cheeger in Ω then $|E| \geq \omega_n \left(\frac{n}{h(\Omega)}\right)^n$.*
(vi) *If E and F are Cheeger in Ω, then $E \cup F$ and $E \cap F$ (if it is not empty) are also Cheeger in Ω.*
(vii) *If E is Cheeger in Ω and Ω has finite perimeter, then $\partial E \cap \Omega$ can meet $\partial^* \Omega$ only in a tangential way, that is, for any $x \in \partial^* \Omega \cap \partial E$ one has that $x \in \partial^* E$ and $\nu_E(x) = \nu_\Omega(x)$.*

Idea of proof (i) and (ii) are immediate consequences of the definition of Cheeger constant and of (15) coupled with $|\lambda \Omega| = \lambda^n |\Omega|$. The proofs of (iii) and (iv) are accomplished by, respectively, Proposition 3.2 and Theorem 3.4. The proof of (v) follows from the isoperimetric inequality (16) and the fact that $P(E) = h(\Omega)|E|$. To prove (vi) we apply (17) and get

$$\begin{aligned} h(\Omega)(|E \cup F| + |E \cap F|) &= h(\Omega)(|E| + |F|) \\ &= P(E) + P(F) \\ &\geq P(E \cup F) + P(E \cap F) \\ &\geq h(\Omega)(|E \cup F| + |E \cap F|), \end{aligned}$$

hence all previous inequalities are actually equalities and this happens if and only if

$$P(E \cup F) = h(\Omega)|E \cup F| \qquad \text{and} \qquad P(E \cap F) = h(\Omega)|E \cap F|,$$

which proves (vi). While the proofs of (i)–(vi) are essentially known and can be found in the previously cited references, for the proof of (vii) we refer to [27]. □

Remark 3.6 We notice that, by Proposition 3.5 (v) and (vi), we can always find minimal Cheeger sets in Ω (possibly not unique) and a unique maximal Cheeger set (this last can be obtained as the union of all minimal Cheeger sets of Ω). An example of a domain with two disjoint minimal Cheeger sets is shown in Fig. 8.

We consider the problem of continuity of the Cheeger constant $h(\Omega)$ with respect to some suitable notions of convergence of domains. In [27] we prove Theorem 3.7 below (see also [30] for the special case of convex domains). Since the proof is particularly simple, we quote it below with full details.

Theorem 3.7 (Continuity properties of the Cheeger constant, [27]) *Let $\Omega, \Omega_j \subset \mathbb{R}^n$ be nonempty open bounded sets for all $j \in \mathbb{N}$. If $\chi_{\Omega_j} \to \chi_\Omega$ in L^1, then*

$$\liminf_{j\to\infty} h(\Omega_j) \geq h(\Omega). \tag{18}$$

If in addition Ω, Ω_j are sets of finite perimeter and $P(\Omega_j) \to P(\Omega)$ as $j \to \infty$, then

$$\lim_{j\to\infty} h(\Omega_j) = h(\Omega). \tag{19}$$

Proof Let E_j be a Cheeger set in Ω_j (whose existence is guaranteed by Proposition 3.5 (iii)). Without loss of generality we assume that $\liminf_{j\to\infty} P(E_j)$ is finite, then by Proposition 3.2 we deduce that $\chi_{E_j} \to \chi_E$ in L^1 as $j \to \infty$, up to subsequences and for some Borel set E with positive volume. Since $E_j \subset \Omega_j$ and $\chi_{\Omega_j} \to \chi_\Omega$ in L^1 as $j \to \infty$, one immediately infers that $E \subset \Omega$ up to null sets. Then by Proposition 3.2 and by the convergence of $|E_j|$ to $|E|$, one has

$$h(\Omega) \leq \frac{P(E)}{|E|} \leq \liminf_{j\to\infty} \frac{P(E_j)}{|E_j|},$$

which proves (18). If in addition $P(\Omega_j) \to P(\Omega)$ as $j \to \infty$, then we consider E Cheeger in Ω and define $E_j = \Omega_j \cap E$. One can easily check that $E_j \to E$ and $E \cup \Omega_j \to \Omega$ in L^1, as $j \to \infty$. Therefore by (17) we find

$$\begin{aligned}\limsup_{j\to\infty} P(E_j) &\leq P(E) + \limsup_{j\to\infty} P(\Omega_j) - \liminf_{j\to\infty} P(E \cup \Omega_j)\\ &\leq P(E) + P(\Omega) - P(\Omega)\\ &= P(E),\end{aligned}$$

which combined with (18) gives (19). □

3.1 The Cheeger Problem in Convex Domains

Further properties of the Cheeger constant and of Cheeger sets are known when the domain Ω is convex. In particular, we refer to [1] and to the references therein for the proof of the following result.

Theorem 3.8 *Let $\Omega \subset \mathbb{R}^n$ be a convex domain. Then there exists a unique Cheeger set E in Ω. Moreover, E is convex and of class $C^{1,1}$.*

We remark that Theorem 3.8 was proved in [8] under stronger assumptions on Ω. The proof is essentially based on exploiting the link between the Cheeger problem and the capillary problem with zero gravity (i.e., with vertical contact at the boundary, see the discussion in the previous section). In particular, one has that a convex domain Ω is self-Cheeger if and only if it is calibrable, and this happens precisely when Ω is of class $C^{1,1}$ and the mean curvature of $\partial\Omega$ is bounded from above by $\frac{P(\Omega)}{(n-1)|\Omega|}$.

More can be said about Cheeger sets of convex domains of the plane. For the proof of the following result, see [23, 34].

Theorem 3.9 *Let Ω be a bounded convex set in $\mathbb{R}^2$. Then the unique Cheeger set E of Ω is the union of all balls of radius $r = h(\Omega)^{-1}$ that are contained in Ω. Moreover, if we define the inner Cheeger set as*

$$E_r = \{x \in \Omega : \ \mathrm{dist}(x, \partial\Omega) > r\} \tag{20}$$

we have $E = E_r + B(0, r)$ (as a Minkowski sum) and it holds

$$|E_r| = \pi r^2. \tag{21}$$

The proof of Theorem 3.9 is essentially based on *Steiner's formulae* for area and perimeter of tubular neighbourhoods of convex sets in the plane ([32]): if $A \subset \mathbb{R}^2$ is a bounded convex set and $\rho > 0$, then setting $A_\rho = A + B_\rho$ we have

$$|A_\rho| = |A| + \rho P(A) + \pi\rho^2, \tag{22}$$

$$P(A_\rho) = P(A) + 2\pi\rho. \tag{23}$$

We recall that Steiner's formula (22) has been generalized by Weyl to n dimensional domains with boundary of class C^2 (the so-called tube formula, see [36]) and then by Federer [16] under the assumption of *positive reach*, that we introduce hereafter. Given $K \subset \mathbb{R}^n$ compact, we define the *reach* of K as

$$\mathcal{R}(K) = \sup\{\varepsilon \geq 0 : \ \mathrm{dist}(x, K) \leq \varepsilon \ \Rightarrow \ x \text{ has a unique projection onto } K\}.$$

We say that K has positive reach if $\mathcal{R}(K) > 0$. Notice that if K is convex, then $\mathcal{R}(K) = +\infty$. It is convenient to introduce the *outer Minkowski content* of an open bounded set A, defined as

$$\mathcal{M}(A) = \lim_{\rho \to 0} \frac{|A_\rho| - |A|}{\rho},$$

provided that the limit exists. Then we have the following result (see [27]).

Proposition 3.10 *Let $A \subset \mathbb{R}^2$ be a bounded open set with Lipschitz boundary. Let us assume that $\mathcal{R}(\overline{A}) > 0$. Then $P(A) < +\infty$ and Steiner's formulae (22), (23) hold for all $0 < \rho < \mathcal{R}(\overline{A})$.*

Remark 3.11 To see how the inner Cheeger formula (21) can be used to derive information on the Cheeger problem for convex planar domains, we compute the Cheeger constant of a unit square $Q = (0, 1)^2$. First, we observe that the inner Cheeger set of Q is a concentric square of side length $1 - 2r$. Therefore (21) becomes

$$(1 - 2r)^2 = \pi r^2,$$

and by coupling this equation with the condition $1 - 2r > 0$ we infer after some elementary computations that

$$r = \frac{1}{2 + \sqrt{\pi}},$$

whence $h(Q) = 2 + \sqrt{\pi}$. A general algorithm for computing the Cheeger constant of a convex polygon (with some extra property, i.e., that there is a one-to-one correspondence between the connected components of the boundary of the Cheeger set in the interior of the polygon and the vertices of the polygon) can be found in [23].

3.2 Some Further Results About Cheeger Sets in $\mathbb{R}^2$

Let E be a Cheeger set inside an open bounded domain $\Omega \subset \mathbb{R}^2$, and set $r = h(\Omega)^{-1}$ as before. Then a first, general fact is that a connected component of $\partial E \cap \Omega$ is an arc of radius r, that cannot be longer than πr (i.e., it can be at most a half-circle).

Lemma 3.12 ([27]) *Let $\partial E \cap \Omega$ be nonempty and let S be one of its connected components. Then S is an arc of circle of radius r, whose length does not exceed πr.*

An apparently, very intuitive property of a planar Cheeger set E could be the fact that E satisfies an internal ball condition of radius $r = \frac{|E|}{P(E)}$ (i.e., that it is a union of balls of radius r). However, this property is false in general (see Fig. 5 and, in particular, Example 5.2). Anyway, the following result holds true: as soon as a maximal Cheeger set E in Ω contains some ball $B_r(x_0)$, one can roll this ball inside Ω following any sufficiently smooth path of centers, and in doing so the moving ball will remain inside E.but without exiting from E.

Theorem 3.13 (Moving ball, [27]) *Let $r = 1/h(\Omega)$ and let E be a maximal Cheeger set in Ω containing a ball $B_r(x_0)$. Assume that there exists a curve $\gamma : [0, 1] \to \Omega$ of class $C^{1,1}$ and curvature bounded by $h(\Omega)$, such that $x_0 = \gamma(0)$ and $B_r(\gamma(t)) \subset \Omega$ for all $t \in [0, 1]$. Then $B_r(\gamma(t)) \subset E$ for all $t \in [0, 1]$.*

Remark 3.14 The requirement in Theorem 3.13 of maximality of E can be dropped whenever the moving ball remains at a positive distance from $\partial\Omega$. In this case, one can prove by using Lemma 3.12 that the moving ball will never intersect ∂E.

4 Characterization of Cheeger Sets in Planar Strips

In [25], D. Krejčiřík and A. Pratelli consider the Cheeger problem for a class of generically non-convex planar domains, called *strips*. Let $\gamma : [0, L] \to \mathbb{R}^2$ be a curve of class $C^{1,1}$ parametrized by arc-length, such that the modulus of its curvature is bounded by 1. For $t \in [0, L]$ we denote by $\sigma(t)$ the relatively open segment of length 2 whose midpoint is $\gamma(t)$ and such that $\dot\gamma(t)$ is orthogonal to $\sigma(t)$. We also assume that $0 \le t_1 < t_2 \le L$ implies $\sigma(t_1) \cap \sigma(t_2) = \emptyset$ (no-crossing condition). Then, the set

$$\mathcal{S} = Int\left(\bigcup_{t\in[0,L]} \sigma(t)\right)$$

is an *open strip* of width 2 and length L (here $Int(A)$ denotes the set of interior points of A). We call γ the *spinal curve* of the strip $\mathcal{S}$. If the no-crossing condition holds for all $t_1 < t_2 \in (0, L)$, but $\sigma(0) = \sigma(L)$, then we say that $\mathcal{S}$ is a *closed strip* (of course, this requires that the curve γ is closed, too). In particular, an open strip is a $C^{1,1}$-diffeomorphic image of $(0, L) \times (-1, 1)$, while a closed strip is a $C^{1,1}$-diffeomorphic image of $[0, L] \times (-1, 1)$ with identification of points $(0, y)$ and (L, y). More precisely we can take $(t, u) \in [0, L] \times (-1, 1)$ and define the map

$$\Psi(t, u) = \gamma(t) + u\,\nu(t),$$

where $\nu(t)$ denotes the counter-clockwise rotation of the unit vector $\dot\gamma(t)$ by 90 degrees. In the following we shall focus on open strips, as the Cheeger problem for closed ones has been completely treated in [25]. One can check that the map Ψ defined above is a diffeomorphism of class $C^{1,1}$ between the rectangle $(0, L)\times(-1, 1)$ and the (open) strip $\mathcal{S}$. Using the (t, u) coordinate system, i.e. the representation of a generic point x of the strip by means of its coordinates $(t, u) = \Psi^{-1}(x)$, can sometimes be of help (Fig. 1).

Up to a scaling, we can more generally define strips of width $2s$ (in this case we must require that the modulus of the curvature of γ is smaller than $1/s$). Without

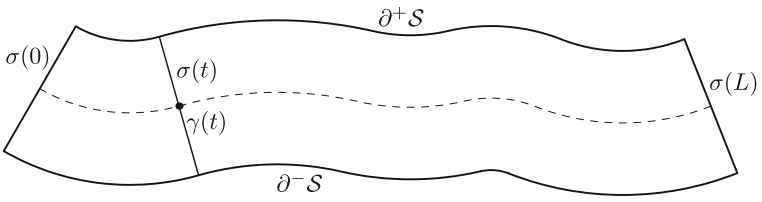


Fig. 1 A planar strip $\mathcal{S}$

loss of generality, we shall consider only strips of width $s = 2$. Moreover, we shall also assume that the curvature of γ is everywhere < 1, as we can recover the case ≤ 1 by approximation, owing to Theorem 3.7. Strips naturally appear in spectral problems, as they model 2-dimensional waveguides (see [15, 24]). In this sense, a number of interesting questions involve the spectral behaviour of a strip when its length L becomes very large. In [25] the authors prove some results specifically on the Cheeger problem for strips. First, they show that closed strips are Cheeger in themselves. Then they prove by means of a suitable symmetrization technique the following bounds on the Cheeger constant of a strip (see Theorem 10 in [25]).

Theorem 4.1 ([25]) *Let $\mathcal{S}$ be a strip of length L and width 2. Then*

$$1 + \frac{1}{400\,L} \leq h(\mathcal{S}) \leq 1 + \frac{2}{L}. \tag{24}$$

In [27] we push forward the analysis done in [25] and, by means of a finer characterization of Cheeger sets inside open strips, we prove the following result.

Theorem 4.2 ([27]) *Let $\mathcal{S}$ be an open strip of length $L \geq \frac{9\pi}{2}$ and width 2. Then*

$$h(\mathcal{S}) = \left(1 + \frac{\pi}{2L} + O(L^{-2})\right) \qquad \text{as } L \to +\infty. \tag{25}$$

The asymptotic estimate (25) is optimal. Its derivation is based on a key result proved in [27]. This result (Theorem 4.3 recalled below) essentially shows that, concerning the Cheeger problem, strips are not too different from convex domains.

Theorem 4.3 ([27]) *Let $\mathcal{S}$ be an open strip of length $L \geq 9\pi/2$, and let $r = h(\mathcal{S})^{-1}$. Assume E is a Cheeger set of $\mathcal{S}$. Then there exists two continuous functions $\rho^+, \rho^- : [0, L] \to [-1, 1]$ such that*

$$E = \Psi\left(\{(t, s) :\ 0 < t < L,\ \rho^-(t) < s < \rho^+(t)\}\right). \tag{26}$$

Moreover, E is unique and coincides with the union of all balls of radius r contained in $\mathcal{S}$, it is simply connected and can be obtained as the Minkowski sum $E = E_r + B_r$, where

$$E_r = \{x \in \mathcal{S} :\ \operatorname{dist}(x, \partial\mathcal{S}) \geq r\}$$

is a set with Lipschitz boundary and positive reach $\mathcal{R}(E_r) \geq r$. *Finally, the inner Cheeger formula*

$$|E_r| = \pi r^2 \tag{27}$$

holds true.

Remark 4.4 We stress that the conclusions of Theorem 4.3 (in particular, the fact that the Cheeger set E is the union of balls of radius r contained in $\mathcal{S}$, and that (27) holds true) are not satisfied by any planar domain. Two examples showing that no inclusion between Cheeger sets and unions of balls of radius r is generally true, are given in the last section (Examples 5.1 and 5.2). Concerning the inner Cheeger formula (27), there exists a star-shaped domain whose Cheeger set is the union of all included balls of radius r, but for which the formula fails (see Example 5.3).

Theorem 4.2 directly follows from Theorem 4.3. Indeed, by the special geometric properties of $\mathcal{S}$ we infer that

$$2(L - 9\pi)(1 - r) \leq |E_r| \leq 2L(1 - r).$$

By combining these two inequalities with the inner Cheeger formula (27), we finally get

$$2(L - 9\pi)(1 - r) \leq \pi r^2 \leq 2L(1 - r),$$

which implies (25) by an elementary computation.

We synthetically present the main ideas and tools, which the proof of Theorem 4.3 is based on. Again, we refer the reader to [27] for the details. We start recalling two key lemmas that are used in the proof of Theorem 4.3. The first lemma states that, if E is a Cheeger set in $\mathcal{S}$, and the length of $\mathcal{S}$ is large enough, then any osculating ball to $\partial E \cap \mathcal{S}$ is entirely contained in $\mathcal{S}$ (see Fig. 2).

Lemma 4.5 (Arc-ball property) *Let E be a Cheeger set inside a strip $\mathcal{S}$ of length $L \geq \frac{9\pi}{2}$. Set $r = h(\mathcal{S})^{-1}$. Then $\partial E \cap \mathcal{S}$ is non-empty, and for any circular arc α contained in $\partial E \cap \mathcal{S}$ the unique ball B_r, such that $\alpha \subset \partial B_r$, is contained in $\mathcal{S}$.*

The second lemma establishes a *ball-to-ball* connectivity property of a generic strip, that is, the possibility of connecting two balls of radius r that are contained in $\mathcal{S}$ by moving one of them towards the other, following a suitable path of centers with controlled curvature and preserving the inclusion in $\mathcal{S}$ (see Fig. 3).

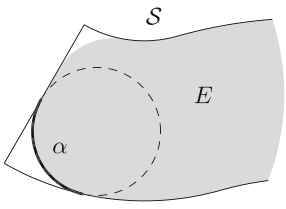


Fig. 2 The arc-ball property

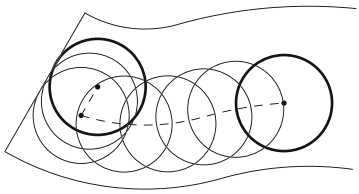


Fig. 3 The ball-to-ball property

Lemma 4.6 (Ball-to-ball property) *If $B_r(x_0)$ and $B_r(x_1)$ are two balls of radius $r \leq 1$, and both are contained in a strip $\mathcal{S}$, then there exists a piece-wise $C^{1,1}$ curve $\beta : [0, 1] \to \mathcal{S}$ such that $\beta(0) = x_0$, $\beta(1) = x_1$, the curvature of β is smaller than r^{-1}, and $B_r(\beta(t)) \subset \mathcal{S}$ for all $t \in (0, 1)$.*

The main difficulties in proving Lemmas 4.5 and 4.6 are of topological type. Roughly speaking, one has to exploit the structural properties of the strip in order to exclude some weird behaviour of its boundary. As it happens for many intuitively clear statements concerning planar objects, proving such lemmas is not as easy as one could imagine at a first sight. For instance, we found no particular simplifications in those proofs by working in the (t, u) coordinate system: this can be understood if one considers that the pre-image of a ball with respect to the map Ψ is no more a ball in the (t, u) coordinates. In several steps of the proofs we find it convenient to argue by contradiction, since a number of (a-posteriori impossible) situations, like for instance the one where an internal ball of radius $\varepsilon < r$ is tangent to more than one point of $\partial^+\mathcal{S}$, or the other where two distinct balls of radius r centered on the spinal curve γ are both tangent to $\sigma(0)$, must be excluded.

With these two lemmas at hand, we can prove Theorem 4.3. Hereafter we provide only a sketch of its proof.

Proof sketch of Theorem 4.3 First of all, we show that there exist exactly four balls of radius r and centers $x_{0,0}, x_{1,0}, x_{0,1}, x_{1,1}$, such that the boundary of $B_r(x_{i,j})$ contains the connected component of $\partial E \cap \mathcal{S}$ that is tangent to

$\sigma(0)$ and $\partial^-\mathcal{S}$ if $i = j = 0$;
$\sigma(0)$ and $\partial^+\mathcal{S}$ if $i = 0$ and $j = 1$;
$\sigma(L)$ and $\partial^-\mathcal{S}$ if $i = 1$ and $j = 0$;
$\sigma(L)$ and $\partial^+\mathcal{S}$ if $i = j = 1$.

This can be accomplished by combining Lemmas 4.5, 4.6, and Theorem 3.13. With this result in force, we are able to define the two functions $\rho^{\pm}$ satisfying (26), which completely characterize the boundary of E. The simply connectedness of E is immediate, as its homeomorphic representation in coordinates (t, u) clearly satisfies this property. On the other hand, we can prove that the union U_r of all balls of radius r that are contained in $\mathcal{S}$ admits in the (t, u) coordinate system the same representation as E:

$$U_r = \Psi\left(\{(t, u) : \ 0 < t < L, \ \rho^-(t) < u < \rho^+(t)\}\right),$$

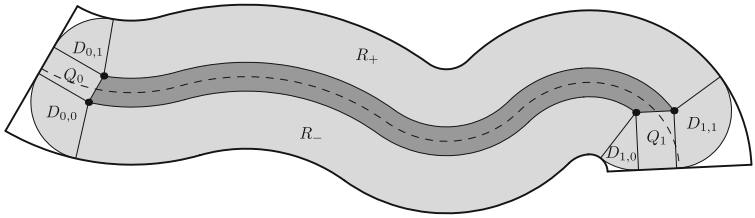


Fig. 4 The decomposition of the Cheeger set. The inner Cheeger set E_r is colored in *dark grey*. One can see the eight regions of the decomposition of $E \setminus E_r$ colored in *light grey*

hence $E = U_r$. Finally, the properties concerning the inner Cheeger set E_r can be proved as follows. For $i, j = 0, 1$ we denote by $a_{i,j}$ the first coordinate of $x_{i,j}$ in the (t, u) representation, then set

$p_{i,j}$ = the orthogonal projection of $x_{i,j}$ onto $\sigma(iL)$;
Q_i = the rectangle of vertices $x_{i,1}, x_{i,0}, p_{i,1}, p_{i,0}$;
$D_{i,j}$ = the circular sector with center $x_{i,j}$ and boundary arc $S_{i,j}$;
R_+ = the region spanned by $\sigma(t; (1 - r, 1))$ as $t \in [a_{0,1}, a_{1,1}]$;
R_- = the region spanned by $\sigma(t; (-1, r - 1))$ as $t \in [a_{0,0}, a_{1,0}]$.

In the above definitions, $\sigma(t; A)$ denotes the set $\{\gamma(t)+u\,\nu(t) : u \in A\}$. Consequently we have the decomposition

$$E \setminus E_r = Q_0 \cup Q_1 \cup \bigcup_{i,j=0}^{1} D_{i,j} \cup R_+ \cup R_-.$$

Finally, we show that E_r has a Lipschitz boundary and that any points of $E \setminus E_r$ has a unique projection onto E_r (which can be more precisely identified according to the above decomposition, see Fig. 4). We can thus apply Steiner's formulae (22) and (23), as in the proof of Theorem 3.9, and obtain the inner Cheeger formula (27), thus concluding the proof of the theorem. □

5 Some Planar Examples

We conclude by collecting some examples of non-convex planar domains, together with their Cheeger sets.

We start from an example of a domain G, whose Cheeger set is strictly contained in the union of balls of radius $r = h(G)^{-1}$ that are contained in G.

Example 5.1 ([23]) Let G be the union of two disjoint balls B_1 and $B_{\frac{2}{3}}$, of radii 1 and $\frac{2}{3}$ respectively (see Fig. 5). One has $\frac{P(G)}{|G|} = \frac{30}{13} > 2$. It is not difficult to check that the Cheeger set E of G coincides with B_1, hence $h(G) = 2$. However, G coincides with

the union of all balls of radius $r = h(G)^{-1} = \frac{1}{2}$ contained in G, which is therefore strictly larger than E.

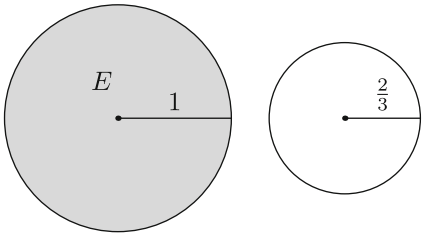


Fig. 5 A union of two disjoint balls B_1 and $B_{\frac{2}{3}}$, whose Cheeger set E coincides with the largest ball B_1

The next example shows a Cheeger set $\mathcal{W}$ *strictly containing* the union of all balls of radius $h(\mathcal{W})^{-1}$ contained in $\mathcal{W}$. This example and the one depicted in Fig. 5 show that, in general, no inclusion holds between a Cheeger set of Ω and the union of all balls of radius $r = h(\Omega)^{-1}$ contained in Ω.

Example 5.2 ([27]) Let us consider a unit-side equilateral triangle T, as in Fig. 6, together with its Cheeger set E_T (depicted in grey). Then, cut T with the vertical line tangent to E and reflect the portion on the left to the right, as shown in the picture. This produces a bow-tie $\mathcal{W}$. Let now $E_{\mathcal{W}}$ be a Cheeger set inside $\mathcal{W}$. By the 2-symmetry of $\mathcal{W}$ one can infer the 2-symmetry of $E_{\mathcal{W}}$. On the other hand, $E_{\mathcal{W}}$ cannot have a connected component F completely contained in T, since otherwise F would be Cheeger inside $\mathcal{W}$ and, at the same time, it would coincide with E_T. But then $E_T \cup E_T'$ (where we have denoted by E_T' the reflected copy of E_T with respect to the cutting line) would be Cheeger in $\mathcal{W}$, which is not possible since $\partial(E_T \cup E_T') \cap \mathcal{W}$ is not everywhere smooth, as it should according to Proposition 3.5. Being necessarily $\partial E_{\mathcal{W}} \cap \mathcal{W}$ equal to a finite union of circular arcs with the same curvature $= h(\mathcal{W})$, it is not difficult to rule out all possibilities except the one in which $\partial E_{\mathcal{W}} \cap \mathcal{W}$ is composed by four congruent arcs, one for each convex corner in the boundary of $\mathcal{W}$. Moreover one has the strict inequality $h(\mathcal{W}) < h(T)$, therefore the union of all balls of radius $h(\mathcal{W})^{-1}$ contained in $\mathcal{W}$ does not contain $E_{\mathcal{W}}$ (indeed, some small region around the two concave corners cannot be covered by those balls).

The next example is obtained as a slight variation of Example 5.2. In this case, the resulting Cheeger set is simply connected, while the inner Cheeger set is disconnected. As a result, we derive the impossibility for the inner Cheeger formula (27) to hold.

Example 5.3 ([27]) Take the bow-tie $\mathcal{W}$ constructed in the previous example and vertically move the two concave corners a bit far apart. By the continuity of the Cheeger constant (see (19)) we infer the existence of some minimal displacement

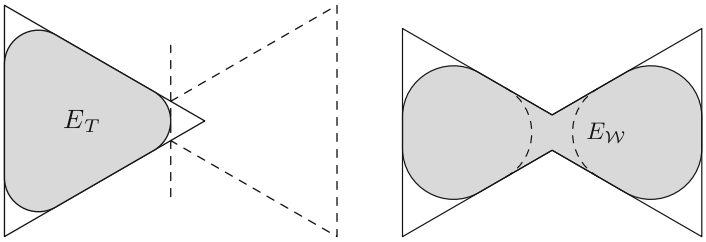


Fig. 6 The construction of the bow-tie $\mathcal{W}$ (*left*) and the Cheeger set $E_{\mathcal{W}}$ in the bow-tie (*right*). Notice that the region between the two *dashed lines* in the picture on the *right* is the difference between the Cheeger set $E_{\mathcal{W}}$ and the (strictly smaller) union of all balls of radius r included in $\mathcal{W}$

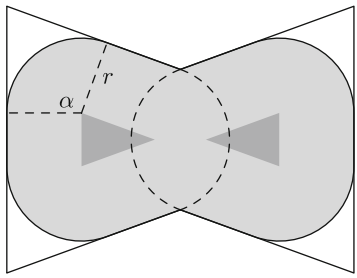


Fig. 7 A loose bow-tie for which the inner Cheeger formula does not hold

of the two corners, such that the Cheeger set in the modified bow-tie $\widetilde{\mathcal{W}}$ actually coincides with the union of all balls of radius $r = h(\widetilde{\mathcal{W}})^{-1}$. This corresponds to the situation represented in Fig. 7. It is then easy to check that the formula $|E_r| = \pi r^2$ does not hold in this case, essentially because the inner Cheeger set E_r (depicted in dark grey) does not satisfy $\mathcal{R}(E_r) \geq r$. We also notice that, while the Cheeger set E is connected, the inner Cheeger set E_r is disconnected. Finally, one can easily check that the true formula, that is satisfied by the inner Cheeger set in this case, is

$$|E_r| = 2\alpha r^2 > \pi r^2,$$

where α is the angle depicted in Fig. 7.

Before getting to the last examples, we recall a result of generic uniqueness for the Cheeger set inside a domain $\Omega \subset \mathbb{R}^n$, proved in [9].

Theorem 5.4 ([9]) *Let $\Omega \subset \mathbb{R}^n$ be any bounded open set, and let $\varepsilon > 0$ be fixed. Then there exists an open set $\Omega_\varepsilon \subset \Omega$, such that $|\Omega \setminus \Omega_\varepsilon| < \varepsilon$ and the Cheeger set of Ω_ε is unique.*

Idea of proof Let E be a minimal Cheeger set of Ω, and let ω_ε be a relatively compact, open subset of Ω with smooth boundary, such that $|\Omega \setminus \omega_\varepsilon| < \varepsilon$. Define $\Omega_\varepsilon = E \cup \omega_\varepsilon$, then by an application of the strong maximum principle for constant mean curvature hypersurfaces one can show that E is the unique Cheeger set of Ω_ε. □

Example 5.5 ([23]) Figure 8 shows a simply connected domain consisting of two congruent squares connected by a small strip. Each square with suitably rounded corners is a minimal Cheeger set, while their union is the maximal Cheeger set of the domain.

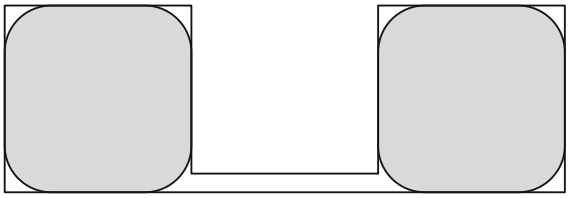


Fig. 8 A simply connected domain whose Cheeger set is not unique

A more sophisticated example of non-uniqueness is constructed below. Indeed one may ask whether it is possible to find a domain admitting infinitely many distinct Cheeger sets. The answer to this question is in the affirmative, as shown by the following example (we point out that a similar example was numerically discussed by E. Parini in [29]).

Example 5.6 ([27]) Let $\mathcal{P}_\theta$ be the union of a unit disc B_1 centered at $(0, 0)$ and a disc of radius $r = \sin\theta$ and center $(\cos\theta, 0)$, where $\theta \in (0, \pi/2)$ will be chosen later. The perimeter of $\mathcal{P}_\theta$ is

$$P(\theta) = 2(\pi - \theta) + \pi \sin\theta,$$

while its area is

$$A(\theta) = (\pi - \theta) + \sin\theta\,\cos\theta + \frac{\pi \sin^2\theta}{2}.$$

Then one shows the existence and uniqueness of $\theta_0 \in (0, \pi/2)$ such that

$$\frac{P(\theta_0)}{A(\theta_0)} = \frac{1}{\sin\theta_0},$$

that is,

$$2(\pi - \theta_0)\sin\theta_0 + \frac{\pi}{2}\sin^2\theta_0 - (\pi - \theta_0) - \frac{\sin(2\theta_0)}{2} = 0. \tag{28}$$

Now we set for brevity $\mathcal{P}_0 = \mathcal{P}_{\theta_0}$ and observe that the ratio $\frac{P(\theta_0)}{A(\theta_0)}$ equals the inverse of the radius of the smaller arc inside $\partial\mathcal{P}_0$. Then by a direct comparison with other possible competitors one infers that $\mathcal{P}_0$ is Cheeger in itself (Fig. 9). Now we consider the one parameter family $\mathcal{P}_t$, $t \in [0, +\infty)$ of sets obtained by "elongating the nose" of $\mathcal{P}_0$ (see Fig. 10). It turns out that $\mathcal{P}_t$ is Cheeger in $\mathcal{P}_\tau$ whenever $t \le \tau$, and this property is stable if one even "bends the nose" of $\mathcal{P}_t$. Indeed, the Cheeger ratio of $\mathcal{P}_t$ is constantly equal to $\frac{1}{\sin\theta_0}$.

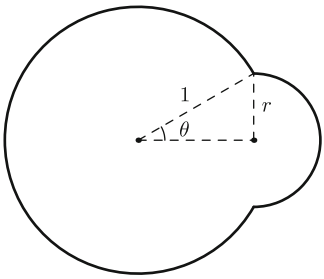


Fig. 9 The set $\mathcal{P}_\theta$

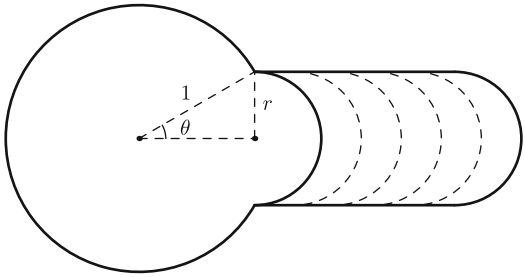


Fig. 10 The one-parameter family of Cheeger sets

Acknowledgments We thank the Department Mathematik—Universität Erlangen-Nürnberg, and in particular Aldo Pratelli, for kind hospitality and financial support through ERC Starting Grant 2010 "AnOptSetCon". We also thank Carlo Nitsch and Antoine Henrot for making us aware of Enea Parini's work.

References

1. F. Alter, V. Caselles, Uniqueness of the Cheeger set of a convex body. Nonlinear Anal. **70**(1), 32–44 (2009)
2. F. Alter, V. Caselles, A. Chambolle, A characterization of convex calibrable sets in $\mathbb{R}^N$. Math. Ann. **332**(2), 329–366 (2005)
3. F. Alter, V. Caselles, A. Chambolle, Evolution of characteristic functions of convex sets in the plane by the minimizing total variation flow. Interfaces Free Bound. **7**(1), 29–53 (2005)
4. L. Ambrosio, N. Fusco, D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, vol. 254 (Clarendon Press, Oxford, 2000)
5. G. Buttazzo, G. Carlier, M. Comte, On the selection of maximal Cheeger sets. Differ. Integral Equ. **20**(9), 991–1004 (2007)
6. G. Carlier, M. Comte, G. Peyré, Approximation of maximal Cheeger sets by projection. M2AN. Math. Model. Numer. Anal. **43**(1), 139–150 (2009)
7. V. Caselles, A. Chambolle, S. Moll, M. Novaga, A characterization of convex calibrable sets in $\mathbb{R}^N$ with respect to anisotropic norms. Ann. Inst. H. Poincaré Anal. Non Linéaire **25**(4), 803–832 (2008)
8. V. Caselles, A. Chambolle, M. Novaga, Uniqueness of the Cheeger set of a convex body. Pac. J. Math. **232**(1), 77–90 (2007)
9. V. Caselles, A. Chambolle, M. Novaga, Some remarks on uniqueness and regularity of Cheeger sets. Rend. Semin. Mat. Univ. Padova **123**, 191–201 (2010)

10. V. Caselles, G. Facciolo, E. Meinhardt, Anisotropic Cheeger sets and applications. SIAM J. Imaging Sci. **2**(4), 1211–1254 (2009)
11. V. Caselles, M.J. Miranda, M. Novaga, Total variation and Cheeger sets in Gauss space. J. Funct. Anal. **259**(6), 1491–1516 (2010)
12. A. Chambolle, P.-L. Lions, Image recovery via total variation minimization and related problems. Numer. Math. **76**(2), 167–188 (1997)
13. J. Cheeger, *A Lower Bound for the Smallest Eigenvalue of the Laplacian*, In Problems in Analysis (Papers Dedicated to Solomon Bochner) (Princeton University Press, New Jersey, 1969). 1970
14. E. De Giorgi, *Selected Papers* (Springer, Berlin, 2006)
15. P. Duclos, P. Exner, Curvature-induced bound states in quantum waveguides in two and three dimensions. Rev. Math. Phys. **7**, 73–102 (1995)
16. H. Federer, Curvature measures. Trans. Amer. Math. Soc. **93**, 418–491 (1959)
17. H. Federer, *Geometric Measure Theory*, Die Grundlehren der mathematischen Wissenschaften, vol. 153 (Springer, New York, 1969)
18. M. Giaquinta, On the Dirichlet problem for surfaces of prescribed mean curvature. Manuscripta Mathematica **12**, 73–86 (1974)
19. E. Giusti, On the equation of surfaces of prescribed mean curvature. Existence and uniqueness without boundary conditions. Invent. Math. **46**(2), 111–137 (1978)
20. D. Grieser, The first eigenvalue of the Laplacian, isoperimetric constants, and the max flow min cut theorem. Arch. Math. **87**(1), 75–85 (2006)
21. I.R. Ionescu, T. Lachand-Robert, Generalized cheeger sets related to landslides. Calc. Var. Partial Differ. Equ. **23**(2), 227–249 (2005)
22. B. Kawohl, V. Fridman, Isoperimetric estimates for the first eigenvalue of the p-Laplace operator and the Cheeger constant. Comment. Math. Univ. Carolin. **44**(4), 659–667 (2003)
23. B. Kawohl, T. Lachand-Robert, Characterization of Cheeger sets for convex subsets of the plane. Pac. J. Math. **225**(1), 103–118 (2006)
24. D. Krejčiřík, J. Kříz, On the spectrum of curved planar waveguides. Publ. Res. Inst. Math. Sci. **41**(3), 757–791 (2005)
25. D. Krejčiřík, A. Pratelli, The Cheeger constant of curved strips. Pac. J. Math. **254**(2), 309–333 (2011)
26. L. Lefton, D. Wei, Numerical approximation of the first eigenpair of the p-Laplacian using finite elements and the penalty method. Numer. Funct. Anal. Optim. **18**(3–4), 389–399 (1997)
27. G.P. Leonardi, A. Pratelli, Cheeger sets in non-convex domains. (2014) arXiv:1409.1376
28. M. Miranda, Superfici cartesiane generalizzate ed insiemi di perimetro localmente finito sui prodotti cartesiani. Ann. Scuola Norm. Sup. Pisa **3**(18), 515–542 (1964)
29. E. Parini, *Cheeger Sets in the Nonconvex Case* (Università degli Studi di Milano, Tesi di Laurea Magistrale, 2006)
30. E. Parini, Asymptotic behaviour of higher eigenfunctions of the p-Laplacian as p goes to 1. Ph.D. thesis, Universität zu Köln (2009)
31. L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. Phys. D: Nonlinear Phenom. **60**(1), 259–268 (1992)
32. J. Steiner, Über parallele flächen. Monatsber. Preuss. Akad. Wiss, 114–118, (1840)
33. G. Strang, Maximum flows and minimum cuts in the plane. J. Global Optim. **47**(3), 527–535 (2010)
34. E. Stredulinsky, W.P. Ziemer, Area minimizing sets subject to a volume constraint in a convex set. J. Geom. Anal. **7**(4), 653–677 (1997)
35. I. Tamanini, Regularity results for almost minimal oriented hypersurfaces in $\mathbb{R}^N$. Quaderni del Dipartimento di Matematica dell'Università di Lecce (1984), http://cvgmt.sns.it/paper/1807/
36. H. Weyl, On the volume of tubes. Am. J. Math. **61**(2), 461–472 (1939)

Shape- and Topology Optimization for Passive Control of Crack Propagation

Günter Leugering, Jan Sokołowski and Antoni Zochowski

Abstract In this review article the theoretical foundations for shape-topological sensitivity analysis of elastic energy functional in bodies with nonlinear cracks and inclusions are presented. The results obtained can be used to determine the location and the shape of inclusions which influence in a desirable way the energy release at the crack tip. In contrast to the linear theory, where in principle, crack lips may mutually penetrate, here we employ nonlinear elliptic boundary value problems in non-smooth domains with cracks with non-penetration contact conditions across the crack lips or faces. A shape-topological sensitivity analysis of the associated variational inequalities is performed for the elastic energy functional. Topological derivatives of integral shape functionals for variational inequalities with unilateral boundary conditions are derived. The closed form results are obtained for the Laplacian and linear elasticity in two and three spatial dimensions. Singular geometrical perturbations in the form of cavities or inclusions are considered. In the variational context the singular perturbations are replaced by regular perturbations of bilinear forms. The obtained expressions for topological derivatives are useful in numerical method of shape optimization for contact problems as well as in passive control of crack propagation.

G. Leugerin (✉)
Institute of Applied Mathematics 2, Friedrich-Alexander University Erlangen-Nürnberg,
Cauerstrasse 11, 91058 Erlangen, Germany
e-mail: leugering@math.fau.de

J. Sokołowski
Institut Élie Cartan, UMR 7502 Nancy-Université-CNRS-INRIA,
Laboratoire de Mathématiques, Université Henri Poincaré Nancy 1, BP 239,
54506 Vandoeuvre Lès Nancy Cedex, France
e-mail: Jan.Sokolowski@univ-lorraine.fr

A. Zochowski
Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6,
01-447 Warszawa, Poland
e-mail: Antoni.Zochowski@ibspan.waw.pl

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_7

Keywords Frictionless contact · Elastic bodies with cracks · Signorini conditions on cracks · Variational inequality · Shape functional · Shape sensitivity · Topological sensitivity · Domain decomposition · Steklov-Poincaré operator · Contact problems

1 Introduction and Overview

Understanding of nucleation, growth and propagation of single cracks and crack patterns in the context of composite materials is a grand challenge in material sciences. This is even more true for the controlled interaction between shapes and geometries of material inclusions on the one side and defects leading to damage, cracks and finally failure on the other side. Within the last century there have been developed a number of theories describing the propagation of cracks in solids. From the point of view of mathematical rigour, the approaches by Griffith and Barenblat are by now well established and widely accepted. Given a particular distribution of inclusions in a matrix material of an elastic body, it is possible, using Griffith's theory, to evaluate the stress concentrations at the tip of an incipient crack. This provides a structure-property map from the shapes and geometries - and of course also the material properties - of the inclusions to the dissipated energy or the energy release rate. This mapping can be described in terms of the Griffith functional. Optimal design of composites with respect to influencing crack-properties is then a matter of 'inverting' that map, in the sense of inverse engineering. Mathematically, this inversion is at the heart of inverse problems and, more precisely in this context, of sensitivity based shape and topology optimization. To this end, directional derivatives of the Griffith functional play a major role in the context of control of crack propagation in brittle materials where the Griffith criterion applies. The idea of designing composites with this aim is not new, some attempts have been made in the literature. See [6] who initialized this field of research in studying a distributed control problem for the Laplacian with a linear crack, that is a crack where no non-penetration condition is assumed to hold. The goal of that paper was to stop the crack propagation under the action of the control. In [16] the problem of crack control has been treated with non-penetration condition along the crack and boundary controls. The authors of [37] consider the shape of inclusions with different material properties as controls but take a linear crack model for a problem in conductivity. See also [51] for examples in mechanical engineering, where sensitivities are typically based on FEM-models. All articles mentioned are concerned with the reduction of the energy release rate. There are only very few articles concerned with shape variations of rigid or elastic inclusions in order to influence the energy release rate associated with non-penetrating cracks. This leads to a problem of shape-optimization in the context of variational inequalities; see [18] for an approach involving obstacles. The maximization of the energy release rate, rather its reduction, is important in some cases, where one wants to release as much energy as possible such that the material does not undergo a global crack. A first attempt towards optimization of the shape of inclusion with respect to maximizing the energy release rate have been reported in [21, 29, 30, 34, 48–50].

However, a rigorous mathematical treatment on the infinite dimensional level is still in its infancy. This article aims at a self-contained description of sensitivity based crack-control in the particular sense that the design of composites is geared towards influencing the crack resistance and, finally, the crack propagation. The sensitivities used in order to optimize the crack propagation are topological and shape derivatives of the Griffith functional with respect to changes in the inclusions constituting the composite.

Topological derivatives of shape functionals are introduced in [54] for linear elliptic boundary value problems. The corresponding expressions depend on pointwise values of solutions as well as of its gradients [46]. Therefore, the expressions for topological derivatives are not well defined on the energy spaces associated with the boundary value problems. In this paper we propose equivalent expressions for the topological derivatives for variational inequalities which are derived by a domain decomposition technique. Such expressions are given by line integrals in two spatial dimensions, or by surface integrals in three spatial dimensions. In addition, the new expressions are well defined on the energy space. In order to derive the topological derivatives by an application of the domain decomposition technique an artificial interface $\Sigma \subset \Omega$ is introduced and $\Omega := \Omega_1 \cup \Sigma \cup \Omega_2$ is decomposed into two subdomains. The functional under consideration is the elastic energy $\mathcal{E}(\Omega)$ of the whole domain Ω. Mixed shape-topological or topological-shape second order derivatives of the energy are evaluated. While shape sensitivity analysis is performed in Ω_2, asymptotic analysis is performed in Ω_1. In the framework of shape-topological sensitivity analysis the velocity method is applied in order to determine the shape functional $J(\Omega) := d\mathcal{E}(\Omega; V)$, where V is the specific vector field in derivation of $V \to d\mathcal{E}(\Omega; V)$. Then an asymptotic expansion of $\epsilon \to J(\Omega_\epsilon)$ is obtained. In the framework of topological-shape sensitivity analysis, first the asymptotic expansion of $\epsilon \to \mathcal{E}(\Omega_\epsilon)$ is performed, and the first order term of such an expansion is called the topological derivative. It turns out [46, 54] that the topological derivative of the energy functional is unbounded in the energy space of the elasticity boundary value problems under considerations. Therefore, we study an equivalent representation of topological derivatives which are well defined in the energy space. These representations can be used as well to modify the state equations by replacing the singular domain perturbations by the regular perturbations of bilinear forms in variational setting.

The asymptotic analysis of the energy functional performed in one subdomain, e.g., Ω_1, can be used in the second subdomain Ω_2 by means of an asymptotic expansion of the Steklov-Poincaré operator on the interface. The method is justified by the fact that the first order expansion of the energy functional in the subdomain leads to the first order asymptotic expansion of the Dirichlet-to-Neumann mapping on the interface between subdomains. Thus, a first order expansion of the Steklov-Poincaré operator on the interface for the second subdomain is obtained. In this way, the first order expansion of the energy functional in the truncated domain Ω_2 is derived. The precision of the obtained expansion is sufficient [56, 58] to replace the original energy functional by its first order expansion, provided the obtained expression is well defined on the energy space. Furthermore, the first order approximation of the

energy functional in Ω is established. We point out that another method of approximation of the state equation by using the so-called self-adjoint extensions of the elliptic operators can be considered [39, 40].

The proposed domain decomposition method is important for variational inequalities [2] related to crack problems with non-penetration conditions across the crack. The arguments can, however, be developed for general variational inequalities. In order to describe the methodology in a nut-shell, before going on to details for elasticity, we consider the following abstract set-up.

$$v \to I(v) = \frac{1}{2}a(v, v) - L(v) \tag{1}$$

over a convex, closed subset $K \subset H$ of the Hilbert space H called the energy space. The function space $H := H(\Omega)$ is a Sobolev space which contains the functions defined over a domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. The singular geometrical perturbation ω_ϵ centered at $\widehat{x} \in \Omega$ of the domain Ω is denoted by Ω_ϵ, the size of the perturbation is governed by a small parameter $\epsilon \to 0$. The quadratic functional defined on $H := H(\Omega_\epsilon)$ becomes

$$v \to I_\epsilon(v) = \frac{1}{2}a_\epsilon(v, v) - L_\epsilon(v) \tag{2}$$

with the minimizers $u_\epsilon \in K := K(\Omega_\epsilon)$. The expansion of the associated energy functional

$$\epsilon \to \mathcal{E}(\Omega_\epsilon) := I_\epsilon(u_\epsilon) = \frac{1}{2}a_\epsilon(u_\epsilon, u_\epsilon) - L_\epsilon(u_\epsilon) \tag{3}$$

is considered at $\epsilon = 0$. Namely, we are looking for its asymptotic expansion

$$\mathcal{E}(\Omega_\epsilon) = \mathcal{E}(\Omega) + \epsilon^d \mathcal{T}(\widehat{x}) + o(\epsilon^d), \tag{4}$$

where $\widehat{x} \to \mathcal{T}(\widehat{x})$ is the topological derivative [46, 54]. We show that there are regular perturbations of the bilinear form defined on the energy space $H(\Omega)$,

$$v \to b(v, v)$$

such that the perturbed quadratic functional defined on the unperturbed function space $H(\Omega)$

$$v \to I^\epsilon(v) = \frac{1}{2}\left[a(v, v) + \epsilon^d b(v, v)\right] - L(v) \tag{5}$$

furnishes the first order expansion (4). In our applications to contact problems in linear elasticity, it turns out that the bilinear form $v \to b(v, v)$ is supported on $\Gamma_R := \{|x - \widehat{x}| = R\} \subset \Omega$ with $R > \epsilon > 0$.

Variational inequalities are used to model contact problems in elasticity. It is known that the solutions to variational inequalities are Lipschitz continuous with respect to the shape [52]. In general, however, the state governed by a variational inequality is not Fréchet differentiable with respect to the shape. For a class of variational inequalities described by unilateral constraints in Sobolev spaces of Dirichlet type, the metric projection onto the constraints turns out to be Hadamard differentiable [12]. This property is used in order to obtain the first order directional differentiability of the associated shape functionals.

In order to show second order shape differentiability for variational inequalities, we have to restrict ourselves to energy-type shape functionals. The energy functional is the so-called marginal function and it is Fréchet differentiable with respect to the shape [12]. The first order shape derivative of the energy functional in the direction of a specific velocity vector field is considered as the shape functional for topological optimization. Thus, its topological derivative is evaluated. The possible applications of shape-topological derivatives include the control of singularities of solutions to variational inequalities by insertion of elastic inclusions far from the singularities.

Example 1 We describe the shape-topological differentiability of the energy shape functional for the Signorini variational inequality in two spatial dimensions. The same idea can be used for the frictionless contact problems in linear elasticity.

Let us consider the Signorini problem posed in $\Omega \subset \mathbb{R}^2$, with boundary $\partial\Omega = \Gamma \cup \Gamma_0$, and $\Gamma_c \subset \Gamma$. Denote $H^1_{\Gamma_0}(\Omega) = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_0 \subset \partial\Omega\}$. The solution $u \in K$ minimizes the quadratic functional

$$I(v) = \frac{1}{2}a(\Omega; v, v) - (f, v)_\Omega$$

over the cone

$$K = \{v \in H^1_{\Gamma_0}(\Omega) \mid v \geqslant 0 \text{ on } \Gamma_c \subset \Gamma \subset \partial\Omega\}.$$

The shape functional is the energy

$$\mathcal{E}(\Omega) = \frac{1}{2}a(\Omega; u, u) - (f, u)_\Omega,$$

where

$$a(\Omega; u, u) = \int_\Omega \nabla u \cdot \nabla u dx,$$
$$(f, u)_\Omega = \int_\Omega f u dx.$$

We assume that $\overline{\Gamma} \cap \overline{\Gamma}_0 = \emptyset$. Let $\Gamma_0^t := T_t(V)(\Gamma_0)$ be the boundary variations [52] of the Dirichlet boundary Γ_0.

Let us consider the decomposition of $\Omega = \Omega_1 \cup \Sigma \cup \Omega_2$, $\overline{\Omega}_1 \cap \overline{\Omega}_2 = \Sigma$, such that $\Gamma_0 \subset \partial\Omega_1$ and $\Gamma_c \subset \partial\Omega_2$. It means that the boundary variations as well as the topological asymptotic analysis are performed in Ω_1, and the unilateral conditions are prescribed in the second subdomain Ω_2.

The shape derivative of the energy functional with respect to the boundary variations of Γ_0 can be written in distributed form [52]

$$d\mathcal{E}(\Omega; V) = \int_{\Omega_1} \langle A'(0) \cdot \nabla u, \nabla u \rangle dx$$

where $A'(0) = \operatorname{div} V I - DV - DV^\top$, under the assumption that the velocity field V is supported in a small neighborhood of Γ_0 and that supp $V \cap$ supp f $= \emptyset$.

The second shape functional for the purposes of topological optimization is simply defined by

$$J(\Omega) := \int_{\Omega_1} \langle A'(0) \cdot \nabla u, \nabla u \rangle dx. \tag{6}$$

We are going to determine the topological derivatives of $\Omega \to J(\Omega)$ for insertion of small inclusions in Ω_1 far from Γ_0. In this way we can control the possible singularities on Γ_0 by topology optimization in Ω.

We consider the domain decomposition method for purposes of the shape-topological differentiability of energy shape functionals. First, the domain Ω is split into two subdomains Ω_1, Ω_2 and the interface Σ. See Fig. 1. The differentiability with respect to small parameter of the Dirichlet-to-Neumann map which lives on the boundary $\Sigma \subset \partial\Omega_1$ is established. This map is called the Steklov-Poincaré operator for subdomain Ω_2.

Once, the derivative of the energy functional is given, we can proceed with the subsequent topological optimization problem. For topological optimization another decomposition $\Omega := \Omega_R \cup \Gamma_R \cup \Omega_c$ is introduced. The small inclusion ω_ε centered at the origin $\widehat{x} := \mathcal{O}$ is located in subdomain $\Omega_R \subset \Omega$ with the interface $\Gamma_R \subset \partial\Omega_R$. See Fig. 2.

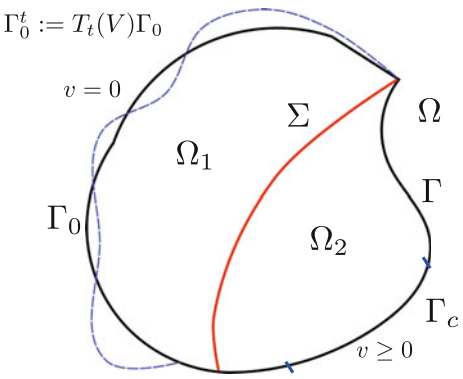
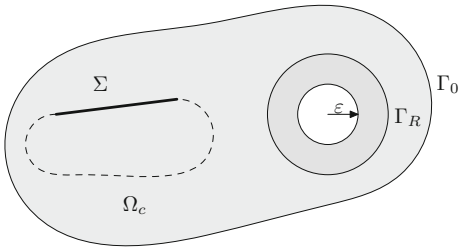


Fig. 1 Signorini problem with domain decomposition

Fig. 2 Domain Ω with crack



In particular, an elastic body weakened by small cracks is considered in the framework of unilateral variational problems in linearized elasticity. The frictionless contact conditions are prescribed on the crack lips in two spatial dimensions, or on the crack faces in three spatial dimensions. The weak solutions of the equilibrium boundary value problem for the elasticity problem are determined by minimization of the energy functional over the cone of admissible displacements. The associated elastic energy functional evaluated for the weak solutions is considered for the purpose of control of crack propagation. The singularities of the elastic displacement field at the crack front are characterized by the shape derivatives of the elastic energy with respect to the crack shape in the framework of Griffith's theory. For example, in two spatial dimensions, the first order shape derivative of the elastic energy functional evaluated in the direction of a velocity field supported in an open neighbourhood of one of crack tips is called the *Griffith functional*. The Griffith functional is minimized with respect to the shape and the location of small inclusions in the body. The inclusions are located far from the crack. In order to minimize the Griffith functional over an admissible family of inclusions, the second order directional, *mixed shape-topological derivatives* of the elastic energy functional are evaluated to determine the locations of inclusions. The boundary value problem for the elastic displacement field takes the form of a variational inequality over the positive cone in a fractional Sobolev space. The sensitivity analysis of variational inequalities under considerations lead to the property of directional differentiability of metric projection operator onto a polyhedric positive cone in fractional Sobolev spaces. Therefore, the concept of *conical or Hadamard differentiability* applies to shape and topological sensitivity analysis of variational inequalities under consideration.

In our framework of shape-topological differentiability we consider:

- Variational inequalities for cracks in solids, the associated Griffith functional is given by the shape derivative of the elastic energy;
- Conical differentiability of metric projection onto positive cone in the fractional Sobolev space of Dirichlet type equipped with natural order;
- Asymptotic analysis of the Dirichlet-to-Neumann map with applications to domain decomposition technique and Steklov-Poincaré nonlocal pseudodifferential boundary operators;
- The second order shape-topological derivatives of elastic energy for the purposes of passive control of crack propagation.

In linearized elasticity the Griffith criterion for crack propagation in two spatial dimensions uses the size of singularity coefficients at the crack tips, called the *stress intensity factors*, in order to forecast the crack propagation. In the pioneering paper [26] this criterion is extended to the nonlinear crack models with a mathematical proof which uses the *Griffith shape functional*, i.e. the shape derivative of the elastic energy with respect to the perturbations of positions of crack tips. The next step in the analysis of nonlinear crack models is the control of crack propagation. For such control the possible strategy is proposed in this paper with full proofs.

- Find the sensitivities of the Griffith functional with respect to the location of inclusions in elastic body;
- These sensitivities are called the topological derivatives [45] of Griffith's shape functional and can be determined by the asymptotic analysis in the singularly perturbed geometrical domains;
- Minimize the topological derivatives and in this way determine the possible locations of inclusions;
- Use the shape sensitivity analysis of Griffith's shape functional and determine optimal shape of inclusions.

The main difficulty of sensitivity analysis of solutions to nonlinear boundary value problems in non-smooth domains under consideration are the unilateral conditions prescribed on the crack lips which lead to the variational inequalities of the first kind. The asymptotic analysis of variational inequalities [57, 58] with respect to small parameter which governs the size of the singular domain perturbation is performed by a domain decomposition technique. In the present paper the mathematical foundation of the passive control strategy for crack propagation by means of shape-topological optimization is described in detail. First, the method of sensitivity analysis used in this paper is explained. The variational inequality in the perturbed domain $\Omega_\epsilon \subset \Omega$ is replaced by another variational inequality in the intact domain Ω. To this end the bilinear form $a(\Omega_\epsilon; \cdot, \cdot)$ is approximated by the bilinear form

$$a(\Omega; \cdot, \cdot) + \epsilon^d b(\Gamma_R; \cdot, \cdot), \tag{7}$$

where $d = 2, 3$ is the space dimension.

We apply the method of boundary variations [53] and the asymptotic analysis [45] in the subdomain Ω_R in order to obtain the expansions of the elastic energy with respect to an inclusion. These expansions are used in the subdomain Ω_c which contains the crack. As a surprising result expansion (7) of the bilinear form, which is well defined on the energy space in the intact domain Ω, is established. To our best knowledge, the bilinear form $b(\Gamma_R; \cdot, \cdot)$ has been employed in asymptotic analysis in singularly perturbed geometrical domains for the first time in [57, 58] for the Laplacian and the planar elasticity.

We now briefly describe the contents of paper, referring to the corresponding sections.

In Sect. 2, frictionless contact problems for the crack are introduced. The elastic energy of the elastic body is considered in the subdomain Ω_c. The contribution of the elastic energy from the subdomain Ω_R is given by the energy of the boundary Steklov-Poincaré operator. The Steklov-Poincaré operator depends on a small parameter $\epsilon \to 0$ which governs the size of singular geometrical perturbation in Ω_R.

In Sect. 3, general results on directional differentiability of metric projection are adapted to crack problems. The conical differentiability of solutions to the variational inequality in Ω_c leads to the main result of the paper, which is the directional differentiability of the Griffith functional with respect to the shape parameter. The abstract results on conical differentiability of the metric projection [12, 53] are adapted to the non-penetration conditions prescribed on the crack.

In Sect. 4, the representative case of cracks in two spatial dimensions are considered for shape-topological sensitivity.

In Sect. 5, the complete proof of shape and topological differentiability of the elastic energy in Ω_R is given. This implies the differentiability of the boundary bilinear form associated with the Steklov-Poincaré operator on Γ_R. Thus, the Griffith functional is differentiable. In this way we show that the main result of the paper applies to the crack control strategy.

In Sect. 6, the bounded perturbations of bilinear forms are presented for elliptic boundary value problems. In such a way the second order shape-topological derivatives of the energy functionals can be evaluated by easily implemented numerical methods.

The expansion of the Steklov-Poincaré operator involves a correction term $\mathbf{B}$, an operator that is made explicit for ring-shaped regions in Sect. 7 for a number of situations.

Finally, in Sect. 8, an asymptotic analysis of the Steklov-Poincaré operator is considered for ring-type walled inclusions, where different material properties apply. This can be seen as an approach for coating of particles included in a matrix material.

In this article, some mathematical aspects of modeling and optimization for nonlinear partial differential equations are required, we refer the reader to the references which can be considered for the specific topics:

- potential theory in Dirichlet spaces and applications to unilateral problems [1, 5, 12, 14, 36, 53]
- mathematical theory of variational inequalities with applications to mechanics and contact problems [7, 8, 14, 15, 22, 23, 36, 53, 57, 58]
- shape optimization in domains with cracks and for variational inequalities [9–12, 18, 22, 53, 57, 58]
- asymptotic analysis and topological optimization for elasticity and variational inequalities [2, 35, 41–45, 57, 58]
- modeling and control of cracks [6, 16, 17, 19, 21, 24–28, 30–34, 37, 48–51]
- numerical methods for variational inequalities and crack problems [3, 4]
- optimization for nonlinear pde's [22, 47]

2 Unilateral Boundary Conditions in Isotropic Elasticity

We consider the following situation:
For the sake of simplicity, it is assumed that

- the crack in two spatial dimensions is given by the interval $\Gamma_c := \{0 < x_1 < 1, x_2 = 0\}$;
- the crack in three spatial dimensions is given by the disk $\Gamma_c := \{0 \leqslant x_1^2 + x_2^2 < 1, x_3 = 0\}$.

Therefore, the function spaces for the crack problem can be identified in Lipschitz domains (see Fig. 3)

- the traces $u^{\pm}$ on Σ of functions $u^{\pm} \in H^1(\Omega^{\pm})$ live in the space $H^{1/2}(\Sigma)$;
- the traces $u^{\pm}$ on Γ_c of functions $u \in H^1(\Omega_c)$ are defined as the restrictions to Γ_c of functions from $H^{1/2}(\Sigma)$;
- the space of traces on the crack $H_{00}^{1/2}(\Gamma_c) \subset H^{1/2}(\Sigma)$ extended by zero outside the crack;
- the jump $[\![u]\!] := u^+ - u^-$ of a function $u \in H^1(\Omega_c)$ is well defined in $H_{00}^{1/2}(\Gamma_c)$;
- the convex constraints for the crack with nonpenetration condition are given by the positive cone in the space $H_{00}^{1/2}(\Gamma_c)$.

2.1 Isotropic Elasticity Boundary Value Problems

For a given displacement vector field $v = (v_1, v_2, v_3)^\top : \Omega \to \mathbb{R}^3$, we define the Jacobian $Dv = (\partial_{x_j} v_i)$ and the gradient is its transpose

$$\nabla v = Dv^\top = (\partial_{x_i} v_j) = [\nabla v_1, \nabla v_2, \nabla v_3]$$

The symmetrized gradient is denoted by

$$\nabla v^s = (\nabla v + \nabla v^\top)/2,$$

and it is called the linearized deformation tensor $\varepsilon(v) := \nabla v^s$.

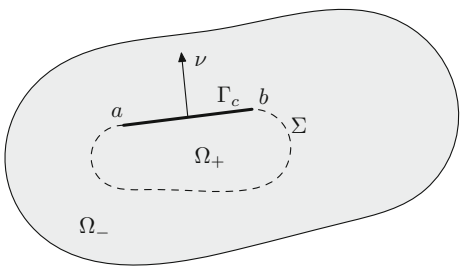


Fig. 3 Domain decomposition of elastic body Ω weakened by crack Γ_c

Given the symmetric and positive definite constitutive tensor $\mathbb{C}$ with the components c_{ijkl}, $i, j, k, l = 1, 2, 3$ and the inverse $\mathbb{S} := \mathbb{C}^{-1}$, the symmetric stress tensor is defined by

$$\sigma(v) = \mathbb{C}\nabla v^s, \quad \text{hence,} \quad \varepsilon(v) = \mathbb{S}\sigma(v)$$

or for the components $\sigma_{ij} = c_{ijrs}\varepsilon_{rs}$, where the summation convention over the repeated indices is used. In the case of the isotropic elasticity

$$c_{ijrs} = \lambda\delta_{ij}\delta_{rs} + \mu(\delta_{ir}\delta_{js} + \delta_{is}\delta_{jr}),$$

whence,

$$\sigma_{ij} = \lambda\delta_{ij}\varepsilon_{kk} + 2\mu\varepsilon_{ij},$$

where λ and μ are the Lame constants, μ is also known as the shear modulus.

Let us assume that the elastic body is given by a torus and let us consider the decomposition of the elastic body into two subdomains $\Omega := \Omega^+ \cup \Sigma \cup \Omega^-$ where Σ is a $C^{1,1}$ regular closed surface. Let $\Gamma_c \subset \Sigma$ be the regular subset of the surface with $C^{1,1}$-boundary given by the curve $\partial\Gamma_c$. We denote by ν the unit normal vector field on Σ which points out of Ω^+, and by n the unit normal vector field on $\partial\Gamma_c$ orthogonal to ν. See Fig. 3.

Given the displacement field $v \in H^1(\Omega)$, and $\sigma := \sigma(v)$, the associated stress field, we introduce the normal and tangential components of the stress field on Σ

$$\sigma_\nu = \sigma_{ij}\nu_j\nu_i, \quad \sigma_\tau = \sigma\nu - \sigma_\nu\nu, \quad \sigma_\tau = (\sigma_{\tau 1}, \sigma_{\tau 2}, \sigma_{\tau 3})^\top.$$

First, we recall the strong form of a general crack boundary value problem in two spatial dimensions.

Remark 2 For the sake of simplicity it is assumed that on the exterior boundary $\Gamma := \partial\Omega$ of the elastic body the homogeneous Dirichlet boundary conditions are prescribed. For a torus boundary conditions disappear. In the case of domain decomposition, the exterior boundary of the subdomain Ω_c is divided into two parts, $\Gamma_R = \partial\Omega_R$ and the exterior boundary $\partial\Omega$.

Problem 3 (*Equilibrium problem for a linear elastic body occupying* Ω_c) In the domain Ω_c with the boundary $\partial\Omega_c := \Gamma \cup \Gamma_c$ we have to find a displacement field $u = (u_1, u_2)$ and stress tensor components $\sigma = \{\sigma_{ij}\}$, $i, j = 1, 2$, such that

$$-\mathrm{div}\sigma = f \quad \text{in} \quad \Omega_c, \tag{8}$$

$$\sigma = \mathbb{C}\varepsilon(u) \quad \text{in} \quad \Omega_c, \tag{9}$$

$$u = 0 \quad \text{on} \quad \Gamma, \tag{10}$$

$$[\![u]\!]\nu \geqslant 0, \quad [\![\sigma_\nu]\!] = 0, \quad \sigma_\nu \cdot [\![u]\!]\nu = 0 \quad \text{on} \quad \Gamma_c, \tag{11}$$

$$\sigma_\nu \leqslant 0, \quad \sigma_\tau = 0 \quad \text{on} \quad \Gamma_c^{\pm}. \tag{12}$$

Here $[\![v]\!] = v^+ - v^-$ is a jump of v on Γ_c, and signs $\pm$ correspond to the positive and negative crack faces with respect to ν, $f = (f_1, f_2) \in L^2(\Omega_c) := L^2(\Omega_c; \mathbb{R}^2)$ is a given function,

$$\sigma_\nu = \sigma_{ij}\nu_j\nu_i, \quad \sigma_\tau = \sigma\nu - \sigma_\nu \cdot \nu, \quad \sigma_\tau = (\sigma_\tau^1, \sigma_\tau^2),$$
$$\sigma\nu = (\sigma_{1j}\nu_j, \sigma_{2j}\nu_j),$$

the strain tensor components are denoted by $\varepsilon_{ij}(u)$,

$$\varepsilon_{ij}(u) = \frac{1}{2}(u_{i,j} + u_{j,i}), \quad \varepsilon(u) = \{\varepsilon_{ij}(u)\}, \quad i, j = 1, 2.$$

The elasticity tensor $\mathbb{C} = \{c_{ijkl}\}$, $i, j, k, l = 1, 2$, is given and satisfies the usual properties of symmetry and positive definiteness

$$c_{ijkl}\xi_{kl}\xi_{ij} \geqslant \alpha|\xi|^2, \quad \forall\, \xi_{ij},\ \xi_{ij} = \xi_{ji}, \quad \alpha > 0,$$

$c_{ijkl} = c_{klij} = c_{jikl}$, $c_{ijkl} \in L^\infty(\Omega_c)$.

Relations (8) are equilibrium equations, and (9) is the generalized Hooke's law, $u_{i,j} = \frac{\partial u_i}{\partial u_j}$, $(x_1, x_2) \in \Omega_c$. All functions with two lower indices are symmetric in those indices, i.e. $\sigma_{ij} = \sigma_{ji}$ etc.

In three spatial dimensions the strong form of the crack boundary value problem is completely analogous. The weak form is given by a variational inequality.

Problem 4 Introduce the Sobolev space

$$H^1_\Gamma(\Omega_c) = \{v = (v_1, v_2) \mid v_i \in H^1(\Omega_c),\ v_i = 0 \text{ on } \Gamma,\ i = 1, 2\}$$

and the closed convex set of admissible displacements

$$K = \{v \in H^1_\Gamma(\Omega_c) \mid [\![v]\!]\nu \geqslant 0 \quad \text{a.e. on} \quad \Gamma_c\}.$$

Find a solution $u \in K$ of the energy minimization problem

$$\min_{v \in K} \left\{ \frac{1}{2} \int_{\Omega_c} \sigma_{ij}(v)\varepsilon_{ij}(v) - \int_{\Omega_c} f_i v_i. \right\}$$

The solution satisfies the variational inequality

$$u \in K, \quad \int_{\Omega_c} \sigma_{ij}(u)\varepsilon_{ij}(v - u) \geqslant \int_{\Omega_c} f_i(v_i - u_i), \quad \forall v \in K, \tag{13}$$

where $\sigma_{ij}(u) = \sigma_{ij}$ are defined in (9).

Remark 5 1. Existence and uniqueness of solutions for the strong Problem 3 and the variational inequality (13) in Problem 4 as optimality conditions for the minimization Problem in 4 are given in e.g. [27].
2. The analysis of shape-topological differentiability of the Griffith functional can be reduced by the proposed domain decomposition approach to the differentiability property of the solution mapping for variational inequality (13)

$$f \to u$$

with respect to the input f. This will be investigated in Sect. 3. We claim that the mapping admits a conical differential. The proof of this claim follows by the Hadamard differentiability of metric projection onto positive cone in fractional Sobolev spaces.

2.2 Control of the Crack Front

The Griffith shape functional is an appropriate indicator in the framework of linear elasticity for the crack propagation scenario. In order to influence the crack propagation, we are going to design the elastic body in such a way that the Griffith functional assumes better properties. In order to improve the design, we consider a finite number of inclusions in the matrix material. Optimization in this context means the best choice of location and shape of inclusions, which can be complemented by optimization of material parameters for the inclusion. To this end, we employ the shape-topological sensitivity analysis [45, 53, 57]. Our analysis is performed for a single inclusion, the same approach works for a finite number of inclusions.

2.2.1 Elastic Body with a Crack and an Inclusion

The domain is divided in two parts as described above. The first part Ω_c which contains the crack, is built up from the matrix material (λ, μ), the second is Ω_R with an inclusion ω. The material properties of ω are denoted by $(\lambda_\omega, \mu_\omega)$. For simplicity, we can consider the inclusion in the form of a ball

$$\omega := B(y, r) = \{x \in \Omega_R \ : \ |x - y| < r\}, \quad \partial\omega = \{|x - y| = r\},$$

however a general shape of the inclusion can be treated in the same way. A finite number of inclusions, far from the crack, is also admissible for our approach.

2.2.2 The Griffith Shape Functional

For a given vector field $V := (V_1, V_2)^\top$ supported in Ω_c, denote $2E_{ij}(V; u) := u_{i,k}V_{k,j} + u_{j,k}V_{k,i}$, where $V_{k,j} := \dfrac{\partial V_k}{\partial x_j}$, $k, j = 1, 2$, and define the shape functional depending on ω,

$$J(\omega) := \frac{1}{2}\int_\Omega \left\{\mathrm{div}V \cdot \varepsilon_{ij}(u) - 2E_{ij}(V; u)\right\}\sigma_{ij}(u) - \int_\Omega \mathrm{div}(Vf_i)u_i.$$

Problem 6 The problem is then to minimize $J(\omega)$ with respect to $\omega \subset \Omega_R$ and solutions u satisfying in the domain $\Omega := \Omega_c \cup \Gamma_R \cup \Omega_R$ the variational inequality

$$u \in K(\omega), \quad \int_\Omega \sigma_{ij}(u)\varepsilon_{ij}(v-u) \geqslant \int_\Omega f_i(v_i - u_i) \quad \forall v \in K(\omega),$$

where

$$K(\omega) = \{v \in H^1_\Gamma(\Omega) \mid [\![v]\!]\nu \geqslant 0 \quad \text{a.e. on} \quad \Gamma_c\}.$$

2.3 Main Results

The shape optimization problem under considerations depends on the *shape of the inclusions* exclusively via the characteristic functions of the inclusions. We are interested in the existence of the shape derivatives of $J(\omega)$ and also of the topological derivatives of this functional. In such a case we speak of the shape-topological differentiability of the Griffith functional.

Theorem 7 *The shape functional $\omega \to J(\omega)$ is directionally shape-topologically differentiable with respect to the inclusion ω in the cracked elastic body Ω.*

We precise the general result for the specific class of circular inclusions. First of all, the simplest choice of the admissible family $\mathcal{U}_{\mathrm{ad}}$ of inclusions with the material properties $(\lambda_\omega, \mu_\omega)$ is

$$\mathcal{U}_{\mathrm{ad}} := \{B(y, r) \subset \Omega_R\}.$$

Such a family, parametrized in a compact subset of $\mathbb{R}^{3+d}$ by $(\lambda_\omega, \mu_\omega, y, r) \in \mathbb{R}^{3+d}$, $d = 2, 3$, is compact with respect to the convergence of characteristic functions. Thus, the existence of an optimal inclusion within this family follows by standard arguments.

Theorem 8 *For given parameters $(\lambda_\omega, \mu_\omega)$ and $\omega = B(y, r) \subset \Omega_R$, the function*

$$r \to I(r) := \frac{1}{2}\int_\Omega \left\{\mathrm{div}V \cdot \varepsilon_{ij}(u) - 2E_{ij}(V; u)\right\}\sigma_{ij}(u) - \int_\Omega \mathrm{div}(Vf_i)u_i$$

is Lipschitz continuous and admits the directional derivatives given by

- *the shape derivative of* $\omega \to J(\omega)$ *for* $r > 0$, r *small enough;*
- *the topological derivative of* $\omega \to J(\omega)$ *for* $r = 0^+$.

3 Applications of Directional Differentiability of Metric Projection in Fractional Sobolev Spaces

Our results on shape-topological sensitivities for the Griffith functional related to crack problems with non-penetration conditions across the crack interfaces depend crucially on the regularity properties of metric projections in Hilbert spaces. This is a classical issue that, due to its importance in this article, nevertheless, deserves a brief but self-contained description. The convex cone for the crack model with non-penetration conditions takes the form

$$\mathbb{K} := \{v \in H^1(\Omega_c) \ : \ [\![v]\!]\nu \in \mathcal{K}(\Gamma_c) \subset H^{1/2}_{00}(\Gamma_c)\},$$

where $\mathcal{K}(\Gamma_c)$ is the positive cone in the fractional Sobolev space $H^{1/2}_{00}(\Gamma_c)$. Therefore, we establish the *Hadamard differentiability* [15, 36] of the *metric projection* in the *Dirichlet space* $H^{1/2}_{00}(\Gamma_c)$ onto its *positive cone* [12]. Let us consider the directional differentiability of the metric projection onto the positive cone in the fractional Sobolev spaces $H^{1/2}(\Sigma)$. In the applications for the crack problem, we would like to have a $C^{1,1}$-surface in three spatial dimensions, and the $C^{1,1}$-curve in two spatial dimensions, selected in the interior of the elastic body Ω in such a way that the crack $\Gamma_c \subset \Sigma$. In order to present the results, we are going to consider a simple geometry of the crack Γ_c. In the general setting the results are obtained in a similar way. Therefore, we consider the subset $B = \{|x| < R\}$, $x = (x_1, \ldots, x_d) \subset \Omega$, of the elastic body Ω, with the crack $\Gamma_c := \{x = (x', x_d) \in \mathbb{R}^d \ : \ x_d = 0, \quad |x'| < R/2\}$ and Σ defined by an extension of the subset $\widetilde{\Sigma} := \{x = (x', x_d) \in B \ : \ x_d = 0\}$. In such a case, the unit normal vector to the crack $\nu := (0, \ldots, 0, 1)$ is constant on the crack, and the unit tangent vector orthogonal to ν on the boundary $\partial\Gamma_c$ of the crack $n := (n_1, \ldots, n_{d-1}, 0)$. For the displacement field $u = (u_1, \ldots, u_d)$ it follows that $u\nu = u_d$, hence, the unilateral constraints for the jump of the normal component over the crack $H^{1/2}_{00}(\Gamma_c) \ni [\![u]\!]\nu = u_d^- - u_d^+ \geqslant 0$. Thus, the convex cone of admissible displacements for the crack problem takes the form

$$\mathcal{U}_{\text{ad}} = \{v = (v_1, \ldots, v_d) \in H^1(\Omega_c) \ : \ v_d^- - v_d^+ \geqslant 0 \text{ on } \Gamma_c\}$$

and our analysis of the metric projection is reduced to the positive cone in $H^{1/2}_{00}(\Gamma_c)$, hence, in $H^{1/2}(\Sigma)$.

Remark 9 We recall that in general for a domain Ω with the boundary Γ, the Sobolev spaces $H^1(\Omega)$ and $H^{1/2}(\Gamma)$ are [1, 14] examples of so-called Dirichlet spaces.

It means that for the scalar product $a(\cdot,\cdot)$, with $v^+ := \sup\{v, 0\}$ and $v^- := \sup\{-v, 0\}$, the property $a(v^+, v^-) \leq 0$ holds for all elements of the Sobolev spaces.

Remark 10 The metric projection in Dirichlet spaces onto the cone of nonnegative elements is considered for the purpose of sensitivity analysis of solutions to frictionless contact problems in [53]. This result is extended to the crack problem. In order to avoid unnecessary technicalities, we restrict ourselves to a model problem. We consider the Hadamard differentiability of metric projection in Dirichlet space onto the cone of positive elements, and recall the result on its conical differentiability.

Consider the convex, closed cone

$$K = \{v \in H^{1/2}(\Sigma) \ : \ v \geqslant 0 \text{ on } \Sigma\}$$

and the metric projection $H^{1/2}(\Sigma) \ni f \to u = P_K(f) \in K$ onto K which is defined by the variational inequality

$$u \in K \ : \ (u - f, v - u)_{1/2,\Sigma} \geqslant 0 \quad \forall v \in K.$$

We denote $v^+ =: v \wedge 0 := \sup\{v, 0\}$ and $v^- =: -v \wedge 0 := \sup\{-v, 0\}$ in $H^{1/2}(\Sigma)$. With the element $u = P_K(f)$ we associate the convex cone

$$C_K(u) = \{v \in H^{1/2}(\Sigma) \ : \ u + tv \in K \text{ for some } t > 0\}$$

and denote by $T_K(u)$ the closure of $C_K(u)$ in $H^{1/2}(\Sigma)$. On the other hand [12] there is a nonnegative Radon measure m such that for all $v \in H^{1/2}(\Sigma)$ we have the equality $\int v\, dm = (u - f, v)_{1/2,\Sigma}$, hence, we denote

$$m[v] := (u - f, v)_{1/2,\Sigma}.$$

Definition 11 The convex cone K is polyhedric [15, 36] at $u \in K$ if

$$T_K(u) \cap m^\perp = \overline{C_K(u) \cap m^\perp}.$$

We recall the result on polyhedricity of the positive cone in a Dirichlet space [12].

Lemma 12 *The convex cone*

$$C_K(u) \cap m^\perp := \{v \in H^{1/2}(\Sigma) \ : \ v \in C_K(u) \text{ such that } (u - f, v)_{1/2,\Sigma} = 0\}$$

is dense in the closed, convex cone

$$T_K(u) \cap m^\perp := \{v \in H^{1/2}(\Sigma) \ : \ v \in T_K(u) \text{such that } (u - f, v)_{1/2,\Sigma} = 0\}.$$

Proof Using the property of the Dirichlet space

$$(v^+, v^-)_{1/2,\Sigma} \leqslant 0 \quad \text{for all } v \in H^{1/2}(\Sigma)$$

then

$$T_K(u) \cap m^\perp = \overline{C_K(u) \cap m^\perp}$$

follows easily. Indeed, let

$$w \in T_K(u) \cap m^\perp.$$

Then $w = 0$ m-a.e. Let $C_K(u) \ni v_n \to w$. Then $v_n^- \to w^-$, $v_n^+ \to w^+$ and $v_n^+ \wedge w^+ - v_n^- \to w$, here $v \wedge w = \inf\{v, w\}$. Now, if $v \in C_K(u)$ then $u + tv \geq 0$. We claim $v_n^+ \wedge w^+ - v_n^- \in C_K(u) \cap m^\perp$. Indeed, $u + t[v_n^+ \wedge w^+ - v_n^-] \geq 0$ so $v_n^+ \wedge w^+ - v_n^- \in C_K(u)$ and $m[v_n^+ \wedge w^+ - v_n^-] = m[v_n^+ \wedge w^+] = 0$, because of $m[w^+] = 0$. □

Remark 13 In [12] the tangent cone $T_K(u)$ is derived for $u \in K$, in the case of the positive cone $K = \{v \in \mathcal{H} \ : \ v \geqslant 0\}$ in the Dirichlet space $\mathcal{H}$ equipped with the scalar product $(u, v)_{\mathcal{H}}$. We have

$$T_K(u) = \{v \in \mathcal{H} \ : \ v \geqslant 0 \text{ on } \{u = 0\}\}.$$

The convex cone $S := T_K(u) \cap m^\perp$ is important for our applications. It is obtained in [12]

$$T_K(u) \cap m^\perp = \{v \in \mathcal{H} \ : \ v \geqslant 0 \text{on } \{u = 0\} \quad \text{and } v = 0 \ m - \text{a.e.}\}.$$

The following result on the directional differentiability of metric projection holds for polyhedric convex sets [15, 36].

Lemma 14 *Let K be a polyhedric cone. For $t > 0$, t small enough,*

$$P_K(u + th) = P_K(u) + t P_S(h) + o(t; h) \text{in } H^{1/2}(\Sigma)$$

where

$$S := T_K(u) \cap m^\perp$$

and the remainder $o(t; h)$ is uniform on compact subsets of $H^{1/2}(\Sigma)$. Hence, the directional derivative of the metric projection is uniquely determined by the variational inequality

$$q := P_S(h) \in S \ : \ (q - h, v - q)_{1/2,\Sigma} \geqslant 0 \quad \forall v \in S.$$

For a crack $\Gamma_c \subset \Sigma$ we introduce the following convex cones

$$\mathcal{K}(\Sigma) := \{v \in H^{1/2}(\Sigma) \ : \ v = 0 \quad \text{on } \Sigma \setminus \overline{\Gamma}_c, \quad v \geqslant 0 \text{ on } \Gamma_c\},$$

and

$$\mathcal{K}(\Gamma_c) := \{v \in H^{1/2}_{00}(\Gamma_c) \ : \ v \geqslant 0 \text{ on } \Gamma_c\}.$$

For the variational problems with unilateral conditions for the jump of normal component of the displacement vector field over the crack, the convex cones $\mathcal{K}(\Gamma_c)$ and $\mathcal{K}(\Sigma)$ are employed in order to show the polyhedricity of the cone of admissible displacements.

Remark 15 The proof of Lemma 12 applies as well to the convex cone $\mathcal{K}(\Gamma_c) \subset H^{1/2}_{00}(\Gamma_c)$ since the space $C^\infty_0(\Gamma_c)$ is dense in $H^{1/2}_{00}(\Gamma_c)$, hence, a nonnegative distribution is a Radon measure. In addition, *contraction operates* [5] for the scalar product (16) in $H^{1/2}_{00}(\Gamma_c)$. Let us note that the scalar products in $H^{1/2}(\Sigma)$ and in $H^{1/2}_{00}(\Gamma_c)$ are not the same, the latter is a weighted space.

We recall an abstract result on shape sensitivity analysis of variational inequalities. The conical differentiability of solutions to variational inequalities for the crack problem follows from the abstract result given by Theorem 17. The general result [53] is adapted here to our setting within the domain decomposition framework. Thus, the bilinear form $a(\cdot,\cdot) + b_t(\cdot,\cdot)$ defined in the subdomain Ω_c is introduced, where $b_t(\cdot,\cdot)$ is the contribution from the Steklov-Poincaré operator on $\Gamma_R = \partial\Omega_R$. The real parameter $t > 0$ governs the shape perturbations of the inclusion $t \to \omega_t$ in Ω_R, where $t \to 0$ governs the topological changes of Ω_R in the framework of asymptotic analysis. The two boundary value problems in two subdomains are coupled by the transmission conditions on the interface Γ_R. The linear boundary value problem in Ω_R furnishes the expansions of the Steklov-Poincaré operators resulting from perturbations of the inclusion in the interior of the subdomain. The sensitivity analysis of solutions to variational inequality in Ω_c is performed for compact perturbations of nonlocal boundary conditions on the interface. As a result, the weak solution to the unilateral elasticity boundary value problem under considerations is directionally differentiable with respect to the parameter $t \to 0$ which governs the perturbations of the inclusion far from the crack. We provide the precise result on the conical differentiability of solutions to variational inequalities [15, 36, 53] (see also [12]) which is given here without proof.

Let $\mathcal{K} \subset \mathcal{H}$ be a convex and closed subset of a Hilbert space $\mathcal{H}$, and let $\langle\cdot,\cdot\rangle$ denote the duality pairing between $\mathcal{H}'$ and $\mathcal{H}$, where $\mathcal{H}'$ denotes the dual of $\mathcal{H}$. Let us assume that there are given symmetric bilinear forms $a(\cdot,\cdot)+b_t(\cdot,\cdot) \ : \ \mathcal{H}\times\mathcal{H} \to \mathbb{R}$ parametrized by $t \geqslant 0$, and the linear form $f \in \mathcal{H}'$, such that

Condition 16 *1. There are $0 < \alpha \leqslant M$ such that*

$$|a(u,v) + b_t(u,v)| \leqslant M\|u\|\|v\|, \quad \alpha\|u\|^2 \leqslant a(v,v) + b_t(v,v) \quad \forall u, v \in \mathcal{H}$$

uniformly with respect to $t \in [0, t_0)$. Furthermore, there exists $\mathcal{Q}' \in \mathcal{L}(\mathcal{H};\mathcal{H}')$ such that

$$\mathcal{Q}_t = \mathcal{Q} + t\mathcal{Q}' + o(t) \quad \text{in} \quad \mathcal{L}(\mathcal{H};\mathcal{H}'),$$

where $\mathcal{Q}_t \in \mathcal{L}(\mathcal{H}; \mathcal{H}')$

$$a(\phi, \varphi) + b_t(\phi, \varphi) = \langle \mathcal{Q}_t(\phi), \varphi \rangle \quad \forall \phi, \varphi \in \mathcal{H}.$$

2. *The set* $\mathcal{K} \subset \mathcal{H}$ *is convex and closed, and the solution operator* $\mathcal{H}' \ni f \to \mathcal{P}(f) \in \mathcal{H}$ *for* (15)

$$\mathcal{P}(f) \in \mathcal{K}: \quad a(\mathcal{P}(f), \varphi - \mathcal{P}(f)) \geq \langle f, \varphi - \mathcal{P}(f) \rangle \quad \forall \varphi \in \mathcal{K}$$

is differentiable in the sense that

$$\forall h \in \mathcal{H}': \quad \mathcal{P}(f + sh) = \mathcal{P}(f) + s\mathcal{P}'(h) + o(s) \quad in \quad \mathcal{H}$$

for $s > 0$, *s small enough, where the mapping* $\mathcal{P}' : \mathcal{H}' \to \mathcal{H}$ *is continuous and positively homogeneous, in addition, the remainder* $o(s)$ *is uniform with respect to the direction* $h \in \mathcal{H}'$ *on compact subsets of* $\mathcal{H}'$.

Let us consider the unique solutions $u_t = \mathcal{P}_t(f)$ to variational inequalities depending on a parameter $t \in [0, t_0)$, $t_0 > 0$,

$$u_t \in \mathcal{K} \ : \ a(u_t, \varphi - u_t) + b_t(u_t, \varphi - u_t) \geqslant \langle f, \varphi - u_t \rangle \qquad \forall \varphi \in \mathcal{K}. \tag{14}$$

In particular, for $t = 0$

$$u \in \mathcal{K} \ : \ a(u, \varphi - u) + b(u, \varphi - u) \geq \langle f, \varphi - u \rangle \qquad \forall \varphi \in \mathcal{K}, \tag{15}$$

with $u = \mathcal{P}(f)$ a unique solution to (15). The mapping $t \to u_t$ is strongly differentiable in the sense of Hadamard at 0^+, and its derivative is given by a unique solution of the auxiliary variational inequality [53].

Theorem 17 *Assume that Condition 16 is satisfied. Then the solutions to the variational inequality (14) are right-differentiable with respect to* t *at* $t = 0$, *i.e. for* $t > 0$, t *small enough,*

$$u_t = u + tu' + o(t) \quad in \quad \mathcal{H},$$

where

$$u' = \mathcal{P}'(-\mathcal{Q}'u).$$

3.1 Metric Projection onto Positive Cone in $H_{00}^{1/2}(\Gamma_c)$

For boundary value problems in domains with cracks, unilateral conditions are prescribed on the crack for the normal component of the displacement field. Hence, the normal component of the displacement field belongs to the positive cone in the fractional Sobolev space $H_{00}^{1/2}(\Gamma_c)$. The sensitivity analysis of variational inequalities

for Signorini problems was reduced in [53] to the directional differentiability of the metric projection onto the positive cone in a fractional space which is the Dirichlet space. This result is further extended in [12] to some crack problem. The method is also used in the present paper, however for the other purposes.

Sensitivity analysis of the crack problem. We are going to explain how the results obtained in [53] for the Signorini problem in linear elasticity can be extended to the crack problems with unilateral constraints. To this end, the abstract analysis performed in [12] for the differentiability of the metric projection onto the cone of nonnegative elements in the Dirichlet space is employed. The framework for analysis is established in function spaces over $\Omega := \Omega^+ \cup \Sigma \cup \Omega^-$, where Σ is a $C^{1,1}$ regular curve without intersections. The regularity assumption can be weakened, if necessary. Let $\Gamma_c \subset \Sigma$ be the segment $\{(x_1, 0) \ : \ 0 < x_1 < 1\}$ included in the curve Σ. We denote by ν the unit normal vector field on Σ which points out of Ω^+, and by n the unit normal vector field on $\partial\Gamma_c$ orthogonal to ν. We consider deformations of the crack in the direction of the vector field V colinear with n in the neighbourhood of the crack tip $A = (1, 0) \in \Omega_c \subset \mathbb{R}^2$. In the Sobolev space defined on the cracked domain Ω_c, the elements enjoy jumps over the crack which are denoted by $[\![v]\!] := v^+ - v^-$, and we have the regularity property of traces $[\![v]\!] \in H^{1/2}_{00}(\Gamma_c)$. In our geometry of Ω_c, the Sobolev space $H^{1/2}_{00}(\Gamma_c)$ coincides with the linear subspace of $H^{1/2}(\Sigma)$

$$H^{1/2}_{00}(\Gamma_c) = \{\varphi \in H^{1/2}(\Sigma) \ : \ \varphi = 0 \quad \text{q.e. on } \Sigma \setminus \Gamma_c\},$$

where q.e. means *quasi-everywhere* with respect to the capacity, see e.g. [47] for the definition and elementary properties of the capacity useful for the existence of optimal shapes in shape optimization problems with nonlinear PDE's constraints. In order to investigate the properties of the metric projection in the space of admissible displacement fields onto the convex cone

$$K := \{v \in H^1(\Omega_c) \ : \ [\![v]\!]\nu \geqslant 0\},$$

where $H^1(\Omega_c) := H^1(\Omega_c; \mathbb{R}^2)$, we need to show that the positive convex cone

$$\mathcal{K} = \{\varphi \in H^{1/2}_{00}(\Gamma_c) \ : \ \varphi \geqslant 0 \quad \text{on} \quad \Gamma_c\}.$$

is polyhedric in the sense of [12, 15, 36]. We consider here the rectilinear crack Γ_c in two spatial dimensions. The scalar product in $H^{1/2}_{00}(\Gamma_c) := H^{1/2}_{00}(0, 1)$ is defined

$$\langle \varphi, \psi \rangle_c = \int_{\Gamma_c} \int_{\Gamma_c} \frac{(\varphi(x) - \varphi(y))(\psi(x) - \psi(y))}{|x - y|^2} dxdy + \int_{\Gamma_c} \left[\varphi(x)\psi(x) + \frac{\varphi(x)\psi(x)}{\text{dist}(x, \partial\Gamma_c)} \right] dx \tag{16}$$

Polyhedricity of the positive cone in $\mathbf{H}_{\mathbf{00}}^{\mathbf{1/2}}(\mathbf{\Gamma_c})$. In order to show the polyhedricity of the nonnegative cone $\mathcal{K}$ in $\mathcal{H} := H_{00}^{1/2}(0,1)$, it is enough to check the property

$$\langle \varphi^+, \varphi^- \rangle_c \leqslant 0 \quad \forall v \in H_{00}^{1/2}(0,1)$$

which is straightforward, here $\varphi^+(x) = \max\{v(x), 0\}$. The full proof of polyhedricity in such a case is provided in [12]. It is easy to check that the polyhedricity with respect to the scalar product implies the polyhedricity with respect to a bilinear form which is equivalent to the scalar product.

Theorem 18 *Let us consider the variational inequality for the metric projection of $f + th \in \mathcal{H}$ onto $\mathcal{K}$*

$$u_t \in \mathcal{K} \ : \ \langle u_t - f - th, v - u_t \rangle \geqslant 0 \quad \forall v \in \mathcal{K},$$

where $f, h \in \mathcal{H}$ are given, denote by $\Xi\{u\} = \{x \in \Gamma_c \ : \ u(x) = 0\}$. Then

$$u_t = u + tq(h) + o(t; h) \ in \quad \mathcal{H},$$

where the remainder $o(t; h)$ is uniform on compact subsets of $\mathcal{H}$, and the conical diffferential of the metric projection $q := q(h)$ is given by the unique solution to the variational inequality

$$q \in \mathcal{S}(u) \ : \ \langle q - h, v - q \rangle \geqslant 0 \quad \forall v \in \mathcal{S}(u)$$

and the closed convex cone

$$\mathcal{S}(u) = \{\varphi \in \mathcal{H} \ : \ \varphi \geqslant 0 q.e. \ on \ \Xi\{u\}, \quad \langle u - f, \varphi \rangle = 0\}.$$

4 Rectilinear Crack in Two Spatial Dimensions

In this section the general method of shape-topological sensitivity analysis is presented in the domain $\Omega := \Omega_c \cup \Gamma_R \cup \Omega_R$, where the first subdomain Ω_c contains the rectilinear crack Γ_c and the second subdomain Ω_R contains the inclusion ω. We denote by $\Omega_{\text{in}} := \Omega_c \cup \overline{\Gamma}_c$, the first subdomain in the elastic body without the crack. We assume that there is a regular $C^{1,1}$-curve $\Sigma \subset \Omega_{\text{in}}$, without intersections, which contains the rectilinear crack $\Gamma_c := \{(x_1, 0) \ : \ 0 \leqslant x_1 \leqslant 1\}$. To simplify the presentation, let us consider a torus $\Omega := \mathbb{T} := \mathbb{T}^2$ with 2π-periodic coordinates $x = (x_1, x_2)$. The deformations of the subdomain Ω_c are defined by the vector field $(x, t) \to V(x, t) = (v(x, t), 0)$, where the $C_0^\infty(\Omega^+)$ function $(x, t) \to v(x, t)$ is supported in $[1 - \delta, 1 + \delta]^2 \times [-t_0, t_0] \subset \Omega^+ \subset \mathbb{R}^2 \times \mathbb{R}$ and $v(x, t) \equiv 1$ on $[1 - \delta/2, 1 + \delta/2]^2 \times [-t_0/2, t_0/2]$. In our notation, the real variable $t \in \mathbb{R}$ is a parameter. It means that the vector field V deforms the reference domain Ω_c^+ to

$t \to T_t(V)(\Omega_c^+)$ just by moving the tip of the crack $X = (1, 0) \to x(t) = (x_1(t), 0)$ in the direction of the x_1-axis. The mapping $T_t : X \to x(t)$ is given by the system of equations

$$\frac{dx}{dt}(t) = V(x(t), t), \quad x(0) = X.$$

The boundary value problem of linear isotropic elasticity in Ω_c is defined by the variational inequality

$$u \in K : a(u, v - u) \geqslant (f, v - u) \quad \forall u \in K,$$

where

$$K = \{v \in H^1(\Omega_c) : [\![v]\!] \cdot \nu := (v^+ - v^-) \cdot \nu \geqslant 0 \quad \text{on} \quad \Gamma_c\}.$$

The bilinear form

$$a(u, v) = \int_{\Omega_c} \left[\frac{\mu}{2} \sum_{j,k=1}^{2} (\partial_j u_k + \partial_k u_j)(\partial_j v_k + \partial_k v_j) + \lambda \text{div}\, u \text{div}\, v \right] dx$$

is associated with the operator

$$Lu := -\mu \Delta u - (\lambda + \mu) \textbf{grad}\, \text{div}\, u.$$

The deformation tensor $2\varepsilon(u) = \partial_j u_k + \partial_k u_j$ as well as the stress tensor $\sigma(u) =$ associated with the displacement field u are useful in the description of the boundary value problems in linear elasticity. The energy functional $\mathcal{E}(\Omega_c) = 1/2a(u, u) - (f, u)_{\Omega_c}$ is twice differentiable [12] in the direction of a vector field V, for the specific choice of the field $V = (v, 0)$. The first order shape derivative

$$V \to d\mathcal{E}(\Omega_c; V) = \frac{1}{t} \lim_{t \to 0} (\mathcal{E}(T_t(\Omega_c)) - (\mathcal{E}(\Omega_c))$$

can be interpreted as the derivative of the elastic energy with respect to the crack length, we refer the reader to [26] for the proof, the same result for the Laplacian is given in [24, 25].

Theorem 19 *We have*

$$d\mathcal{E}(\Omega_c; V) = \frac{1}{2} \int_{\Omega_c} \left\{ \text{div} V \cdot \varepsilon_{ij}(u) - 2E_{ij}(V; u) \right\} \sigma_{ij}(u) - \int_{\Omega_c} \text{div}(V f_i) u_i. \quad (17)$$

Now we restrict our consideration to the perturbation of the crack tip only in the direction which coincides with the crack direction. The derivative is evaluated in the framework of the velocity method [53] for a specific velocity vector field V

selected in such a way that the result $d\mathcal{E}(\Omega_c; V)$ is independent of the field V and it depends only on the perturbation of the crack tip. That is why, this derivative is called the *Griffith functional* $J(\Omega_c) := d\mathcal{E}(\Omega_c; V)$ defined for the elastic energy in a domain with crack. We are interested in the dependence of this functional on domain perturbations far from the crack. As a result, shape and topological derivatives of the nonsmooth Griffith shape functional are obtained with respect to the boundary variations of an inclusion.

4.1 Green Formulae and Steklov-Poincaré Operators

The Steklov-Poincaré operator on the interface for the domain $\Omega_c \cup \Gamma_R \cup \Omega_R$ is defined by the Green formula, first as the Dirichlet-to-Neumann map in Ω_R, then it is used on the interface as nonlocal boundary operator. Therefore, we recall here the Green formula for linear elasticity operators in two and three spatial dimensions. We start with analysis in two spatial dimensions. To simplify the presentation let us consider the reference domain without a crack in the form of the torus $\mathbb{T} := \mathbb{T}^2$ with 2π-periodic coordinates $x = (x_1, x_2)$. For the purpose of shape-topological sensitivity analysis we assume that the elastic body without the crack is decomposed into two subdomains, $\Omega_{\rm in}$ and Ω_R, separated from each other by the interface Γ_R. Thus, the elastic body with the crack Γ_c is written as

$$\Omega := \Omega_c \cup \Gamma_R \cup \Omega_R.$$

The rectilinear crack $\Gamma_c \subset \Sigma \subset \Omega_{in}$ is an open set, where the fictitious interface $\Sigma \subset \Omega_{in}$ is a closed $C^{1,1}$-curve without intersections. In our notation $\Omega_c = \Omega_{\rm in} \setminus \overline{\Gamma}_c$. The bilinear form of the linear isotropic elasticity is associated with the operator

$$Lu := -\mu \Delta u - (\lambda + \mu)\, \mathbf{grad}\ \mathrm{div}\, u$$

for given Lame coefficients $\mu > 0, \lambda \geqslant 0$. The displacement field u in the elastic body Ω is given by the unique solution of the variational inequality

$$u \in K \ : \ a(u, v - u) \geqslant (f, v - u) \quad \forall u \in K,$$

where

$$K = \{v \in H^1(\Omega_c) \ : \ [\![v]\!] \cdot \nu := (v^+ - v^-) \cdot \nu \geqslant 0 \text{ on } \ \Gamma_c\}.$$

Given the unique solution $u \in K$ of the variational inequality and the admissible vector field V compactly supported in Ω_c, we consider the associated shape functional (17) evaluated in Ω_c, which is called the Griffith functional

$$J(\Omega_c) := d\mathcal{E}(\Omega_c; V). \tag{18}$$

Let $\omega \subset \Omega_R$ be an elastic inclusion. Introduce the family of inclusions $t \to \omega_t \subset \Omega_R$ governed by the velocity field W compactly supported in Ω_R. The elastic energy in Ω_R with the inclusion ω_t is denoted by

$$\omega_t \to \mathcal{E}_t(\Omega_R) := \frac{1}{2} a_t(\Omega_R\ ;\ u, u) - (f, u)_{\Omega_R}.$$

Its shape derivative $d\mathcal{E}(\Omega_R; W)$ in the direction W is obtained by differentiation at $t = 0$ of the function

$$t \to \mathcal{E}_t(\Omega_R) := \frac{1}{2} a_t(\Omega_R\ ;\ u, u) - (f, u)_{\Omega_R}.$$

Proposition 20 *Assume that the energy shape functional in the subdomain Ω_R,*

$$\omega \to \mathcal{E}(\Omega_R) := \frac{1}{2} a(\Omega_R\ ;\ u, u) - (f, u)_{\Omega_R}$$

is differentiable in the direction of the velocity field W compactly supported in a neighbourhood of the inclusion $\overline{\omega} \subset \Omega_R$, then the Griffith functional (18) is directionally differentiable in the direction of the velocity field W. Therefore, the second order directional shape derivative $d\mathcal{E}(\Omega; V, W)$ of the energy functional in Ω in the direction of fields V, W is obtained.

This result can be proved by the domain decomposition technique:

- the shape differentiability of the energy functional in the subdomain Ω_R implies the differentiability of the associated Steklov-Poincaré operator defined on the Lipschitz curve given by the interface $\overline{\Omega}_R \cap \overline{\Omega}_c$ with respect to the scalar parameter $t \to 0$ which governs the boundary variations of the inclusion ω;
- the expansion of the Steklov-Poincaré nonlocal boundary pseudodifferential operator obtained in the subdomain Ω_R is used in the boundary conditions for the variational inequality defined in the cracked subdomain Ω_c and leads to the conical differential of the solution to the unilateral problem in the subdomain;
- the one term expansion of the solution to the unilateral problem is used in the Griffith functional in order to obtain the directional derivative with respect to the boundary variations of the inclusion.

Remark 21 For the circular inclusion $\omega := \{x \in \Omega_R\ :\ |x - y| < r_0\}$, $r_0 > 0$, the scalar parameter $t \to 0$ which governs the shape perturbations of $\partial\omega$ in the direction of a field W [53] can be replaced by the parameter $r \to r_0$. Thus, the moving domain $t \to \omega_t$ is replaced by the moving domain $r \to \{x \in \Omega_R\ :\ |x - y| < r\}$. In this way the shape sensitivity analysis [53] for $r_0 > 0$ and the topological sensitivity analysis [45] for $r_0 = 0^+$ are performed in the same framework for the simple case of circular inclusion.

5 Shape and Topological Derivatives of Elastic Energy in Two Spatial Dimensions for an Inclusion

In the subdomain Ω_c the unique weak solutions

$$\varepsilon \to u := u_\varepsilon$$

of the elasticity boundary value subproblem are given by the variational inequality

$$u \in K \ : \ a(\Omega_c; u, v-u) + b_\varepsilon(\Gamma_R; u, v-u) \geqslant (f, v-u)_{\Omega_c} \quad \forall v \in K.$$

In order to differentiate the solution mapping for this variational inequality, it is required to differentiate the bilinear form $\varepsilon \to b_\varepsilon(\Gamma_R; u, v)$, which is performed in this section.

5.1 Shape and Topological Derivatives of the Energy Functional in Ω_R with Respect to the Inclusion ω

In order to evaluate the topological derivative of energy functional in isotropic elasticity, the shape sensitivity analysis is combined with the asymptotic analysis [45]. In this section the small parameter is denoted by $\varepsilon \to 0$, and the circular inclusion $\varepsilon \to \omega_\varepsilon := B_\varepsilon$ is considered. The general shape of the inclusion $\varepsilon \to \omega_\varepsilon$ can be considered in the same way for shape sensitivity analysis [53] and asymptotic analysis [45]. For the sake of simplicity, the subscript R is omitted, thus, we denote $\Omega := \Omega_R$, since the inclusion is located in the subdomain Ω_R. We also allow for the Neumann Γ_N and Dirichlet Γ_D pieces of the boundary $\partial\Omega := \partial\Omega_R$, thus, $\partial\Omega_R := \Gamma_N \cup \Gamma_D \cup \Gamma$. Thus, we evaluate the shape and topological derivative [45] of the total potential energy associated to the plane stress linear elasticity problem, considering the nucleation of a small inclusion, represented by $B_\varepsilon \subset \Omega$, as the topological perturbation. In this way the expansion of the Steklov-Poincaré operator on the interface $\Gamma := \Gamma_R$ is obtained.

5.1.1 Steklov-Poincaré Operator

Let us consider the nonhomogeneous Dirichlet linear elasticity boundary value problem in the domain Ω with the boundary $\partial\Omega := \Gamma_N \cup \Gamma_D \cup \Gamma$.

$$\begin{cases} \text{Find } u, \text{ such that} \\ \operatorname{div}\sigma(u) = 0 & \text{in } \Omega, \\ \sigma(u) = \mathbb{C}\nabla u^s, \\ u = 0 & \text{on } \Gamma_D, \\ u = \overline{u} & \text{on } \Gamma, \\ \sigma(u)n = 0 & \text{on } \Gamma_N, \end{cases}$$

where the only columnists term is the Dirichlet condition $u = \overline{u}$ on the interface Γ. Let

$$a(u, u) := \int_\Omega \sigma(u) \cdot \nabla u^s$$

stands for the associated bilinear form, thus the elastic energy of the solution u is given by

$$\mathcal{E}(\Omega; u) = \frac{1}{2} a(u, u).$$

Then by Green's formula

$$\mathcal{E}(\Omega; u) = \langle \mathcal{T}(\overline{u}), \overline{u} \rangle_\Gamma.$$

In the case of an inclusion $\omega_\varepsilon \subset \Omega$, the formula becomes

$$\mathcal{E}_\varepsilon(\Omega; u) = \langle \mathcal{T}_\varepsilon(\overline{u}), \overline{u} \rangle_\Gamma. \tag{19}$$

Hence, the expansion of the energy functional in Ω, on the left hand side of (19) with respect to the parameter $\varepsilon \to 0$ can be used in order to determine the associated expansion of the Steklov-Poincaré operator $\overline{u} \to \mathcal{T}(\overline{u})$ on the right hand side of (19). Therefore, let us consider the smooth domain Ω with the boundary $\partial\Omega := \Gamma_N \cup \Gamma_D \cup \Gamma$, here Γ is the interface on which the Steklov-Poincaré operator introduced in our domain decomposition method is defined.

5.2 *Shape and Topological Differentiability of the Energy Functional for Expansion of Steklov-Poincaré Operator*

The notation of monograph [45] is used in this section. We recall the known results [45, 53] on the shape gradient of the energy functional $\varepsilon \to \mathcal{E}_\varepsilon(\Omega)$ with respect to moving interface $\varepsilon \to \partial\omega_\varepsilon$ which is the boundary of inclusion $\omega_\varepsilon \subset \Omega$. Finally, the topological derivative of the energy functional with respect to $\varepsilon \to 0^+$ is obtained [45]. In this way, the shape and topological differentiability of the Steklov-Poincaré operator on the fictitious interface Γ is established. Let us consider the subdomain

Ω_R with the interface $\Gamma_R \subset \partial\Omega_R$, which are denoted by Ω and Γ, respectively. Let us consider a circular inclusion in Ω. The inclusion $\omega_\varepsilon := B_\varepsilon(y) \subset \Omega_R$ depends on the parameter $\varepsilon \in [0, \varepsilon_0)$ $\varepsilon_0 \gg 0$. The energy functional $\varepsilon \to \mathcal{E}_\varepsilon(\Omega)$ is *shape differentiable* for $\varepsilon > 0$ and *topologically differentiable* for $\varepsilon = 0^+$. In this way the expansion of the Steklov-Poincaré operator is obtained on the interface Γ_R. The energy shape functional associated to the unperturbed domain with $\varepsilon = 0$, i.e., without inclusion, which we are dealing with is defined as

$$\psi(\chi) = \frac{1}{2}\int_\Omega \sigma(u) \cdot \nabla u^s,$$

where χ stands for the characteristic function of Ω, and the vector function u is the solution to the variational problem:

$$\begin{cases} \text{Find } u \in \mathcal{U}, \text{ such that} \\ \displaystyle\int_\Omega \sigma(u) \cdot \nabla \eta^s = 0 \quad \forall \eta \in \mathcal{V}, \\ \qquad \text{with} \quad \sigma(u) = \mathbb{C}\nabla u^s. \end{cases} \tag{20}$$

In the above equation, $\mathbb{C}$ is the constitutive tensor given by

$$\mathbb{C} = \frac{E}{1-\nu^2}((1-\nu)\mathbb{I} + \nu \mathrm{I} \otimes \mathrm{I}),$$

where I and $\mathbb{I}$ are the second and fourth order identity tensors, respectively, E is the Young modulus and ν the Poisson ratio, both considered constants everywhere. For the sake of simplicity, we also assume that the thickness of the elastic body is constant and equal to one. The convex set $\mathcal{U}$ written for the columnists Dirichlet boundary condition on the interface and the associated space of test functions $\mathcal{V}$ are respectively defined as

$$\mathcal{U} := \{\varphi \in H^1(\Omega; \mathbb{R}^2) : \varphi_{|\Gamma_D} = 0, \quad \varphi_{|\Gamma} = \overline{u}\},$$
$$\mathcal{V} := \{\varphi \in H^1(\Omega; \mathbb{R}^2) : \varphi_{|\Gamma_D} = 0 \quad \varphi_{|\Gamma} = 0\}.$$

In addition, $\partial\Omega = \Gamma \cup \Gamma_D \cup \Gamma_N$ with $\Gamma_D \cap \Gamma_N = \varnothing$, $\Gamma \cap \Gamma_N = \varnothing$, and $\Gamma_D \cap \Gamma = \varnothing$, where Γ_D and Γ_N are Dirichlet and Neumann boundaries, respectively. Thus, $\overline{u}$ is a Dirichlet data on Γ, and there are homogeneous Dirichlet data on Γ_D and Neumann data on Γ_N. The strong system associated to the variational problem (20) reads:

$$\begin{cases} \text{Find } u, \text{ such that} \\ \text{div}\sigma(u) = 0 & \text{in } \Omega, \\ \sigma(u) = \mathbb{C}\nabla u^s, \\ u = \overline{u} & \text{on } \Gamma, \\ u = 0 & \text{on } \Gamma_D, \\ \sigma(u)n = 0 & \text{on } \Gamma_N. \end{cases}$$

Remark 22 Since the Young modulus E and the Poisson ratio ν are constants, the above boundary value problem reduces to the well-known Navier system, namely

$$-\mu\Delta u - (\lambda+\mu)\nabla(\mathrm{div} u) = 0 \quad \text{in} \quad \Omega,$$

with the Lame's coefficients μ and λ respectively given by

$$\mu = \frac{E}{2(1+\nu)} \quad \text{and} \quad \lambda = \frac{\nu E}{1-\nu^2}.$$

Now, let us state the same problem in the perturbed domain which contains the inclusion B_ε. More precisely, the perturbed domain is obtained if a circular hole $B_\varepsilon(y)$ is introduced inside $\Omega \subset \mathbb{R}^2$, where $\overline{B_\varepsilon(y)} \Subset \Omega$ denotes a ball of radius ε and center at $y \in \Omega$. Then, $B_\varepsilon(y)$ is filled by an inclusion with different material property compared to the unperturbed domain Ω. The material properties are characterized by a piecewise constant function γ_ε of the form

$$\gamma_\varepsilon = \gamma_\varepsilon(x) := \begin{cases} 1 \text{ if } x \in \Omega \setminus \overline{B_\varepsilon}, \\ \gamma \text{ if } x \in B_\varepsilon, \end{cases} \tag{21}$$

where $\gamma \in \mathbb{R}_+$ is the contrast coefficient. In this case, the shape functional reads

$$\psi(\chi_\varepsilon) := \frac{1}{2}\int_\Omega \sigma_\varepsilon(u_\varepsilon)\cdot\nabla u_\varepsilon^s, \tag{22}$$

where the vector function u_ε solves the variational problem:

$$\begin{cases} \text{Find } u_\varepsilon \in \mathcal{U}_\varepsilon, \text{ such that} \\ \displaystyle\int_\Omega \sigma_\varepsilon(u_\varepsilon)\cdot\nabla\eta^s = 0 \quad \forall \eta \in \mathcal{V}_\varepsilon, \\ \qquad \text{with} \quad \sigma_\varepsilon(u_\varepsilon) = \gamma_\varepsilon \mathbb{C}\nabla u_\varepsilon^s. \end{cases} \tag{23}$$

with γ_ε given by (21). The set $\mathcal{U}_\varepsilon$ and the space $\mathcal{V}_\varepsilon$ are defined as

$$\mathcal{U}_\varepsilon := \{\varphi \in \mathcal{U} : [\![\varphi]\!] = 0 \text{ on } \partial B_\varepsilon\},$$
$$\mathcal{V}_\varepsilon := \{\varphi \in \mathcal{V} : [\![\varphi]\!] = 0 \text{ on } \partial B_\varepsilon\},$$

where the operator $[\![\varphi]\!]$ is used to denote the jump of function φ on the boundary of the inclusion ∂B_ε, namely $[\![\varphi]\!] := \varphi_{|_{\Omega\setminus\overline{B_\varepsilon}}} - \varphi_{|_{B_\varepsilon}}$ on ∂B_ε. The *strong system* associated to the variational problem (23) reads:

$$\left\{\begin{array}{rcll} \text{Find } u_\varepsilon, \text{ such that} & & & \\ \text{div}\sigma_\varepsilon(u_\varepsilon) & = & 0 & \text{in } \Omega, \\ \sigma_\varepsilon(u_\varepsilon) & = & \gamma_\varepsilon \mathbb{C}\nabla u_\varepsilon^s, & \\ u_\varepsilon & = & \overline{u} & \text{on } \Gamma, \\ u_\varepsilon & = & 0 & \text{on } \Gamma_D, \\ \sigma(u_\varepsilon)n & = & 0 & \text{on } \Gamma_N, \\ [\![u_\varepsilon]\!] & = & 0 & \\ [\![\sigma_\varepsilon(u_\varepsilon)]\!]n & = & 0 & \text{on } \partial B_\varepsilon. \end{array}\right. \tag{24}$$

The *transmission condition* on the boundary of the inclusion ∂B_ε comes out from the variation formulation (23).

5.3 Shape Derivative of Steklov-Poincaré Operator

The next step consists in evaluating the shape derivative of functional $\psi(\chi_\varepsilon)$ with respect to an uniform expansion of the inclusion B_ε. In the particular case of circular inclusions, for a given $y \in \Omega$ and $0 < \varepsilon < \ell$, with $\ell := \text{dist}(y, \partial\Omega)$, we can construct a shape change velocity field $\mathfrak{V}$ that represents uniform expansion of $B_\varepsilon(y)$. In fact, it is sufficient to define $\mathfrak{V}$ on the boundary ∂B_ε i.e., $\mathfrak{V}_{|_{\partial B_\varepsilon(y)}} = -n$, where $n = -(x-y)/\varepsilon$, with $x \in \partial B_\varepsilon$, is the normal unit vector field pointing toward the center of the circular inclusion B_ε Let us introduce the *Eshelby energy-momentum tensor* [45], namely

$$\mathbb{E}_\varepsilon = \frac{1}{2}(\sigma_\varepsilon(u_\varepsilon) \cdot \nabla u_\varepsilon^s)\mathrm{I} - \nabla u_\varepsilon^\top \sigma_\varepsilon(u_\varepsilon). \tag{25}$$

In addition, we note that after considering the constitutive relation $\sigma_\varepsilon(u_\varepsilon) = \gamma_\varepsilon \mathbb{C}\nabla u_\varepsilon^s$ in (22), with the contrast γ_ε given by (21), the shape functional $\psi(\chi_\varepsilon)$ can be written as follows

$$\psi(\chi_\varepsilon) = \frac{1}{2}\left(\int_{\Omega\setminus\overline{B_\varepsilon}} \sigma(u_\varepsilon) \cdot \nabla u_\varepsilon^s + \int_{B_\varepsilon} \gamma\sigma(u_\varepsilon) \cdot \nabla u_\varepsilon^s\right), \tag{26}$$

where $\sigma(u_\varepsilon) = \mathbb{C}\nabla u_\varepsilon^s$. Therefore, the explicit dependence with respect to the parameter ε arises, and we recall the following result [45]

Proposition 23 *Let $\psi(\chi_\varepsilon)$ be the energy shape functional defined by (22). Then, the shape derivative of $\psi(\chi_\varepsilon)$ with respect to the small parameter $\varepsilon > 0$ is given by*

$$\dot{\psi}(\chi_\varepsilon) = \int_\Omega \mathbb{E}_\varepsilon \cdot \nabla\mathfrak{V},$$

where $\mathfrak{V}$ is the shape change velocity field defined by an extension of the normal vector field $n = -(x-y)/\varepsilon$, with $x \in \partial B_\varepsilon$, and $\mathbb{E}_\varepsilon$ is the Eshelby energy-momentum tensor given by (25).

Proof Before starting, let us recall that the constitutive operator is defined as $\sigma_\varepsilon(\varphi) = \gamma_\varepsilon \mathbb{C}\nabla\varphi^s$. Thus, by making use of the Reynolds' transport theorem and the concept of material derivative of spatial fields [45], the derivative with respect to ε of the shape functional (26) is given by

$$\begin{aligned}\dot{\psi}(\chi_\varepsilon) &= \frac{1}{2}\left(\int_{\Omega\setminus\overline{B_\varepsilon}} \sigma(u_\varepsilon)\cdot\nabla u_\varepsilon^s + \int_{B_\varepsilon} \gamma\sigma(u_\varepsilon)\cdot\nabla u_\varepsilon^s\right)\\ &= \int_{\Omega\setminus\overline{B_\varepsilon}} \sigma(u_\varepsilon)\cdot\nabla \dot{u}_\varepsilon^s + \int_{B_\varepsilon} \gamma\sigma(u_\varepsilon)\cdot\nabla \dot{u}_\varepsilon^s\\ &\quad + \frac{1}{2}\int_{\Omega\setminus\overline{B_\varepsilon}} ((\sigma(u_\varepsilon)\cdot\nabla u_\varepsilon^s)\mathrm{I} - 2\nabla u_\varepsilon^\top \sigma(u_\varepsilon))\cdot\nabla\mathfrak{V}\\ &\quad + \frac{1}{2}\int_{B_\varepsilon} \gamma((\sigma(u_\varepsilon)\cdot\nabla u_\varepsilon^s)\mathrm{I} - 2\nabla u_\varepsilon^\top \sigma(u_\varepsilon))\cdot\nabla\mathfrak{V}.\end{aligned}$$

Then,

$$\begin{aligned}\dot{\psi}(\chi_\varepsilon) &= \frac{1}{2}\int_{\Omega} ((\sigma_\varepsilon(u_\varepsilon)\cdot\nabla u_\varepsilon^s)\mathrm{I} - 2\nabla u_\varepsilon^\top \sigma_\varepsilon(u_\varepsilon))\cdot\nabla\mathfrak{V}\\ &\quad + \int_{\Omega} \sigma_\varepsilon(u)\cdot\nabla \dot{u}_\varepsilon^s.\end{aligned}$$

Since $\dot{u}_\varepsilon$ is a variation of u_ε in the direction of the velocity field $\mathfrak{V}$, then $\dot{u}_\varepsilon \in \mathcal{V}_\varepsilon$ [53]. Finally, by taking $\dot{u}_\varepsilon$ as test function in the variational problem (23), we have that the last two terms of the above equation vanish. □

The shape gradient of energy functional is supported on the moving interface $\varepsilon \to \partial B_\varepsilon$ as it is predicted by the structure theorem of the shape gradient [45, 53].

Proposition 24 *Let $\psi(\chi_\varepsilon)$ be the shape functional defined by (22). Then, its derivative with respect to the small parameter ε is given by*

$$\dot{\psi}(\chi_\varepsilon) = \int_{\partial B_\varepsilon} [\![\mathbb{E}_\varepsilon]\!] n\cdot\mathfrak{V}, \tag{27}$$

with $\mathfrak{V}$ standing for the shape change velocity field compactly supported in a neighbourhood of ∂B_ε and tensor $\mathbb{E}_\varepsilon$ given by (25).

Proof Before starting, let us recall the constitutive operator $\sigma_\varepsilon(\varphi) = \gamma_\varepsilon\mathbb{C}\nabla\varphi^s$ and the relation between material and spatial derivatives of vector fields $\dot{\varphi} = \varphi' + (\nabla\varphi)\mathfrak{V}$. By making use of the Reynolds' transport theorem [45], the shape derivative of the functional (22) results in

$$\begin{aligned}\dot{\psi}(\chi_\varepsilon) &= \left(\frac{1}{2}\int_\Omega \sigma_\varepsilon(u_\varepsilon)\cdot\nabla u_\varepsilon^s\right)\\ &= \int_\Omega \sigma_\varepsilon(u_\varepsilon)\cdot(\nabla u_\varepsilon')^s + \frac{1}{2}\int_{\partial\Omega}(\sigma_\varepsilon(u_\varepsilon)\cdot\nabla u_\varepsilon^s)n\cdot\mathfrak{V}\\ &\quad + \frac{1}{2}\int_{\partial B_\varepsilon}[\![\sigma_\varepsilon(u_\varepsilon)\cdot\nabla u_\varepsilon^s]\!]n\cdot\mathfrak{V}.\end{aligned}$$

In addition, we have

$$\begin{aligned}\dot{\psi}(\chi_\varepsilon) &= \frac{1}{2}\int_{\partial\Omega}(\sigma_\varepsilon(u_\varepsilon)\cdot\nabla u_\varepsilon^s)n\cdot\mathfrak{V} + \frac{1}{2}\int_{\partial B_\varepsilon}[\![\sigma_\varepsilon(u_\varepsilon)\cdot\nabla u_\varepsilon^s]\!]n\cdot\mathfrak{V}\\ &\quad - \int_\Omega \sigma_\varepsilon(u_\varepsilon)\cdot\nabla((\nabla u_\varepsilon)\mathfrak{V})^s + \int_\Omega \sigma_\varepsilon(u)\cdot\nabla\dot{u}_\varepsilon^s.\end{aligned}$$

Since $\dot{u}_\varepsilon$ is a variation of u_ε in the direction of the velocity field $\mathfrak{V}$, then $\dot{u}_\varepsilon \in \mathcal{V}_\varepsilon$ [53]. Now, by taking into account that u_ε is the solution to the variational problem (23), we have that the last two terms of the above equation vanish. From integration by parts

$$\begin{aligned}\dot{\psi}(\chi_\varepsilon) &= \frac{1}{2}\int_{\partial\Omega}(\sigma_\varepsilon(u_\varepsilon)\cdot\nabla u_\varepsilon^s)n\cdot\mathfrak{V} + \frac{1}{2}\int_{\partial B_\varepsilon}[\![\sigma_\varepsilon(u_\varepsilon)\cdot\nabla u_\varepsilon^s]\!]n\cdot\mathfrak{V}\\ &\quad - \int_{\partial\Omega}(\nabla u_\varepsilon^\top\sigma_\varepsilon(u_\varepsilon))n\cdot\mathfrak{V} - \int_{\partial B_\varepsilon}[\![\nabla u_\varepsilon^\top\sigma_\varepsilon(u_\varepsilon)]\!]n\cdot\mathfrak{V}\\ &\quad + \int_\Omega \mathrm{div}(\sigma_\varepsilon(u_\varepsilon))\cdot(\nabla u_\varepsilon)\mathfrak{V},\end{aligned}$$

and rewriting the above equation in the compact form, we obtain

$$\dot{\psi}(\chi_\varepsilon) = \int_{\partial\Omega}\mathbb{E}_\varepsilon n\cdot\mathfrak{V} + \int_{\partial B_\varepsilon}[\![\mathbb{E}_\varepsilon]\!]n\cdot\mathfrak{V} + \int_\Omega \mathrm{div}(\sigma_\varepsilon(u_\varepsilon))\cdot(\nabla u_\varepsilon)\mathfrak{V}.$$

Finally, taking into account that u_ε is the solution to the state equation (24), namely $\mathrm{div}\sigma_\varepsilon(u_\varepsilon) = 0$, we have that the last term in the above equation vanishes, which leads to the result. □

Corollary 25 *We have*

$$\dot{\psi}(\chi_\varepsilon) = \int_{\partial B_\varepsilon}[\![\mathbb{E}_\varepsilon]\!]n\cdot\mathfrak{V} - \int_\Omega \mathrm{div}\mathbb{E}_\varepsilon\cdot\mathfrak{V}.$$

Since the above equation and (27) remain valid for all velocity fields $\mathfrak{V}$, we have that the last term of the above equation must satisfy

$$\int_\Omega \mathrm{div}\mathbb{E}_\varepsilon\cdot\mathfrak{V} = 0 \quad \forall\mathfrak{V} \quad\Rightarrow\quad \mathrm{div}\mathbb{E}_\varepsilon = 0.$$

hence

$$\frac{d}{d\varepsilon}\psi(\chi_\varepsilon) = \dot{\psi}(\chi_\varepsilon) = -\int_{\partial B_\varepsilon} [\![\mathbb{E}_\varepsilon]\!] n \cdot n. \tag{28}$$

5.4 Application to Conical Differentiability for Model Problem

We return to the variational inequality with regularly perturbed bilinear form, see (32) for an application. Therefore, for given $\varepsilon > 0$ we consider the variational inequality in Ω_c,

$$u_t \in K \ : \ a(u_t, v - u_t) + b_t(u_t, v - u_t) = (f, v - u_t) \quad \forall v \in K,$$

where for $t > 0$, t small enough, the symmetric, boundary bilinear form b_t is defined on $\Gamma := \Gamma_R$ by the elastic energy in Ω_R,

$$t \to b_t(u, u) := \langle \mathcal{T}_{\varepsilon+t}(\overline{u}), \overline{u} \rangle_\Gamma.$$

Here, $\overline{u}$ stands for the trace of u on Γ. Thus, the shape derivative of this bilinear form with respect to the deformations of interface ∂B_ε governed by $t \to 0$ is given by

$$b'(u, u) := \dot{\psi}(\chi_\varepsilon) = -\int_{\partial B_\varepsilon} [\![\mathbb{E}_\varepsilon]\!] n \cdot n.$$

In this case Lemma 14 applies and we have

Proposition 26 *For $t > 0$, t small enough,*

$$u_t = u + tq + o(t),$$

where

$$q \in S \ : \ a(q, v - q) + b(q, v - q) + b'(q, v - q) \geqslant 0 \quad \forall v \in S.$$

Remark 27 For $\varepsilon = 0^+$ the result remain valid with the modification that

$$u_\varepsilon = u + \varepsilon^2 q + o(\varepsilon^2),$$

and with the shape derivative of the Steklov-Poincaré replaced by the topological derivative which is evaluated in the section below.

Remark 28 Given the one term expansion of the solution to variational inequality in Ω_c with respect to ε, it is straightforward to obtain the directional derivative of the Griffith functional.

5.5 Topological Derivative of the Steklov-Poincaré Operator

We recall known results on topological sensitivity analysis given in [45] which are adapted to our setting. The shape derivative of functional $\dot{\psi}(\chi_\varepsilon)$ is given in terms of an integral over the boundary of the inclusion ∂B_ε (28). The formula for the topological derivative $\mathcal{J}_\psi$ of the shape functional ψ is obtained by asymptotic analysis of u_ε with respect to ε. The asymptotic expansion of the solution u_ε is associated to the transmission condition on the inclusion. We start with an *ansatz* for u_ε

$$u_\varepsilon(x) = u(x) + w_\varepsilon(x) + \widetilde{u}_\varepsilon(x).$$

After applying the operator σ_ε, we have

$$\begin{aligned}\sigma_\varepsilon(u_\varepsilon(x)) &= \sigma_\varepsilon(u(x)) + \sigma_\varepsilon(w_\varepsilon(x)) + \sigma_\varepsilon(\widetilde{u}_\varepsilon(x)) \\ &= \sigma_\varepsilon(u(y)) + \nabla\sigma_\varepsilon(u(\widehat{y}))(x-y) + \sigma_\varepsilon(w_\varepsilon(x)) + \sigma_\varepsilon(\widetilde{u}_\varepsilon(x)),\end{aligned}$$

where $\widehat{y}$ is an intermediate point between x and y. On the boundary of the inclusion ∂B_ε we have

$$[\![\sigma_\varepsilon(u_\varepsilon)]\!]n = 0 \quad \Rightarrow \quad (\sigma(u_\varepsilon)_{|_{\Omega\setminus\overline{B_\varepsilon}}} - \gamma\sigma(u_\varepsilon)_{|_{B_\varepsilon}})n = 0,$$

with $\sigma_\varepsilon(\varphi) = \gamma_\varepsilon \mathbb{C}\nabla\varphi^s$ and $\sigma(\varphi) = \mathbb{C}\nabla\varphi^s$. The above expansion, evaluated on ∂B_ε, leads to

$$(1-\gamma)\sigma(u(y))n - \varepsilon(1-\gamma)(\nabla\sigma(u(y))n)n + [\![\sigma_\varepsilon(w_\varepsilon(x))]\!]n + [\![\sigma_\varepsilon(\widetilde{u}_\varepsilon(x))]\!]n = 0.$$

Thus, we can choose $\sigma_\varepsilon(w_\varepsilon)$ such that

$$[\![\sigma_\varepsilon(w_\varepsilon(x))]\!]n = -(1-\gamma)\sigma(u(y))n \quad \text{on} \quad \partial B_\varepsilon.$$

Now, the following exterior problem is considered, and formally obtained as $\varepsilon \to 0$:

$$\begin{cases} \text{Find } \sigma_\varepsilon(w_\varepsilon), \text{ such that} \\ \operatorname{div}\sigma_\varepsilon(w_\varepsilon) = 0 \text{ in } \mathbb{R}^2, \\ \sigma_\varepsilon(w_\varepsilon) \to 0 \text{ at } \infty, \\ [\![\sigma_\varepsilon(w_\varepsilon)]\!]n = \widehat{u} \text{ on } \partial B_\varepsilon, \end{cases}$$

with $\widehat{u} = -(1-\gamma)\sigma(u(y))n$. The above boundary value problem admits an explicit solution, which will be used later to construct the expansion for $\sigma_\varepsilon(u_\varepsilon)$. Now we can construct $\sigma_\varepsilon(\widetilde{u}_\varepsilon)$ in such a way that it compensates the discrepancies introduced by the higher order terms in ε as well as by the boundary layer $\sigma_\varepsilon(w_\varepsilon)$ on the exterior boundary $\partial\Omega$. It means that the remainder $\widetilde{u}_\varepsilon$ must be solution to the following boundary value problem:

$$\left\{\begin{array}{rcll} \text{Find } \widetilde{u}_\varepsilon, \text{ such that} & & & \\ \operatorname{div}\sigma_\varepsilon(\widetilde{u}_\varepsilon) &=& 0 & \text{in } \Omega, \\ \sigma_\varepsilon(\widetilde{u}_\varepsilon) &=& \gamma_\varepsilon \mathbb{C}\nabla \widetilde{u}^s_\varepsilon, & \\ \widetilde{u}_\varepsilon &=& -w_\varepsilon & \text{on } \Gamma_D, \\ \sigma(\widetilde{u}_\varepsilon)n &=& -\sigma(w_\varepsilon)n & \text{on } \Gamma_N, \\ \left.\begin{array}{r}[\![\widetilde{u}_\varepsilon]\!] = \\ [\![\sigma_\varepsilon(\widetilde{u}_\varepsilon)]\!]n =\end{array}\right. & \left.\begin{array}{l} 0 \\ \varepsilon h \end{array}\right. & & \text{on } \partial B_\varepsilon, \end{array}\right. \tag{29}$$

with $h = (1-\gamma)(\nabla\sigma(u(y))n)n$. The following lemma is proved in [45]:

Lemma 29 *Let $\widetilde{u}_\varepsilon$ be the solution to (29) or equivalently the solution to the following variational problem:*

$$\left\{\begin{array}{l} \textit{Find } \widetilde{u}_\varepsilon \in \widetilde{\mathcal{U}}_\varepsilon, \textit{ such that} \\ \displaystyle\int_\Omega \sigma_\varepsilon(\widetilde{u}_\varepsilon)\cdot\nabla\eta^s = \varepsilon^2 \int_{\Gamma_N} \sigma(g)n\cdot\eta + \varepsilon \int_{\partial B_\varepsilon} h\cdot\eta \quad \forall \eta \in \widetilde{\mathcal{V}}_\varepsilon, \\ \qquad\qquad \textit{with} \quad \sigma_\varepsilon(\widetilde{u}_\varepsilon) = \gamma_\varepsilon \mathbb{C}\nabla\widetilde{u}^s_\varepsilon, \end{array}\right.$$

where the set $\widetilde{\mathcal{U}}_\varepsilon$ and the space $\widetilde{\mathcal{V}}_\varepsilon$ are defined as

$$\widetilde{\mathcal{U}}_\varepsilon := \{\varphi \in H^1(\Omega;\mathbb{R}^2) : [\![\varphi]\!] = 0 \textit{ on } \partial B_\varepsilon,\ \varphi_{|\Gamma_D} = \varepsilon^2 g\},$$
$$\widetilde{\mathcal{V}}_\varepsilon := \{\varphi \in H^1(\Omega;\mathbb{R}^2) : [\![\varphi]\!] = 0 \textit{ on } \partial B_\varepsilon,\ \varphi_{|\Gamma_D} = 0\},$$

with functions $g = -\varepsilon^{-2} w_\varepsilon$ and $h = (1-\gamma)(\nabla\sigma(u(y))n)n$ independent of the small parameter ε. Then, we have the estimate $\|\widetilde{u}_\varepsilon\|_{H^1(\Omega;\mathbb{R}^2)} = O(\varepsilon^2)$ for the remainder.

Therefore, the *expansion* for $\sigma_\varepsilon(u_\varepsilon)$ can be written [45] in a polar coordinate system (r,θ) centered at the point y as:

- For $r \geqslant \varepsilon$ (outside the inclusion)

$$\begin{aligned} \sigma_\varepsilon^{rr}(u_\varepsilon(r,\theta)) &= \varphi_1\left(1 - \tfrac{1-\gamma}{1+\gamma\alpha}\tfrac{\varepsilon^2}{r^2}\right) \\ &\quad + \varphi_2\left(1 - 4\tfrac{1-\gamma}{1+\gamma\beta}\tfrac{\varepsilon^2}{r^2} + 3\tfrac{1-\gamma}{1+\gamma\beta}\tfrac{\varepsilon^4}{r^4}\right)\cos 2\theta + O(\varepsilon^2), \\ \sigma_\varepsilon^{\theta\theta}(u_\varepsilon(r,\theta)) &= \varphi_1\left(1 + \tfrac{1-\gamma}{1+\gamma\alpha}\tfrac{\varepsilon^2}{r^2}\right) \\ &\quad - \varphi_2\left(1 + 3\tfrac{1-\gamma}{1+\gamma\beta}\tfrac{\varepsilon^4}{r^4}\right)\cos 2\theta + O(\varepsilon^2), \\ \sigma_\varepsilon^{r\theta}(u_\varepsilon(r,\theta)) &= -\varphi_2\left(1 + 2\tfrac{1-\gamma}{1+\gamma\beta}\tfrac{\varepsilon^2}{r^2} - 3\tfrac{1-\gamma}{1+\gamma\beta}\tfrac{\varepsilon^4}{r^4}\right)\sin 2\theta + O(\varepsilon^2). \end{aligned}$$

- For $0 < r < \varepsilon$ (inside the inclusion)

$$\sigma_\varepsilon^{rr}(u_\varepsilon(r,\theta)) = \varphi_1\left(\tfrac{2}{1-\nu}\tfrac{\gamma}{1+\gamma\alpha}\right) + \varphi_2\left(\tfrac{4}{1+\nu}\tfrac{\gamma}{1+\gamma\beta}\right)\cos 2\theta + O(\varepsilon^2),$$
$$\sigma_\varepsilon^{\theta\theta}(u_\varepsilon(r,\theta)) = \varphi_1\left(\tfrac{2}{1-\nu}\tfrac{\gamma}{1+\gamma\alpha}\right) - \varphi_2\left(\tfrac{4}{1+\nu}\tfrac{\gamma}{1+\gamma\beta}\right)\cos 2\theta + O(\varepsilon^2),$$
$$\sigma_\varepsilon^{r\theta}(u_\varepsilon(r,\theta)) = -\varphi_2\left(\tfrac{4}{1+\nu}\tfrac{\gamma}{1+\gamma\beta}\right)\sin 2\theta + O(\varepsilon^2).$$

Some terms in the above formulae require explanations. The coefficients φ_1 and φ_2 are given by

$$\varphi_1 = \frac{1}{2}(\sigma_1(u(y)) + \sigma_2(u(y))), \quad \varphi_2 = \frac{1}{2}(\sigma_1(u(y)) - \sigma_2(u(y))),$$

where $\sigma_1(u(y))$ and $\sigma_2(u(y))$ are the eigenvalues of tensor $\sigma(u(y))$, which can be expressed as

$$\sigma_{1,2}(u(y)) = \frac{1}{2}\left(\operatorname{tr}\sigma(u(y)) \pm \sqrt{2\sigma^D(u(y))\cdot\sigma^D(u(y))}\right),$$

with $\sigma^D(u(y))$ standing for the deviatory part of the stress tensor $\sigma(u(y))$, namely

$$\sigma^D(u(y)) = \sigma(u(y)) - \frac{1}{2}\operatorname{tr}\sigma(u(y))\mathrm{I}.$$

In addition, the constants α and β are given by

$$\alpha = \frac{1+\nu}{1-\nu} \quad \text{and} \quad \beta = \frac{3-\nu}{1+\nu}. \tag{30}$$

Finally, $\sigma_\varepsilon^{rr}(u_\varepsilon)$, $\sigma_\varepsilon^{\theta\theta}(u_\varepsilon)$ and $\sigma_\varepsilon^{r\theta}(u_\varepsilon)$ are the components of tensor $\sigma_\varepsilon(u_\varepsilon)$ in the polar coordinate system, namely $\sigma_\varepsilon^{rr}(u_\varepsilon) = e^r \cdot \sigma_\varepsilon(u_\varepsilon)e^r$, $\sigma_\varepsilon^{\theta\theta}(u_\varepsilon) = e^\theta \cdot \sigma_\varepsilon(u_\varepsilon)e^\theta$ and $\sigma_\varepsilon^{r\theta}(u_\varepsilon) = \sigma_\varepsilon^{\theta r}(u_\varepsilon) = e^r \cdot \sigma_\varepsilon(u_\varepsilon)e^\theta$, with e^r and e^θ used to denote the canonical basis associated to the polar coordinate system (r, θ), such that, $\|e^r\| = \|e^\theta\| = 1$ and $e^r \cdot e^\theta = 0$.

5.6 Formulae for Topological Derivative

Now, we can evaluate the integral in formula (28). With this result, we can perform the limit passage $\varepsilon \to 0$. The integral in (28) can be evaluated by using the expansion for $\sigma_\varepsilon(u_\varepsilon)$ given by (30). The idea is to introduce a polar coordinate system (r, θ) with center at y. Then, we can write u_ε in this coordinate system to evaluate the integral explicitly. In particular, the integral in (28) yields

$$\int_{\partial B_\varepsilon} [\![\mathbb{E}_\varepsilon]\!] n \cdot n = 2\pi\varepsilon \mathbb{P}_\gamma \sigma(u(y)) \cdot \nabla u^s(y) + o(\varepsilon).$$

Finally,

$$\mathcal{J}_\psi(y) = -\lim_{\varepsilon \to 0} \frac{1}{f'(\varepsilon)} \left(2\pi\varepsilon \mathbb{P}_\gamma \sigma(u(y)) \cdot \nabla u^s(y) + o(\varepsilon) \right),$$

where the *polarization tensor* $\mathbb{P}_\gamma$ is given by the following fourth order isotropic tensor

$$\mathbb{P}_\gamma = \frac{1}{2} \frac{1-\gamma}{1+\gamma\beta} \left((1+\beta)\mathbb{I} + \frac{1}{2}(\alpha - \beta) \frac{1-\gamma}{1+\gamma\alpha} \mathrm{I} \otimes \mathrm{I} \right),$$

with the parameters α and β given by (30). Now, in order to extract the leading term of the above expansion, we choose

$$f(\varepsilon) = \pi\varepsilon^2,$$

which leads to the final formula for the *topological derivative*, namely

$$\mathcal{J}_\psi(y) = -\mathbb{P}_\gamma \sigma(u(y)) \cdot \nabla u^s(y).$$

Remark 30 Polarization tensors for cracks are considered e.g., in [41–44].

Finally, the topological asymptotic expansion of the energy shape functional takes the form

$$\psi(\chi_\varepsilon(y)) = \psi(\chi) - \pi\varepsilon^2 \mathbb{P}_\gamma \sigma(u(y)) \cdot \nabla u^s(y) + o(\varepsilon^2),$$

whose mathematical justification is given in [45].

Remark 31 We note that the obtained polarization tensor is isotropic because we are dealing with circular inclusions. Some results on the polarization tensor associated to arbitrary shaped inclusions can be found in [35, 43].

Remark 32 Formally, we can consider the limit cases $\gamma \to 0$ and $\gamma \to \infty$. For $\gamma \to 0$, the inclusion leads to a void and the transmission condition on the boundary of the inclusion degenerates to homogeneous Neumann boundary condition. In fact, in this case the polarization tensor is given by

$$\mathbb{P}_0 = \frac{1}{2}(1+\beta)\mathbb{I} + \frac{1}{4}(\alpha - \beta)\mathrm{I} \otimes \mathrm{I} = \frac{2}{1+\nu}\mathbb{I} - \frac{1-3\nu}{2(1-\nu^2)}\mathrm{I} \otimes \mathrm{I}.$$

In addition, for $\gamma \to \infty$, the elastic inclusion leads to a rigid one and the polarization tensor is given by

$$\mathbb{P}_\infty = -\frac{1+\beta}{2\beta}\mathbb{I} + \frac{\alpha - \beta}{4\alpha\beta}\mathrm{I} \otimes \mathrm{I} = -\frac{2}{3-\nu}\mathbb{I} - \frac{1-3\nu}{2(1+\nu)(3-\nu)}\mathrm{I} \otimes \mathrm{I}.$$

6 Asymptotic Analysis with Bounded Perturbations of Variational Inequalities

The bounded perturbations of bilinear forms in variational inequalities resulting from the approximation of the energy by (5) are employed in asymptotic analysis of variational inequalities in singularly perturbed geometrical domains. The proposed method of asymptotic analysis is sufficiently precise for the first order topological differentiability [56].

6.1 Applications of Steklov-Poincaré Operators in Asymptotic Analysis

We analyse the precision of the proposed method of approximation for variational inequalities in singularly perturbed geometrical domains. We assume for simplicity that the singular perturbation is a disc $B_\epsilon = \{|x| < \epsilon\}$. The Signorini variational inequality in $\Omega_\epsilon := \Omega \setminus \overline{B}_\epsilon$,

$$u_\epsilon \in K(\Omega_\epsilon) \ : \ a(\Omega_\epsilon; u_\epsilon, v - u_\epsilon) - L(\Omega_\epsilon; v - u_\epsilon) \geqslant 0 \quad \forall v \in u_\epsilon \in K(\Omega_\epsilon),$$

can be considered in the truncated domain $\Omega_c := \Omega \setminus \overline{B}_R$ for $R > \epsilon > 0$, R small enough. It is assumed that the source or linear form $v \to L(\Omega; v)$ is supported in Ω_c. Hence the restriction $u_\epsilon \in K(\Omega_c)$ of $u_\epsilon \in K(\Omega_\epsilon)$ to the truncated domain is given by the solution to variational inequality

$$u_\epsilon \in K(\Omega_c) \ : \ a(\Omega_c; u_\epsilon, v - u_\epsilon) + \langle \mathcal{A}_\epsilon(u_\epsilon), v - u_\epsilon \rangle - L(\Omega_c; v - u_\epsilon) \geqslant 0 \ \forall v \in K(\Omega_\epsilon), \quad (31)$$

where $\mathcal{A}_\epsilon$ stands for the Steklov-Poincaré operator which *replaces* the portion of bilinear form over the ring $C(R, \epsilon) := \{R > |x| > \epsilon\}$.

Proposition 33 *Assume that the Steklov-Poincaré operator admits the one-term expansion*

$$\langle \mathcal{A}_\epsilon(v), v \rangle = \langle \mathcal{A}(v), v \rangle + \epsilon^2 \langle \mathcal{B}(v), v \rangle + o(\epsilon^2; v, v)$$

with the compact remainder $o(\epsilon^2; v, v)$, then we can replace in (31) the Steklov-Poincaré operator by its one term approximation

$$\begin{aligned} &\widetilde{u}_\epsilon \in K(\Omega_c) \ : \\ &a(\Omega_c; \widetilde{u}_\epsilon, v - \widetilde{u}_\epsilon) + \langle \mathcal{A}(\widetilde{u}_\epsilon), v - \widetilde{u}_\epsilon \rangle + \\ &\qquad \epsilon^2 \langle \mathcal{B}(\widetilde{u}_\epsilon), v - \widetilde{u}_\epsilon \rangle - L(\Omega_c; v - \widetilde{u}_\epsilon) \geqslant 0 \quad \forall v \in K(\Omega_\epsilon), \end{aligned}$$

with the estimate

$$\|\widetilde{u}_\epsilon - u_\epsilon\| = o(\epsilon^2).$$

Remark 34 From Proposition 33 it follows that for the shape-topological differentiability of the energy functional we can consider the variational inequality

$$\widehat{u}_\epsilon \in K(\Omega) \ : \ a(\Omega; \widehat{u}_\epsilon, v - \widehat{u}_\epsilon) + \epsilon^2 \langle \mathcal{B}(\widehat{u}_\epsilon), v - \widehat{u}_\epsilon \rangle - L(\Omega; v - \widehat{u}_\epsilon) \geqslant 0 \quad \forall v \in K(\Omega), \tag{32}$$

since $\|\widehat{u}_\epsilon - u_\epsilon\| = o(\epsilon^2)$ in Ω_c. In this way, the approximation (5) of quadratic functional (2) is justified for the first order topological derivatives of variational inequalities in truncated domains.

For the quadratic functional (1) and the associated boundary value problem, the bilinear form

$$v \to b(\Gamma_R; v, v) := \langle \mathcal{B}(v), v \rangle$$

is determined. The linear operator $\mathcal{B}$ is obtained from the one term expansion of the Steklov-Poincaré operator $\mathcal{A}_\epsilon$, the expansion results from the energy expansion in the subdomain Ω_R. Therefore, the perturbed quadratic functional (3) can be replaced by its approximation given by (5). For the Signorini problem in two spatial dimensions it means that the variational inequality is obtained for minimization of perturbed functional (3) over the energy space in unperturbed domain Ω, and the associated energy functional

$$\mathcal{E}_\epsilon(\Omega) = \frac{1}{2} a(\Omega; u_\epsilon, u_\epsilon) + \frac{\epsilon^2}{2} b(\Gamma_R; u_\epsilon, u_\epsilon) - (f, u_\epsilon)_\Omega,$$

is evaluated for the solution of variational inequality

$$u_\epsilon \in K(\Omega) \ : \ a(\Omega; u_\epsilon, v - u_\epsilon) + \epsilon^2 b(\Gamma_R; u_\epsilon, v - u_\epsilon) - (f, v - u_\epsilon)_\Omega \geqslant 0 \quad \forall v \in K(\Omega).$$

6.2 Asymptotic Analysis by Domain Decomposition Method

In order to apply the domain decomposition technique to topological differentiability $\omega_\epsilon \to J_\epsilon(\Omega)$ in topologically perturbed domains $\Omega := \Omega_\epsilon$ for the shape functionals $\Omega \to J(\Omega)$ we need the appropriate results on topological differentiability $\epsilon \to \mathcal{B}_\epsilon$ of the Steklov-Poincaré pseudodifferential boundary operators defined on the artificial interface Σ. In the particular case of holes $\epsilon \to \omega_\epsilon$ the notation is straighforward, with the singularly perturbed domain $\Omega_\epsilon := \Omega \setminus \overline{\omega}_\epsilon$ and with the shape functional to be analysed with respect to small parameter $\epsilon \to J_\epsilon(\Omega) := J(\Omega \setminus \overline{\omega}_\epsilon)$. In the case of inclusions $\epsilon \to \omega_\epsilon$ the shape functional depends on the characteristic functions

$\epsilon \to \chi_\epsilon$ of the domain perturbation ω_ϵ. For inclusions the state solution $\epsilon \to u_\epsilon \in H(\Omega)$ is obtained by solving boundary value problems with operator coefficients depending on the small parameter $\epsilon \to 0$. In both cases the asymptotics of Steklov-Poincaré operators are obtained by asymptotic analysis of the energy functional for linear elliptic boundary value problems in subdomains Ω_2 which contains the perturbations $\epsilon \to \omega_\epsilon$. Let us consider the direct method of sensitivity analysis in subdomain Ω_1 which contains the contact subset $\Gamma_c \subset \partial\Omega$. This is possible due to the conical differentiability of metric projection onto the convex set K which is valid under some assumptions (e.g., the convex, closed cone K is polyhedric in the Dirichlet space $H(\Omega)$ [12]).

Example 35 In the case of the Signorini problem in two spatial dimensions the direct method of asymptotic analysis for the shape functional (6)

$$J_\epsilon(\Omega_\epsilon) := \int\limits_{\Omega_1} \langle A'(0) \cdot u_\epsilon, u_\epsilon \rangle dx$$

can be described as follows for the disc $\omega_\epsilon := B(\epsilon) = \{|x| < \epsilon\}$ located at the origin.

1. We solve the variational inequality in Ω_1 : determine $u \in K$ and its coincidence set $\Xi := \{x \in \Gamma_c \ : \ u(x) = 0\}$. Thus, the convex cone

$$S = \{v \in H^1_{\Gamma_0}(\Omega) \ : \ v \geqslant 0 \text{ on } \Xi \quad a(\Omega; u, v) = (f, v)_\Omega\}$$

   used in conical differentiability of the element u with respect to the shape can be determined.
2. The asymptotic analysis of solutions to variational inequality in singularly perturbed domain $\Omega(\epsilon) : \Omega \setminus \overline{B(\epsilon)}$ with respect to small parameter $\epsilon \to 0$ which governs the size of the hole $B(\epsilon)$ leads to the expansion

$$u_\epsilon = u + \epsilon^2 q + o(\epsilon^2)$$

   obtained by the domain decomposition method with the Steklov-Poincaré boundary operators, where

$$q \in S \ : \ a(\Omega; q, v - q) + \epsilon^2 \langle \mathcal{B} q, v - q \rangle_R \geqslant \quad \forall v \in S.$$

3. The shape functional

$$J_\epsilon(\Omega_\epsilon) := \int\limits_{\Omega_1} \langle A'(0) \cdot u_\epsilon, u_\epsilon \rangle dx$$

   can be expanded in Ω_1, the expansion is valid in the whole domain Ω,

$$J_\epsilon(\Omega_\epsilon) = \int_\Omega \langle A'(0)\cdot u, u\rangle dx + 2\epsilon^2 \int_\Omega \langle A'(0)\cdot q, u\rangle dx + o(\epsilon^2),$$

however the obtained expression for the topological derivative may not be constructive in numerical methods. We want to obtain an equivalent expression, when possible, which replaces the topological derivative

$$\mathcal{T}(\mathcal{O}) = 2\int_\Omega \langle A'(0)\cdot q, u\rangle dx$$

in the first order expansion of the energy functional for Signorini problem. In the linear boundary value problems such an expression can always be obtained by the introduction of an appropriate adjoint state. We point out that for variational inequalities the existence of an adjoint state in general cannot be expected.

7 Asymptotic Analysis of Boundary Value Problems in Rings or Spherical Shells

In this section we shall consider asymptotic corrections to the energy functional for the elasticity boundary value problems or the Laplace equation in $\mathbb{R}^d$, where $d = 2, 3$. The dependence of the energy on small parameter is caused by creating a small ball-like void of variable radius ϵ in the interior of the domain Ω, with the homogeneous Neumann boundary conditions for the boundary value problems on its surface. We assume that this void has its centre at the origin $\mathcal{O}$. In order to eliminate the variability of the domain, we take as Ω_R the open ball $B(\mathcal{O}, R) = B(R)$ with fixed R. In this way the void $B(\epsilon)$ is surrounded by $B(R) \subset \text{int}\,\Omega$. We denote also the ring or spherical shell as $C(R, \epsilon) = B(R) \setminus \overline{B}(\epsilon)$, $\Omega(R) = \Omega \setminus \overline{B}(R)$ and $\Gamma_R = \partial B(R)$. Using these notations we define our main tool, namely the Dirichlet-to-Neumann mapping for linear elasticity or the Steklov-Poincaré operator

$$\mathcal{A}_\epsilon : \mathbf{H}^{1/2}(\Gamma_R) \longmapsto \mathbf{H}^{-1/2}(\Gamma_R)$$

by means of the boundary value problem:

$$\begin{aligned}
&(1-2\nu)\Delta\mathbf{w} + \mathbf{grad}\,\mathrm{div}\,\mathbf{w} = 0, \quad \text{in } C(R,\epsilon),\\
&\mathbf{w} = \mathbf{v} \ \text{ on } \Gamma_R,\\
&\boldsymbol{\sigma}(\mathbf{w}).\mathbf{n} = 0 \ \text{ on } \Gamma_\epsilon
\end{aligned}$$

so that

$$\mathcal{A}_\epsilon \mathbf{v} = \boldsymbol{\sigma}(\mathbf{w}).\mathbf{n} \ \text{ on } \Gamma_R.$$

Domain decomposition—Steklov-Poincaré operator. Let $\mathbf{u}^R$ be the restriction of $\mathbf{u}$ to $\Omega(R)$ and $\gamma^R\varphi$ the projection of φ on Γ_R. We may then define the functional

$$I_\epsilon^R(\varphi_\epsilon) = \frac{1}{2}\int_{\Omega(R)} \boldsymbol{\sigma}(\varphi_\epsilon) : \varepsilon(\varphi_\epsilon)\, dx - \int_{\Gamma_N} \mathbf{h}.\varphi_\epsilon\, ds + \frac{1}{2}\int_{\Gamma_R} (\mathcal{A}_\epsilon \gamma^R \varphi_\epsilon).\gamma^R\varphi_\epsilon\, ds$$

and the solution $\mathbf{u}_\epsilon^R$ as a minimal argument for

$$I_\epsilon^R(\mathbf{u}_\epsilon^R) = \inf_{\varphi_\epsilon \in K \subset V_\epsilon} I_\epsilon^R(\varphi_\epsilon),$$

Here lies the essence of the domain decomposition concept: we have replaced the the variable domain by a fixed one, at the price of introducing variable boundary operator $\mathcal{A}_\epsilon$. The above expressions have even simpler form in case of a single Laplace equation. It is enough to replace the displacement by the scalar function u, elasticity operator by $-\Delta$, and

$$\boldsymbol{\sigma}(\mathbf{u}) := \mathbf{grad}\, u, \qquad \varepsilon(\mathbf{u}) := \mathbf{grad}\, u, \qquad \boldsymbol{\sigma}(\mathbf{u}).\mathbf{n} := \partial u/\partial \mathbf{n}.$$

The goal is to find the expansion

$$\mathcal{A}_\epsilon = \mathcal{A} + \epsilon^d \mathcal{B} + \mathcal{R}_\epsilon, \tag{33}$$

where the remainder $\mathcal{R}_\epsilon$ is of order $o(\epsilon^d)$ in the operator norm in the space $L(\mathbf{H}^{1/2}(\Gamma_R), \mathbf{H}^{-1/2}(\Gamma_R))$, and the operator $\mathcal{B}$ is regular enough, namely it is bounded and linear:

$$\mathcal{B} \in L\big(\mathbf{L}_2(\Gamma_R), \mathbf{L}_2(\Gamma_R)\big).$$

Under this assumption the following propositions hold.

Proposition 36 *Assume that (33) holds in the operator norm. Then strong convergence takes place*

$$\mathbf{u}_\epsilon^R \to \mathbf{u}^R$$

in the norm of $\mathbf{H}^1(\Omega(R))$.

Proposition 37 *The energy functional has the representation*

$$I_\epsilon^R(\mathbf{u}_\epsilon^R) = I^R(\mathbf{u}^R) + \epsilon^d \langle \mathcal{B}(\mathbf{u}^R), \mathbf{u}^R\rangle_R + o(\epsilon^3),$$

where $o(\epsilon^d)/\epsilon^d \to 0$ *with* $\epsilon \to 0$ *in the same energy norm.*

Here $I^R(\mathbf{u}^R)$ denotes the functional I_ϵ^R on the intact domain, i.e. $\epsilon := 0$ and $\mathcal{A}_\epsilon := \mathcal{A}$, applied to truncation of $\mathbf{u}$. Generally, the energy correction for both the elasticity system and the Laplace operator has the form

$$\langle \mathcal{B}(\mathbf{u}^R), \mathbf{u}^R \rangle_R \;=\; -c_d e_u(\mathcal{O}),$$

where $c_d = \mathrm{vol}(B(1))$. The energy-like density function $e_u(\mathcal{O})$ has the form:

- In case of the Laplace operator

$$e_u(\mathcal{O}) \;=\; \frac{1}{2}\|\nabla u^R(\mathcal{O})\|^2$$

  for both $d = 2$ and $d = 3$, see [56].
- In case of the elasticity system

$$e_u(\mathcal{O}) \;=\; \frac{1}{2}\mathbb{P}\boldsymbol{\sigma}(\mathbf{u}^R(\mathcal{O}) : \boldsymbol{\varepsilon}(\mathbf{u}^R(\mathcal{O}),$$

  where for $d = 2$ and plain stress

$$\mathbb{P} \;=\; \frac{1}{1-\nu}(4\mathbb{I} - \mathrm{I}\otimes\mathrm{I})$$

  and for $d = 3$

$$\mathbb{P} \;=\; \frac{1-\nu}{7-5\nu}\left(10\mathbb{I} - \frac{1-5\nu}{1-2\nu}\mathrm{I}\otimes\mathrm{I}\right)$$

  see [46, 55]. Here $\mathbb{I}$ is the fourth order identity tensor, and I is the second order identity tensor.

This approach is important for variational inequalities since it allows us to derive the formulas for topological derivatives which are similar to the expressions obtained for the corresponding linear boundary value problems.

Explicit form of the operator $\mathcal{B}$—the Laplace operator in two spatial dimensions. If the function u is harmonic in a ball $B(R) \subset \mathbb{R}^2$, of radius $R > 0$ and centre at $\mathbf{x}_0 = \mathcal{O}$, then the exact expressions for the first order derivatives of u take on the following form [56]

$$u_{/1}(\mathcal{O}) = \frac{1}{\pi R^3}\int_{\Gamma_R} u \cdot x_1\, ds,$$
$$u_{/2}(\mathcal{O}) = \frac{1}{\pi R^3}\int_{\Gamma_R} u \cdot x_2\, ds.$$

Since the line integrals on Γ_R are well defined for functions in $L_2(\Gamma_R)$, it follows that the operator $\mathcal{B}$ can be extended to a bounded operator on $L_2(\Gamma_R)$,

$$\mathcal{B} \in \mathcal{L}(L_2(\Gamma_R) \to L_2(\Gamma_R)).$$

The symmetric bilinear form for this operator, given by

$$\langle \mathcal{B}u, v\rangle_R = -\tfrac{1}{2\pi R^6}\left[\left(\int_{\Gamma_R} ux_1\,ds\right)\left(\int_{\Gamma_R} vx_1\,ds\right) + \left(\int_{\Gamma_R} ux_2\,ds\right)\left(\int_{\Gamma_R} vx_2\,ds\right)\right],$$

is continuous for all $u, v \in L_2(\Gamma_R)$. In fact, the bilinear form

$$L_2(\Gamma_R) \times L_2(\Gamma_R) \ni (u, v) \mapsto b(\Gamma_R; u, v) \in \mathbb{R}$$

is continuous with respect to the weak convergence because of the simple structure

$$b(\Gamma_R; u, v) = l_1(u)l_1(v) + l_2(u)l_2(v) \quad u, v \in L_1(\Gamma_R)$$

with two linear forms $v \to l_i(v), i = 1, 2,$

$$l_i(u) \;=\; \frac{1}{\sqrt{2\pi}} R^{-3} \int_{\Gamma_R} ux_i\,ds$$

defined as line integrals on Γ_R. This gives an additional regularity for the regular non-local perturbation $\mathcal{B}$ of the pseudo-differential Steklov-Poincaré boundary operator $\mathcal{A}_\epsilon$.

Explicit form of the operator $\mathcal{B}$—the Laplace operator in three spatial dimensions. Similarly as in two spatial dimensions, for harmonic functions in $\mathbb{R}^3$ it may be proved [56] that

$$u_{/1}(\mathcal{O}) = \frac{3}{4\pi R^4}\int_{S(R)} ux_1\,ds,$$
$$u_{/2}(\mathcal{O}) = \frac{3}{4\pi R^4}\int_{S(R)} ux_2\,ds,$$
$$u_{/3}(\mathcal{O}) = \frac{3}{4\pi R^4}\int_{S(R)} ux_3\,ds.$$

Using this one can easily write down the bilinear form

$$b(\Gamma_R; u, v) = \langle \mathcal{B}u, v\rangle_R = l_1(u)l_1(v) + l_2(u)l_2(v) + l_3(u)l_3(v)$$

where

$$l_i(u, v) \;=\; \sqrt{\frac{3}{8\pi}} R^{-4} \int_{S(R)} ux_i\,ds.$$

From the computational point of view, the effort in comparison to two spatial dimensions grows similarly as the difficulty of computing integrals over circle versus integrals over sphere.

Explicit form of the operator $\mathcal{B}$—elasticity in two spatial dimensions. Let us denote for the plain stress case

$$k = \frac{\lambda + \mu}{\lambda + 3\mu}.$$

It has been proved in [56] that the following exact formulae hold

$$\begin{aligned}
\varepsilon_{11}(\mathcal{O}) + \varepsilon_{22}(\mathcal{O}) &= \frac{1}{\pi R^3} \int_{\Gamma_R} (u_1 x_1 + u_2 x_2)\, ds, \\
\varepsilon_{11}(\mathcal{O}) - \varepsilon_{22}(\mathcal{O}) &= \frac{1}{\pi R^3} \int_{\Gamma_R} \left[(1 - 9k)(u_1 x_1 - u_2 x_2) + \frac{12k}{R^2}(u_1 x_1^3 - u_2 x_2^3) \right] ds, \\
2\varepsilon_{12}(\mathcal{O}) &= \frac{1}{\pi R^3} \int_{\Gamma_R} \left[(1 + 9k)(u_1 x_2 + u_2 x_1) - \frac{12k}{R^2}(u_1 x_2^3 + u_2 x_1^3) \right] ds.
\end{aligned}$$

These expressions are easy to compute numerically, but contain additional integrals of third powers of x_i. Therefore, strains $\varepsilon_{ij}(\mathcal{O})$ may be expressed as linear combinations of integrals over circle which have the form

$$\int_{\Gamma_R} u_i x_j \, ds, \qquad \int_{\Gamma_R} u_i x_j^3 \, ds.$$

The same is true, due to Hooke's law, for stresses $\sigma_{ij}(\mathcal{O})$. They may then be substituted into expression for the operator B, yielding

$$\langle \mathcal{B}(\mathbf{u}^R), \mathbf{v}^R \rangle_R = -\frac{1}{2} c_2 \mathbb{P}\boldsymbol{\sigma}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}).$$

These formulas are quite similar to the ones obtained for Laplace operator and easy to compute numerically.

Explicit form of the operator $\mathcal{B}$ for elasticity in three spatial dimensions. It turns out that similar situation holds in three spatial dimensions, but obtaining the formulas is more difficult. Assuming given values of $\mathbf{u}$ on Γ_R, the solution of elasticity system in $B(R)$ may be expressed as

$$\mathbf{u} = \sum_{n=0}^{\infty} [\mathbf{U}_n + (R^2 - r^2) k_n(\nu) \mathbf{grad}\, \mathrm{div}\, \mathbf{U}_n].$$

where $k_n(\nu) = 1/2[(3 - 2\nu)n - 2(1 - \nu)]$ and $r = \|\mathbf{x}\|$. In addition

$$\mathbf{U}_n = \frac{1}{R^n} [\mathbf{a}_{n0} d_n(\mathbf{x}) + \sum_{m=1}^{n} (\mathbf{a}_{nm} c_n^m(\mathbf{x}) + \mathbf{b}_{nm} s_n^m(\mathbf{x}))].$$

The vectors

$$\mathbf{a}_{n0} = (a^1_{n0}, a^2_{n0}, a^3_{n0})^\top,$$

$$\mathbf{a}_{nm} = (a^1_{nm}, a^2_{nm}, a^3_{nm})^\top,$$

$$\mathbf{b}_{nm} = (b^1_{nm}, b^2_{nm}, b^3_{nm})^\top$$

are constant and the set of functions

$$\{d_0;\ d_1, c^1_1, s^1_1;\ d_2, c^1_2, s^1_2, c^2_2, s^2_2;\ d_3, c^1_3, s^1_3, c^2_3, s^2_3, c^3_3, s^3_3;\ \ldots\}$$

constitutes the complete system of orthonormal harmonic polynomials on Γ_R, related to Laplace spherical functions, see next paragraph. Specifically,

$$c^l_k(\mathbf{x}) = \frac{\hat{P}^{l,c}_k(\mathbf{x})}{\|\hat{P}^{l,c}_k\|_R}, \quad s^l_k(\mathbf{x}) = \frac{\hat{P}^{l,s}_k(\mathbf{x})}{\|\hat{P}^{l,s}_k\|_R}, \quad d_k = \frac{P_k((\mathbf{x})}{\|\hat{P}_k\|_R}.$$

For example,

$$c^2_3(\mathbf{x}) = \frac{1}{R^4}\sqrt{\frac{7}{240\pi}}(15x^2_1x_3 - 15x^2_2x_3),$$

If the value of $\mathbf{u}$ on Γ_R is assumed as given, then, denoting

$$\langle\phi, \psi\rangle_R = \int_{\Gamma_R} \phi\psi\, ds,$$

we have for $n \geqslant 0$, $m = 1..n$, $i = 1, 2, 3$:

$$\begin{aligned} a^i_{n0} &= R^n\langle u_i, d_n(\mathbf{x})\rangle_R, \\ a^i_{nm} &= R^n\langle u_i, c^m_n(\mathbf{x})\rangle_R, \\ b^i_{nm} &= R^n\langle u_i, s^m_n(\mathbf{x})\rangle_R. \end{aligned} \tag{34}$$

Since we are looking for $\varepsilon_{ij}(\mathcal{O})$, only the part of $\mathbf{u}$ which is linear in $\mathbf{x}$ is relevant. It contains two terms:

$$\hat{\mathbf{u}} = \mathbf{U}_1 + R^2k_3(\nu)\mathbf{grad}\,\mathrm{div}\,\mathbf{U}_3.$$

For any $f(\mathbf{x})$, $\mathbf{grad}\,\mathrm{div}\,(\mathbf{a}f) = H(f)\cdot\mathbf{a}$, where $\mathbf{a}$ – constant vector and $H(f)$ is the Hessian matrix of f. Therefore

$$\hat{\mathbf{u}} = \frac{1}{R}[\mathbf{a}_{10}d_1(\mathbf{x}) + \mathbf{a}_{11}c_1^1(\mathbf{x}) + \mathbf{b}_{11}s_1^1(\mathbf{x}))] + R^2k_3(\nu)\frac{1}{R^3}\left[H(d_3)(\mathbf{x})\mathbf{a}_{30} + \sum_{m=1}^{3}\left(H(c_3^m)(\mathbf{x})\mathbf{a}_{3m} + H(s_3^m)(\mathbf{x})\mathbf{b}_{3m}\right)\right]$$

From the above we may single out the coefficients standing at x_1, x_2, x_3 in u_1, u_2, u_3. For example,

$$\varepsilon_{11}(\mathcal{O}) = \frac{1}{R^3}\sqrt{\frac{3}{4\pi}}a_{11}^1 + \frac{1}{R^5}k_3(\nu)\left[-3\sqrt{\frac{7}{4\pi}}a_{30}^3 - 9\sqrt{\frac{7}{24\pi}}a_{31}^1 - 3\sqrt{\frac{7}{24\pi}}b_{31}^2 + 30\sqrt{\frac{7}{240\pi}}a_{32}^3 + 90\sqrt{\frac{7}{1440\pi}}a_{33}^1 + 90\sqrt{\frac{7}{1440\pi}}b_{33}^2\right],$$

$$\varepsilon_{12}(\mathcal{O}) = \frac{1}{R^3}\sqrt{\frac{3}{4\pi}}(b_{11}^1 + a_{11}^2) + \frac{1}{R^5}k_3(\nu)\left[-3\sqrt{\frac{7}{24\pi}}a_{31}^2 - \sqrt{\frac{7}{24\pi}}b_{31}^1 + 15\sqrt{\frac{7}{60\pi}}b_{32}^3 - 90\sqrt{\frac{7}{1440\pi}}a_{33}^2 + 90\sqrt{\frac{7}{1440\pi}}b_{33}^1\right].$$

Observe that

$$\varepsilon_{11}(\mathcal{O}) + \varepsilon_{22}(\mathcal{O}) + \varepsilon_{33}(\mathcal{O}) = \frac{1}{R^3}\sqrt{\frac{3}{4\pi}}\left(R\langle u_1, c_1^1\rangle_R + R\langle u_2, s_1^1\rangle_R + R\langle u_3, d_1\rangle_R\right)$$

and $c_1^1 = \frac{1}{R^2}\sqrt{\frac{3}{4\pi}}x_1$, $s_1^1 = \frac{1}{R^2}\sqrt{\frac{3}{4\pi}}x_2$, $d_1 = \frac{1}{R^2}\sqrt{\frac{3}{4\pi}}x_3$, exactly the same as for the case of Laplace equation. This should be expected, since $\operatorname{tr}\varepsilon$ is a harmonic function.

As a result, the operator $\mathbf{B}$ may be defined by the formula

$$\langle \mathcal{B}\mathbf{u}, \mathbf{u}\rangle_R = -c_3\mathbb{P}\boldsymbol{\sigma}(\mathbf{u}((\mathcal{O}))) : \varepsilon(\mathbf{u}((\mathcal{O})))$$

but the right-hand side consists of integrals of $\mathbf{u}$ multiplied by first and third order polynomials in x_i over Γ_R resulting from (34). This is a very similar situation as in two spatial dimensions. Thus, the new expressions for strains make possible to rewrite $\mathcal{B}$ in the form possessing the desired regularity.

Laplace spherical polynomials. For $n = 1$:

$$\hat{P}_1(\mathbf{x}) = x_3, \quad \hat{P}_1^{1,c}(\mathbf{x}) = x_1, \quad \hat{P}_1^{1,s}(\mathbf{x}) = x_2,$$
$$\|\hat{P}_1\|_R = \|\hat{P}_1^{1,c}\|_R = \|\hat{P}_1^{1,s}\|_R = R^2\sqrt{\frac{4\pi}{3}},$$

and for $n = 3$:

$$\hat{P}_3(\mathbf{x}) = x_3^3 - \frac{3}{2}x_2^2 x_3 - \frac{3}{2}x_1^2 x_3, \qquad \|\hat{P}_3\|_R = R^4\sqrt{\frac{4\pi}{7}},$$

$$\hat{P}_3^{1,c}(\mathbf{x}) = 6x_1 x_3^2 - \frac{3}{2}x_1^3 - \frac{3}{2}x_1 x_2^2, \qquad \|\hat{P}_3^{1,c}\|_R = R^4\sqrt{\frac{24\pi}{7}},$$

$$\hat{P}_3^{1,s}(\mathbf{x}) = 6x_2 x_3^2 - \frac{3}{2}x_2^3 - \frac{3}{2}x_1^2 x_2, \qquad \|\hat{P}_3^{1,s}\|_R = R^4\sqrt{\frac{24\pi}{7}},$$

$$\hat{P}_3^{2,c}(\mathbf{x}) = 15x_1^2 x_3 - 15x_2^2 x_3, \qquad \|\hat{P}_3^{2,c}\|_R = R^4\sqrt{\frac{240\pi}{7}},$$

$$\hat{P}_3^{2,s}(\mathbf{x}) = 15x_1 x_2 x_3, \qquad \|\hat{P}_3^{2,s}\|_R = R^4\sqrt{\frac{60\pi}{7}},$$

$$\hat{P}_3^{3,c}(\mathbf{x}) = 15x_1^3 - 45x_1 x_2^2, \qquad \|\hat{P}_3^{3,c}\|_R = R^4\sqrt{\frac{1440\pi}{7}},$$

$$\hat{P}_3^{3,s}(\mathbf{x}) = 45x_1^2 x_2 - 15x_2^3, \qquad \|\hat{P}_3^{3,s}\|_R = R^4\sqrt{\frac{1440\pi}{7}},$$

8 Asymptotic Analysis of Steklov-Poincaré Operators in Reinforced Rings in Two Spatial Dimensions

In this section the similar asymptotic analysis of elliptic boundary value problems in subdomain $\Omega_R \in \mathbb{R}^2$ is performed, but we modify the situation, assuming that the hole is filled only partially, different material constituting a fixed part of it. In this way, we may consider double asymptotic transition, where both the size of the hole, as well as the proportion of the different material contained in it can vary. Mechanically this situation corresponds e.g. to the hole with hardened walls.

The analysis is based again on exact representation of solutions and allows to obtain the perturbation of solutions, using the fact that these solutions may be considered as minimizers of energy functional. The method is also suitable for double asymptotic expansions of solutions as well as energy form. The ultimate goal is to use obtained formulas in the evaluation of topological derivatives for elliptic boundary value problems.

8.1 Model Problem

Let us consider the the domain Ω containing the hole with boundary made of modified material. For simplicity the hole is located at the origin of the coordinate system. In

order to write down the model problem, we introduce some notations.

$$\begin{aligned} B_s &= \{\, x \in \mathbb{R}^2 \mid \|x\| < s \,\} \\ C_{s,t} &= \{\, x \in \mathbb{R}^2 \mid s < \|x\| < t \,\} \\ \Gamma_s &= \{\, x \in \mathbb{R}^2 \mid \|x\| = s \,\} \\ \Omega_s &= \Omega \setminus B_s \end{aligned}$$

Then the problem in the intact domain Ω has the form

$$\begin{aligned} k_1 \Delta w_0 &= 0 \quad \text{in} \quad \Omega \\ w_0 &= g_0 \quad \text{on} \quad \partial\Omega \end{aligned} \tag{35}$$

The model problem in the modified domain reads:

$$\begin{aligned} k_1 \Delta w_\rho &= 0 \quad \text{in} \quad \Omega_\rho \\ w_\rho &= g_0 \quad \text{on} \quad \partial\Omega \\ w_\rho &= v_\rho \quad \text{on} \quad \Gamma_\rho \\ k_2 \Delta v_\rho &= 0 \quad \text{in} \quad C_{\lambda\rho,\rho} \end{aligned} \tag{36}$$

$$\begin{aligned} k_2 \frac{\partial v_\rho}{\partial n_2} &= 0 \quad \text{on} \quad \Gamma_{\lambda\rho} \\ k_1 \frac{\partial w_\rho}{\partial n_1} + k_2 \frac{\partial v_\rho}{\partial n_2} &= 0 \quad \text{on} \quad \Gamma_\rho, \end{aligned}$$

where n_1—exterior normal vector to Ω_ρ, n_2—exterior normal vector to $C_{\lambda\rho,\rho}$, and $0 < \lambda < 1$. We want to investigate the influence of the small ring-like inclusion made of another material on the difference $w_\rho - w_0$ in Ω_R, where Γ_R surrounds $C_{\lambda\rho,\rho}$ and R is fixed. We assume that $\rho \to 0+$ and λ is considered temporarily constant.

If we define

$$u_\rho = \begin{cases} w_\rho & \text{in} \quad \Omega_\rho \\ v_\rho & \text{in} \quad C_{\lambda\rho,\rho} \end{cases}$$

then the problem (36) reduces to finding minimum of the energy functional

$$\mathcal{E}_1(u_\rho) = \frac{1}{2} \int_{\Omega_\rho} k_1 \nabla u_\rho \cdot \nabla u_\rho \, dx + \frac{1}{2} \int_{C_{\lambda\rho,\rho}} k_2 \nabla u_\rho \cdot \nabla u_\rho \, dx$$

for $u_\rho \in H^1(\Omega_\rho)$, $u_\rho = g_0$ on $\partial\Omega$. This expression may be rewritten as

$$\begin{aligned}\mathcal{E}_1(u_\rho) = &\frac{1}{2}\int_{\Omega_R} k_1 \nabla w_\rho \cdot \nabla w_\rho \, dx \\ &+ \frac{1}{2}\int_{C_{\rho,R}} k_1 \nabla w_\rho \cdot \nabla w_\rho \, dx \\ &+ \frac{1}{2}\int_{C_{\lambda\rho,\rho}} k_2 \nabla v_\rho \cdot \nabla v_\rho \, dx.\end{aligned}$$

Using integration by parts we obtain

$$\begin{aligned}\mathcal{E}_1(u_\rho) = &\frac{1}{2}\int_{\Omega_R} k_1 \nabla w_\rho \cdot \nabla w_\rho \, dx \\ &+ \frac{1}{2}\int_{\Gamma\rho} \left(w_\rho k_1 \frac{\partial w_\rho}{\partial n_1} + v_\rho k_2 \frac{\partial v_\rho}{\partial n_2} \right) ds \\ &+ \frac{1}{2}\int_{\Gamma_R} k_1 w_\rho \frac{\partial w_\rho}{\partial n_3} \, ds,\end{aligned}$$

where n_3—exterior normal to Ω_R. Hence, due to boundary and transmission condition,

$$\mathcal{E}_1(u_\rho) = \frac{1}{2}\int_{\Omega_R} k_1 \nabla w_\rho \cdot \nabla w_\rho \, dx + \frac{1}{2}\int_{\Gamma_R} k_1 w_\rho \frac{\partial w_\rho}{\partial n_3} \, ds \tag{37}$$

8.2 Steklov-Poincaré Operator

Observe that $\mathcal{E}_1(w_0)$ corresponds to the problem (35). Therefore the main goal is to find the Steklov-Poincaré operator

$$\mathcal{A}_{\lambda,\rho} \; : \; w \in H^{1/2}(\Gamma_R) \; \longmapsto \; \frac{\partial w_\rho}{\partial n_3} \in H^{-1/2}(\Gamma_R)$$

where the normal derivative is computed from auxiliary problem

$$\begin{aligned} k_1 \Delta w_\rho &= 0 \quad \text{in} \quad C_{\rho,R} \\ w_\rho &= w \quad \text{on} \quad \Gamma_R \\ w_\rho &= v_\rho \quad \text{on} \quad \Gamma_\rho \\ k_2 \Delta v_\rho &= 0 \quad \text{in} \quad C_{\lambda\rho,\rho} \\ k_2 \frac{\partial v_\rho}{\partial n_2} &= 0 \quad \text{on} \quad \Gamma_{\lambda\rho} \\ k_1 \frac{\partial w_\rho}{\partial n_1} + k_2 \frac{\partial v_\rho}{\partial n_2} &= 0 \quad \text{on} \quad \Gamma_\rho \end{aligned}$$

The geometry of domains of definition for functions is shown in Fig. 2. Now let us adopt the polar coordinate system around origin and assume the Fourier series form for w on Γ_R.

$$w = C_0 + \sum_{k=1}^{\infty} (A_k \cos k\varphi + B_k \sin k\varphi)$$

The general form of the solution w_ρ is

$$w_\rho = A^w + B^w \log r + \sum_{k=1}^{\infty} \left(w_k^c(r) \cos k\varphi + w_k^s(r) \sin k\varphi \right),$$

where

$$w_k^c(r) = A_k^c r^k + B_k^c \frac{1}{r^k}, \quad w_k^s(r) = A_k^s r^k + B_k^s \frac{1}{r^k}.$$

Similarly for v_ρ:

$$v_\rho = A^v + B^v \log r + \sum_{k=1}^{\infty} \left(v_k^c(r) \cos k\varphi + v_k^s(r) \sin k\varphi \right),$$

where

$$v_k^c(r) = a_k^c r^k + b_k^c \frac{1}{r^k}, \quad v_k^s(r) = a_k^s r^k + b_k^s \frac{1}{r^k}.$$

Additionally, we denote the Fourier expansion of v_ρ on Γ_ρ by

$$v_\rho = c_0 + \sum_{k=1}^{\infty} (a_k \cos k\varphi + b_k \sin k\varphi)$$

From boundary conditions on $\Gamma_{\lambda\rho}$ it follows easily $B^v = 0$, $A^v = c_0$, and then $B^w = 0$, $A^w = A^v = c_0 = C_0$. There remains to find a_k, b_k, a_k^c, b_k^c, a_k^s, b_k^s, A_k^c, B_k^c, A_k^s, B_k^s assuming A_k, B_k as given.

8.3 Asymptotic Expansion

In order to eliminate the above mentioned coefficients we consider first the terms at $\cos k\varphi$. From boundary and transmission conditions we have for $k = 1, 2, \ldots$

$$\begin{aligned} A_k^c R^k + B_k^c \frac{1}{R^k} &= A_k \\ A_k^c \rho^k + B_k^c \frac{1}{\rho^k} - a_k &= 0 \\ a_k^c \rho^k + b_k^c \frac{1}{\rho^k} - a_k &= 0 \\ a_k^c (\lambda\rho)^{k-1} - b_k^c \frac{1}{(\lambda\rho)^{k+1}} &= 0 \\ k_1 A_k^c \rho^{k-1} - k_1 B_k^c \frac{1}{\rho^{k+1}} - k_2 a_k^c \rho^{k-1} + k_2 b_k^c \frac{1}{\rho^{k+1}} &= 0 \end{aligned}$$

This may be rewritten in the matrix form: grouping unknown parameters into a vector $\mathbf{p}_k = [A_k^c, B_k^c, a_k^c, b_k^c, a_k]^\top$ we obtain

$$T(k_1, k_2, R, \lambda, \rho)\mathbf{p}_k = R^k A_k \mathbf{e}_1$$

where

$$T = \begin{bmatrix} R^{2k} & 1 & 0 & 0 & 0 \\ \rho^{2k} & 1 & 0 & 0 & -\rho^k \\ 0 & 0 & (\lambda\rho)^{2k} & 1 & -\rho^k \\ 0 & 0 & (\lambda\rho)^{2k} & -1 & 0 \\ k_1\rho^{2k} & -k_1 & -k_2\rho^{2k} & k_2 & 0 \end{bmatrix}$$

where $\mathbf{e}_1 = [0, 0, 0, 0, 1]^\top$. It is easy to see that

$$\mathbf{p}_k = \mathbf{p}_k^0 A_k + \rho^{2k}\mathbf{p}_k^1 A_k + o(\rho^{2k})$$

where

$$\mathbf{p}_k^0 = \lim_{\rho\to 0+}\lim_{\lambda\to 0+} \frac{\mathbf{p}_k(k_1, k_2, R, \lambda, \rho)}{A_k}$$

and $\mathbf{p}_k^0 = [1/R^k, 0, 0, 0, 0]^\top$, which corresponds to the ball B_R filled completely with material k_1. Similar reasoning may be conducted for terms containing $\sin k\varphi$.

As a result,

$$\mathcal{A}_{\lambda,\rho} = \mathcal{A}_{0,0} + \rho^2 \mathcal{A}^1_{\lambda,\rho}(k_1, k_2, R, \lambda, \rho, A_1, B_1) + o(\rho^2).$$

The exact form of $\mathcal{A}^1_{\lambda,\rho}(k_1, k_2, R, \lambda, \rho, A_1, B_1)$ is obtained from inversion of matrix T, but, what is crucial, it is linear in both A_1 and B_1. They in turn are computed as line integrals

$$A_1(w) = \frac{1}{\pi R^2}\int_{\Gamma_R} wx_1\,ds, \quad B_1(w) = \frac{1}{\pi R^2}\int_{\Gamma_R} wx_2\,ds.$$

As a result, for computing u_ρ we may use the following energy form

$$\mathcal{E}(u_\rho) = \frac{1}{2}\int_\Omega k_1 \nabla u_\rho \cdot \nabla u_\rho\,dx + \rho^2 Q(k_1, k_2, R, \lambda, \rho, A_1, B_1) + o(\rho^2),$$

where $A_1 = A_1(u_\rho)$, $B_1 = B_1(u_\rho)$ and Q is a quadratic function of A_1, B_1. This constitutes a regular perturbation of the energy functional which allows computing perturbations of any functional depending on this solution and caused by small inclusion of the described above form.

8.4 Extension to Linear Elasticity

Let us consider the plane elasticity problem in the ring $C_{R,\rho}$. We use polar coordinates (r, θ) with $\mathbf{e}_r$ pointing outwards and $\mathbf{e}_\theta$ perpendicularly in the counter-clockwise direction. Then there exists an exact representation of both solutions, using the complex variable series. It has the form [20, 38]

$$\begin{aligned} \sigma_{rr} - i\sigma_{r\theta} &= 2\Re\phi' - e^{2i\theta}(\bar{z}\phi'' + \psi') \\ \sigma_{rr} + i\sigma_{\theta\theta} &= 4\Re\phi' \\ 2\mu(u_r + iu_\theta) &= e^{-i\theta}(\kappa\phi - z\bar{\phi}' - \bar{\psi}). \end{aligned} \tag{38}$$

The functions $\phi,\ \psi$ are given by complex series

$$\begin{aligned} \phi &= A\log(z) + \sum_{k=-\infty}^{k=+\infty} a_k z^k \\ \psi &= -\kappa\bar{A}\log(z) + \sum_{k=-\infty}^{k=+\infty} b_k z^k. \end{aligned} \tag{39}$$

Here μ—the Lame constant, ν—the Poisson ratio, $\kappa = 3 - 4\nu$ in the plain strain case, and $\kappa = (3-\nu)/(1+\nu)$ for plane stress.

Similarly as in the simple case described in former sections, the displacement data may be given in the form of Fourier series,

$$2\mu(u_r + iu_\theta) = \sum_{k=-\infty}^{k=+\infty} A_k e^{ik\theta}$$

The traction-free condition on some circle means $\sigma_{rr} = \sigma_{r\theta} = 0$. From (38) and (39) we get for displacements the formula

$$2\mu(u_r + iu_\theta) = 2\kappa A r \log(r)\frac{1}{z} - \bar{A}\frac{1}{r}z + \sum_{p=-\infty}^{p=+\infty}\left[\kappa r a_{p+1} - (1-p)\bar{a}_{1-p}r^{-2p+1} - \bar{b}_{-(p+1)}r^{-2p-1}\right]z^p.$$

Similarly we obtain representation of tractions on some circle

$$\sigma_{rr} - i\sigma_{r\theta} = 2A\frac{1}{z} + (\kappa+1)\frac{1}{r^2}\bar{A}z + \sum_{p=-\infty}^{p=+\infty}(1-p)\left[(1+p)a_{p+1} + \bar{a}_{1-p}r^{-2p} + \frac{1}{r^2}b_{p-1}\right]z^p.$$

As we see, in principle it is possible to repeat the same procedure again, glueing solutions in two rings together and eliminating the intermediary Dirichlet data on the interface. The only difference lies in considerably more complicated calculations, see e.g. [13]. This could be applied for making double asymptotic expansion, in term of both ρ and λ. However, in our case λ does not need to be small in comparison to ρ.

8.5 Summary of Results for Particular Cases

The explicit form of solutions in B_R allows us to conclude that for

$$\|w_\rho\|_{H^{1/2}(\Gamma_R)} \leqslant \Lambda_0$$

the correction to the energy functional contains part proportional to ρ^d and the remainder of order $o(\rho^d)$. This in turn [56, 58] implies the possibility of representation

$$w_\rho = w_0 + \rho^2 q + o(\rho^2) \quad \text{in} \quad H^1(\Omega_R)$$

for both standard and contact problems, justifying computations of topological derivatives. It is well known that the singularities of solutions to Partial Differential Equations due to the singularities of geometrical domains can be characterized by specific shape derivatives of the associated energy shape functionals [12]. Therefore, the influence of topological changes in domains on the singularities can be measured by the appropriate second-order topological derivatives of the energy functionals. It means that we evaluate the shape derivatives of the energy functional by using the

velocity field method, and subsequently the second order topological derivatives of the functionals by an application of the domain decomposition method,

- the portion Γ_0 of the boundary with the homogeneous Dirichlet boundary conditions is deformed to obtain $t \to T_t(V)(\Gamma_0)$ as well as $t \to \mathcal{E}(\Omega_t)$ for the energy shape functional; as a result the first order shape derivative $J(\Omega) := d\mathcal{E}(\Omega; V)$ is obtained in the distributed form as a volume integral.
- the second order derivative of the energy functional is evaluated with respect to small parameter $\varepsilon \to 0$, the parameter governs the size of small inclusion with the material defined by a contrast parameter $\gamma \in [0, \infty)$.

We consider the energy shape functional $\Omega \to \mathcal{E}(\Omega)$ for Signorini problems for the Laplacian as well for the frictionless contact. The shape derivative $J(\Omega) := d\mathcal{E}(\Omega; V)$ of this functional is evaluated with respect to the boundary variations of the portion $\Gamma_0 \subset \partial\Omega$. In another words the velocity vector field V is supported in a small neighbourhood of Γ_0. The topological derivatives of $J(\Omega)$ are evaluated with respect to nucleation of small inclusions far from Γ_0. The domain decomposition method is applied in order to obtain the robust expressions for topological derivatives.

9 Conclusions

In the paper the review of mathematical techniques required for shape-topological sensitivity analysis for variational inequalities is presented. The singular geometrical perturbations depending on small parameter $\epsilon \to 0$ are considered. It is shown that the singular geometrical domain perturbations can be replaced, without any loss of precision, by the regular perturbations of bilinear forms depending on the small parameter. Non-smooth analysis is employed in order to obtain the second order topological derivatives. The proposed method can be now used in numerical methods of topology optimization as well in passive control of crack propagation.

Acknowledgments This work has been supported by the DFG EC315 'Engineering of Advanced Materials' and by the ANR-12- BS01-0007 Optiform.

References

1. A. Ancona, Sur les espaces de Dirichlet: principes, fonction de Green. J. Math. Pures Appl. **54**, 75–124 (1975)
2. I.I. Argatov, J. Sokołowski, Asymptotics of the energy functional in the Signorini problem under small singular perturbation of the domain. Zh. Vychisl. Mat. Mat. Fiz. **43**, 744–758, translation in Comput. Math. Math. Phys. **43**, 710–724 (2003)
3. Z. Belhachmi, J.-M. Sac-Epée, J. Sokołowski, Approximation par la méthode des élement finit de la formulation en domaine régulière de problems de fissures. Comptes Rendus Acad. Sci. Paris, Ser. I **338**, 499–504 (2004)

4. Z. Belhachmi, J.M. Sac-Epée, J. Sokołowski, Mixed finite element methods for smooth domain formulation of crack problems. SIAM J. Numer. Anal. **43**, 1295–1320 (2005)
5. A. Beurling, J. Deny, Dirichlet spaces. Proc. Natl. Acad. Sci. USA **45**, 208–215 (1959)
6. P. Destuynder, Remarques sur le contrôle de la propagation des fissures en régime stationnaire. Comptes Rendus Acad. Sci. Paris Sèr. II Méc. Phys. Chim. Sci. Univers Sci. Terre **308**, 697–701 (1989)
7. G. Fichera, Boundary value problems of elasticity with unilateral constraints, in *Festkörpermechanik/Mechanics of Solids, Handbuch der Physik*, ed. by S. Flügge, C.A. Truesdell (Springer, New York, 1984), pp. 391–424
8. G. Fichera, Existence theorems in elasticity, in *Festkörpermechanik/Mechanics of Solids, Handbuch der Physik*, ed. by S. Flügge, C.A. Truesdell (Springer, Berlin, 1984), pp. 347–389
9. G. Frémiot, Eulerian semiderivatives of the eigenvalues for Laplacian in domains with cracks. Adv. Math. Sci. Appl. **12**, 115–134 (2002)
10. G. Frémiot, J. Sokołowski, Hadamard formula in nonsmooth domains and applications, in *Partial Differential Equations on Multistructures (Luminy, 1999)*. Lecture Notes in Pure and Applied Mathematics, vol. 219 (Dekker, New York, 2001), pp. 99–120
11. G. Frémiot, J. Sokołowski, Shape sensitivity analysis of problems with singularities, in *Shape Optimization and Optimal Design (Cambridge, 1999)*. Lecture Notes in Pure and Applied Mathematics, vol. 216 (Dekker, New York, 2001), pp. 255–276
12. G. Frémiot, W. Horn, A. Laurain, M. Rao, J. Sokołowski, On the analysis of boundary value problems in nonsmooth domains. Diss. Math. **462**, 149 (2009)
13. W.A. Gross, The second fundamental problem of elasticity applied to a plane circular ring. Z. für Angew. Math. Phys. **8**, 71–73 (1957)
14. B. Hanouzet, J.-L. Joly, Méthodes d'ordre dans l'interprétation de certaines inéquations variationnelles et applications. J. Funct. Anal. **34**, 217–249 (1979)
15. A. Haraux, How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities. J. Math. Soc. Jpn. **29**, 615–631 (1977)
16. P. Hild, A. Münch, Y. Ousset, On the active control of crack growth in elastic media. Comput. Methods Appl. Mech. Eng. **198**, 407–419 (2008)
17. M. Hintermüller, V.A. Kovtunenko, From shape variation to topology changes in constrained minimization: a velocity method based concept, in Special issue on advances in shape and topology optimization: theory, numerics and new applications areas, ed. by C. Elliott, M. Hintermüller, G. Leugering, J. Sokołowski, *Optimization Methods and Software*, vol. 26 (2011), pp. 513–532
18. M. Hintermüller, A. Laurain, Optimal shape design subject to variational inequalities. SIAM J. Control Optim. **49**, 1015–1047 (2011)
19. D. Hömberg, A.M. Khludnev, J. Sokołowski, Quasistationary problem for a cracked body with electrothermoconductivity. Interfaces Free Bound. **3**, 129–142 (2001)
20. M. Kachanov, B. Shafiro, I. Tsukrov, *Handbook of Elasticity Solutions* (Kluwer Academic Publishers, Dordrecht, 2003)
21. A.M. Khludnev, Optimal control of crack growth in elastic body with inclusions. Eur. J. Mech. A Solids **29**, 392–399 (2010)
22. A.M. Khludnev, J. Sokołowski, *Modelling and Control in Solid Mechanics* (Birkhäuser, Basel 1997), reprinted by Springer in 2012
23. A.M. Khludnev, J. Sokołowski, On solvability of boundary value problems in elastoplasicity. Control Cybern. **27**, 311–330 (1998)
24. A.M. Khludnev, J. Sokołowski, The Griffith formula and the Rice-Cherepanov integral for crack problems with unilateral conditions in nonsmooth domains. Eur. J. Appl. Math. **10**, 379–394 (1999)
25. A.M. Khludnev, J. Sokołowski, Griffith's formula and Rice-Cherepanov's integral for elliptic equations with unilateral conditions in nonsmooth domains, in *Optimal Control of Partial Differential Equations (Chemnitz, 1998)*. International Series of Numerical Mathematics, vol. 133 (Birkhäuser, Basel, 1999), pp. 211–219

26. A.M. Khludnev, J. Sokołowski, Griffith's formulae for elasticity systems with unilateral conditions in domains with cracks. Eur. J. Mech. A Solids **19**, 105–119 (2000)
27. A.M. Khludnev, V.A. Kovtunenko, *Analysis of Cracks in Solids* (WIT Press, Southampton, 2000)
28. A.M. Khludnev, J. Sokołowski, Smooth domain method for crack problems. Q. Appl. Math. **62**, 401–422 (2004)
29. A.M. Khludnev, G. Leugering, Optimal control of cracks in elastic bodies with thin rigid inclusions. Z. Angew. Math. Mech. **91**, 125–137 (2011)
30. A.M. Khludnev, G. Leugering, M. Specovius-Neugebauer, Optimal control of inclusion and crack shapes in elastic bodies. J. Optim. Theory Appl. **155**, 54–78 (2012)
31. A.M. Khludnev, K. Ohtsuka, J. Sokołowski, On derivative of energy functional for elastic bodies with a crack and unilateral conditions. Q. Appl. Math. **60**, 99–109 (2002)
32. A.M. Khludnev, A.A. Novotny, J. Sokołowski, A. Żochowski, Shape and topology sensitivity analysis for cracks in elastic bodies on boundaries of rigid inclusions. J. Mech. Phys. Solids **57**, 1718–1732 (2009)
33. A.M. Khludnev, J. Sokołowski, K. Szulc, Shape and topological sensitivity analysis in domains with cracks. Appl. Math. **55**, 433–469 (2010)
34. G. Leugering, M. Prechtel, P. Steinmann, M. Stingl, A cohesive crack propagation model: mathematical theory and numerical solution. Commun. Pure Appl. Anal. **12**, 1705–1729 (2013)
35. T. Lewiński, J. Sokołowski, Energy change due to the appearance of cavities in elastic solids. Int. J. Solids Struct. **40**, 1765–1803 (2003)
36. F. Mignot, Contrôle dans les inéquations variationelles elliptiques. J. Funct. Anal. **22**, 130–185 (1976)
37. A. Münch, P. Pedregal, Relaxation of an optimal design problem in fracture mechanics: the anti-plane case. ESAIM Control Optim. Calc. Var. **16**, 719–743 (2010)
38. N.I. Muskhelishvili, *Some Basic Problems on the Mathematical Theory of Elasticity* (Noordhoff, Groningen, 1952)
39. S.A. Nazarov, J. Sokołowski, Asymptotic analysis of shape functionals. J. Math. Pures Appl. **82**(9), no. 2, pp. 125–196 (2003)
40. S.A. Nazarov, J. Sokołowski, Self-adjoint extensions for the Neumann Laplacian and applications. Acta Math. Sin. (Engl. Ser.) **22**(3), pp. 879–906 (2006)
41. S.A. Nazarov, J. Sokołowski, Shape sensitivity analysis of eigenvalues revisited. Control Cybern. **37**, 999–1012 (2008)
42. S.A. Nazarov, J. Sokołowski, Spectral problems in the shape optimisation. Singular boundary perturbations. Asymptot. Anal. **56**, 159–204 (2008)
43. S.A. Nazarov, J. Sokołowski, M. Specovius-Neugebauer, Polarization matrices in anisotropic heterogeneous elasticity. Asymptot. Anal. **68**, 189–221 (2010)
44. S.A. Nazarov, J. Sokołowski, On asymptotic analysis of spectral problems in elasticity. Lat. Am. J. Solids Struct. **8**, 27–54 (2011)
45. A.A. Novotny, J. Sokołowski, *Topological Derivatives in Shape Optimization*. Interaction of Mechanics and Mathematics (Springer, New York, 2013)
46. A.A. Novotny, J. Sokołowski, *Topological Derivative in Shape Optimization* (Springer, Berlin, 2013)
47. P. Plotnikov, J. Sokołowski, *Compressible Navier-Stokes Equations. Theory and Shape Optimization*. Mathematics Institute of the Polish Academy of Sciences. Mathematical Monographs (New Series), vol. 73 (Birkhäuser/Springer Basel AG, Basel, 2012)
48. M. Prechtel, P. Leiva Ronda, R. Janisch, G. Leugering, A. Hartmaier, P. Steinmann, *Cohesive Element Model for Simulation of Crack Growth in Composite Materials, International Conference on Crack Paths*, Vicenza (2009)
49. M. Prechtel, P. Leiva Ronda, R. Janisch, A. Hartmaier, G. Leugering, P. Steinmann, M. Stingl, Simulation of fracture in heterogeneous elastic materials with cohesive zone models. Int. J. Fract. **168**, 15–29 (2010)
50. M. Prechtel, G. Leugering, P. Steinmann, M. Stingl, Towards optimization of crack resistance of composite materials by adjusting of fiber shapes. Eng. Fract. Mech. **78**, 944–960 (2011)

51. V.V. Saurin, Shape design sensitivity analysis for fracture conditions. Comput. Struct. **76**, 399–405 (2001)
52. J. Sokołowski, J.-P. Zolésio, *Introduction to Shape Optimization, Shape Sensitivity Analysis*. Springer Series in Computational Mathematics, vol. 16 (Springer, Berlin, 1992)
53. J. Sokołowski, J.-P. Zolésio, *Introduction to Shape Optimization. Shape Sensitivity Analysis*, Springer Series in Computational Mathematics, vol. 16 (Springer, Berlin, 1992), reprinted in 2013
54. J. Sokołowski, A. Żochowski, On the topological derivative in shape optimization. SIAM J. Control Optim. **37**(4), 1251–1272 (1999)
55. J. Sokołowski, A. Żochowski, Optimality conditions for simultaneous topology and shape optimization. SIAM J. Control Optim. **42**(4), 1198–1221 (2003)
56. J. Sokołowski, A. Zochowski, Modelling of topological derivatives for contact problems. Numer. Math. **102**(1), 145–179 (2005)
57. J. Sokołowski, A. Żochowski, Modelling of topological derivatives for contact problems. Numer. Math. **102**, 145–179 (2005)
58. J. Sokołowski, A. Żochowski, Topological derivatives for optimization of plane elasticity contact problems. Eng. Anal. Bound. Elem. **32**, 900–908 (2008)

Recent Existence Results for Spectral Problems

Dario Mazzoleni

Abstract In this survey we present the new techniques developed for proving existence of optimal sets when one minimizes functionals depending on the eigenvalues of the Dirichlet Laplacian with a measure constraint, the most important being:

$$\min\big\{\lambda_k(\Omega) : \Omega \subset \mathbb{R}^N,\ |\Omega| = 1\big\}.$$

In particular we sketch the main ideas of some recent works, which allow to extend the now classic result by Buttazzo and Dal Maso to $\mathbb{R}^N$.

Keywords Shape optimization · Eigenvalues · Dirichlet laplacian

Mathematics Subject Classification (2010) Primary 49Q10 · Secondary 49R50

1 Introduction

The aim of this note is to report some recent existence results for classical shape optimization problems involving eigenvalues of the Dirichlet Laplacian. More precisely, we consider minimization problems of the following form:

$$\min\{\lambda_k(\Omega) : \Omega \in \mathcal{A}\}, \tag{1.1}$$

where $k \in \mathbb{N}$, λ_i denotes the ith eigenvalue of the Dirichlet Laplacian (counted with multiplicity) and $\mathcal{A}$ is the class of admissible shapes. A natural choice for this class, that we use in Sects. 3 and 4, is:

This work was done while the author was a Ph.D. student at the University of Pavia and at FAU of Erlangen. It has been partially supported by the ERC Starting Grant no. 258685 "AnOptSetCon".

D. Mazzoleni (✉)
Dipartimento di Matematica "G. Peano", Via Carlo Alberto, 10, 10123 Torino, Italy
e-mail: dmazzole@unito.it

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_8

$$\mathcal{A} := \left\{ \Omega \subset \mathbb{R}^N, \text{ quasi-open, } |\Omega| \leq 1 \right\}, \tag{1.2}$$

where $|\cdot|$ denotes the Lebesgue measure in $\mathbb{R}^N$, $N \in \mathbb{N}$. We need to have a bound on the measure of admissible sets, otherwise the monotonicity of Dirichlet eigenvalues would trivialize the problem; moreover the bound on the measure is taken less or equal to 1 only for simplicity: with every other positive constant the setting is unchanged. Then, since eigenvalues are decreasing with respect to set inclusion, it is equivalent to consider the problem with the equality constraint. An alternative (less common) choice, instead of the measure constraint, is a bound on the perimeter, which was studied only recently in [18]. The choice of *quasi-open*[1] sets is made in order to get compactness with a suitable topology and will be enlightened in Sect. 2. At last, one can consider also shapes contained in (see Sect. 2) or containing (see Sect. 5) a (quasi-)open bounded set.

Optimization problems like (1.1) naturally arise in the study of many physical phenomena, e.g. heat diffusion or wave propagation inside a domain $\Omega \subset \mathbb{R}^N$, and the literature is very wide (see [8, 13, 21, 22] for an overview), with many works in the last few years. Problem (1.1) in the class (1.2) was studied first by Lord Rayleigh in his treatise *The theory of sound* of 1877 (see [28]) and he conjectured the ball to be the optimal set when $k = 1$. This was proved by Faber [19] and Krahn [23, 24] in the 1920s, using techniques based on spherical decreasing rearrangements. From that result the case $k = 2$ follows with little additional effort: Krahn [23, 24] and Szegö [29] proved two disjoint equal balls of half measure each to be optimal. The situation for $k \geq 3$ becomes more complicate and it is not known what are the optimal shapes, yet. The only other functionals of eigenvalues for which the optimal shape is known are λ_1/λ_2 and λ_2/λ_3; Ashbaugh and Benguria (see [2]) proved that the minimizers are the unit ball and two equal disjoint balls of half measure each respectively.

Since the search for explicit optimal shapes did not give other results, it is natural to study at least whether a minimizer for (1.1) exists, and this subject turns out to be a difficult one, too. It is natural to attack an existence problem using the direct method of the Calculus of Variations. One first difficulty in order to apply it in this setting consists in finding a suitable notion of convergence for sets, which "behaves well" with respect to eigenvalues of Dirichlet Laplacian. More important, one needs also to find out how to suitably choose the class of admissible sets. It is immediately clear that the convergence in measure (or L^1 convergence of the characteristic functions) does not fit well, since it neglects sets of positive capacity: as an example one can consider a ball and the same ball minus a radius (in $\mathbb{R}^2$), which are the same set for this topology, but have different Dirichlet eigenvalues.

The search for a "right" notion of convergence in this setting was a main problem for many years. In the 1980s Dal Maso and Mosco (see [16, 17]) proposed the notion of γ-convergence, which has the "good" property that Dirichlet eigenvalues are continuous with respect to it. This was the main tool used by Buttazzo and Dal

[1] Quasi-open sets are superlevels of Sobolev functions.

Maso in 1993 (see [14]) for proving a fundamental existence result for a very general class of functions of eigenvalues, in the class of *quasi-open* sets inside a fixed bounded box. More precisely, they fix $D \subset \mathbb{R}^N$ bounded and open, and consider $F : \mathbb{R}^k \to \mathbb{R}$ a functional increasing in each variable and lower semicontinuous (l.s.c.). Then there exists a minimizer for the problem

$$\min \{F(\lambda_1(\Omega), \dots, \lambda_k(\Omega)) : \Omega \subset D, \text{ quasi-open}, |\Omega| \leq 1\}. \tag{1.3}$$

The above result gives a definitive answer to the existence problem for a very general class of spectral functionals in a bounded ambient space (actually it is sufficient to suppose D to have finite measure). We give the main ideas of the proof of this result in Sect. 2, together with some preliminaries about γ-convergence. The extension of the result by Buttazzo and Dal Maso to generic domains in $\mathbb{R}^N$ is a non trivial topic, because minimizing sequences, in principle, could have a significant portion of volume moving to infinity.

A first partial result in the direction of an extension to unbounded domains was obtained by Bucur and Henrot in 2000 (see [11]); they proved the existence of a minimizer for λ_3, using a concentration-compactness argument (see [5]). Moreover they showed that, given $k \geq 1$, if there exists a bounded minimizer for λ_j for all $j = 1, \dots, k-1$, then there exists a minimizer for λ_k (and more in general for Lipschitz functionals of the first k eigenvalues). Unfortunately this boundedness hypothesis was not known even for λ_3, till Dorin Bucur in a very recent paper (see [7]) was able to study the regularity of *energy shape subsolutions*. Employing techniques coming from the theory of free boundaries, it is possible to prove boundedness and finiteness of the perimeter for this class of sets, stable with respect to internal perturbations. Since optimal sets for (1.1) can be proved to be energy shape subsolutions, the existence of a minimizer for λ_k for all $k \in \mathbb{N}$ follows easily from the result by Bucur and Henrot. We present the ideas behind the proof of these results in Sect. 3.

In the same period another independent proof of existence of a solution for problem (1.3) in $\mathbb{R}^N$, with F satisfying the same hypotheses as in the result by Buttazzo and Dal Maso, was given by Mazzoleni and Pratelli (see [27]). Their idea consists in showing that, given a minimizing sequence for the problem

$$\min \left\{F(\lambda_1(\Omega), \dots, \lambda_k(\Omega)) : \Omega \subset \mathbb{R}^N, \text{ quasi-open}, |\Omega| \leq 1\right\}, \tag{1.4}$$

it is then possible to find a new one made of sets with diameter bounded by a constant depending only on k, N (but not on the particular functional) and with all the first k eigenvalues not increased. This argument, roughly speaking, works because sets with long "tails" can not have the first k eigenvalues very small. Moreover, with minor changes in the proof, it is also possible to deduce that every minimizer for (1.4) is bounded, provided that F is *weakly strictly increasing* (see [26]). This more "direct" method is presented in Sect. 4.

In recent years the existence of optimal sets was studied also for another kind of shape optimization problem (among sets with a measure constraint) involving

eigenvalues of Dirichlet Laplacian: when there is an internal *obstacle*, that is,

$$\min\left\{\lambda_k(\Omega) : D \subset \Omega \subset \mathbb{R}^N,\ \text{quasi-open},\ |\Omega| \leq 1\right\}, \tag{1.5}$$

where D is a fixed quasi-open box with $|D| \leq 1$. Bucur et al. in [10], using a concentration-compactness argument similar to the one in [5], proved existence of a solution for $k = 1$, gave a characterization of the cases when $k \geq 2$ and provided a partial regularity result for the solutions. In Sect. 5 we deal with the main ideas of their work.

The results exposed above give a quite complete understanding for the problem of existence of minimizers for spectral functionals involving eigenvalues of the Dirichlet Laplacian with a measure constraint. On the other hand the study of the regularity of solutions is still a main subject of research, both in the bounded (see [3]) and in the unbounded case (see the recent work [12]). In particular it is not known in general whether the minimizers for λ_k are open sets and not only quasi-open. This is one major open problem in spectral shape optimization.

It is also possible to consider minimization problems like (1.1) with perimeter constraint instead of volume constraint. This kind of problem was studied in the recent paper by De Philippis and Velichkov [18], where they prove that there exists a minimizer for

$$\min\left\{\lambda_k(\Omega) : \Omega \subset \mathbb{R}^N,\ \text{measurable},\ P(\Omega) \leq 1\right\}.$$

They use techniques to some extent analogous to those used by Bucur in [7], combining a concentration compactness argument and the study of the regularity for *perimeter* shape subsolutions. The perimeter constraint turns out to have a better *regularizing effect* than the volume constraint. In fact De Philippis and Velichkov are able to give many informations about regularity of optimal shapes: first of all the optimal shapes are open, so the above problem can be formulated among open sets. Moreover every optimal set Ω is bounded, has finite perimeter and its boundary $\partial\Omega$ is $C^{1,\alpha}$ for all $\alpha \in (0, 1)$, outside a closed set of Hausdorff dimension at most $N - 8$.

2 Preliminaries and Existence in a Bounded Box

First of all we need to recall some basic tools, which you can find in more detail in the books [8, 21, 22]. We define the Sobolev space $H^1_0(\Omega)$ as

$$H^1_0(\Omega) = \left\{u \in H^1(\mathbb{R}^N) :\ \text{cap}(\{u \neq 0\} \setminus \Omega) = 0\right\}, \tag{2.1}$$

where for every $E \subset \mathbb{R}^N$ the capacity of E is defined as

$$\mathrm{cap}(E) := \min \quad \left\{ \|v\|^2_{H^1(\mathbb{R}^N)} : v \in H^1(\mathbb{R}^N),\ v \geq 1 \text{ a.e. in a neighborhood of } E \right\}.$$

Then, given a function $u \in H^1_0(\Omega)$, its *quasi-continuous* representative is defined as

$$\widetilde{u}(x) := \lim_{r \to 0} \fint_{B_r(x)} u(y)\, dy.$$

Since outside a set of zero capacity every point is Lebesgue for u (see [22] for example), then the quasi-continuous representative is defined up to zero capacity and we identify every H^1 function with its quasi-continuous representative.

A set Ω is called *quasi-open* if for all $\varepsilon > 0$ there exists an open set Ω_ε such that $\mathrm{cap}\,(\Omega_\varepsilon \Delta \Omega) < \varepsilon$; for example superlevels of H^1 functions are quasi-open sets. Moreover, given a bounded open box D, we call $R_\Omega \colon L^2(D) \to L^2(D)$ the resolvent operator for the Dirichlet Laplacian, that is,

$$R_\Omega(f) := \arg\min \left\{ \frac{1}{2} \int_D |Du|^2 - \int_D uf,\ u \in H^1_0(\Omega) \right\},$$

for all $f \in L^2(D)$. The definition above can be extended also to capacitary[2] measures:

$$R_\mu(f) := \arg\min \left\{ \frac{1}{2} \int_D |Du|^2 + \int_D u^2\, d\mu - \int_D uf,\ u \in H^1_0(\Omega) \cap L^2_\mu(D) \right\}.$$

When $f = 1$, $R_\Omega(1) =: w_\Omega$ is called *torsion function* and it is an important tool for proving existence results. In particular w_Ω is the solution of

$$\begin{cases} -\Delta w = 1 \quad \text{in } \Omega, \\ \quad w \in H^1_0(\Omega), \end{cases}$$

and hence a minimizer for the so called *torsion energy functional*

$$E(\Omega) := \min_{u \in H^1_0(\Omega)} \left\{ \frac{1}{2} \int_D |Du|^2 - \int_D u \right\}.$$

After that, given a sequence of quasi-open sets contained in D, $(\Omega_n)_{n \in \mathbb{N}}$, we say that Ω_n γ-*converge* to a quasi-open set $\Omega \subset D$ as $n \to \infty$ when $w_{\Omega_n} \rightharpoonup w_\Omega$ in $H^1_0(D)$. Moreover Dal Maso and Mosco proved (see [16, 17]) that the convergence above implies for all $f \in L^2(D)$ $R_{\Omega_n}(f) \to R_\Omega(f)$ in $L^2(D)$, hence also $R_{\Omega_n} \to R_\Omega$ in the operator norm $\mathcal{L}(L^2(D))$ and hence the full spectrum converges. Thus eigenvalues of the Dirichlet Laplacian are continuous with respect to γ-convergence. Unfortunately, γ is a rather strong convergence and it is not compact in the class $\mathcal{A}(D) = \{\Omega \subset D,\ \text{quasi-open},\ |\Omega| \leq 1\}$; it is then necessary to weaken it, in order

[2] A Borel measure μ is called capacitary if, for every set E, $\mathrm{cap}\,(E) = 0$ implies $\mu(E) = 0$.

to apply the direct method of the calculus of variations to problem (1.3). A natural choice is the following.

A sequence $\Omega_n \in \mathcal{A}(D)$ is said to *weak γ-converge* to a domain $\Omega \in \mathcal{A}(D)$ if $w_{\Omega_n} \rightharpoonup w$ in $H^1_0(D)$ as $n \to \infty$, with $\Omega := \{w > 0\}$. Note that w coincide with $w_\Omega = R_\Omega(1)$ if and only if the convergence is γ and not only weak γ. More precisely, for some capacitary measure μ, $w = R_\mu(1)$: in fact we can say that the γ-convergence is compact in the class of capacitary measures, where a set Ω corresponds to the following measure:

$$\infty_\Omega(E) = \begin{cases} +\infty & \text{if } \operatorname{cap}(E \setminus \Omega) > 0, \\ 0 & \text{if } \operatorname{cap}(E \setminus \Omega) = 0. \end{cases}$$

A well known example of a sequence of quasi-open sets γ-converging to a measure which is not a quasi-open set is due to Cioranescu and Murat [15].

Buttazzo and Dal Maso used the compactness properties of the weak γ-convergence and the lower semicontinuity of Dirichlet eigenvalues with respect to it for proving a very general existence result.

Theorem 2.1 (Buttazzo–Dal Maso) *Let $D \subset \mathbb{R}^N$ be a bounded, open set and $F : \mathbb{R}^k \to \mathbb{R}$ be a functional increasing in each variable and lower semicontinuous (l.s.c.). Then there exists a minimizer for the problem*

$$\min\{F(\lambda_1(\Omega), \dots, \lambda_k(\Omega)) : \Omega \subset D, \ \textit{quasi-open}, \ |\Omega| \le 1\}. \tag{2.2}$$

First of all, the weak γ-convergence is built in order to be compact in the class $\mathcal{A}(D)$ and so a minimizing sequence converges, up to subsequences. Then it is easy to see that the weak γ-convergence is l.s.c. with respect to the Lebesgue measure, so the constraint $|\Omega| \le 1$ is satisfied by the limit of a weak γ-converging sequence of sets. It is then necessary to study the lower semicontinuity of the weak γ-convergence with respect to eigenvalues and this turns out to be a crucial point in the argument by Buttazzo and Dal Maso for proving Theorem 2.1. The following proposition gives a positive answer, for a (quite large) class of functionals.

Proposition 2.2 *A functional $J : \mathcal{A}(D) \to \mathbb{R}$ non decreasing with respect to set inclusion is γ l.s.c if and only if it is weak γ l.s.c.*

The hypothesis on the functional J to be nondecreasing with respect to set inclusion is quite strong, but it is satisfied by eigenvalues of the Dirichlet Laplacian and hence also by increasing functions of them. Thus the above Proposition can be applied in the hypothesis of Theorem 2.1.

The proof of Proposition 2.2 is based on the following (non trivial) key points, whose proof relies also on the maximum principle for the Dirichlet Laplacian.

(a) If w_{Ω_n} converge weakly in $H_0^1(D)$ to w and $v_n \in H_0^1(\Omega_n)$ converge weakly in $H_0^1(D)$ to v, then $v \in H_0^1(\{w > 0\})$.
(b) Let $\Omega_n \subset D$ be quasi-open sets such that w_{Ω_n} converge weakly in $H_0^1(D)$ to $w \in H_0^1(\Omega)$ for some quasi-open set $\Omega \subset D$. Then there exist a subsequence (not relabeled) and a sequence of quasi-open sets $\widetilde{\Omega}_n$ that γ-converge to Ω satisfying $\Omega_n \subset \widetilde{\Omega}_n \subset D$.

Then the Buttazzo and Dal Maso Theorem follows easily from Proposition 2.2 using the direct method of the Calculus of Variations. Given a minimizing sequence (Ω_n) of quasi-open sets for problem (1.3), by the compactness of the weak γ-convergence we can extract a subsequence (not relabeled) that weak γ-converges to a quasi-open set $\Omega \in \mathcal{A}(D)$. Using the properties of the weak γ-convergence highlighted above, the hypotheses on F and Proposition 2.2, we have that

$$|\Omega| \le \liminf_{n\to\infty} |\Omega_n| \le 1,$$
$$F(\lambda_1(\Omega), \dots, \lambda_k(\Omega)) \le \liminf_{n\to\infty} F(\lambda_1(\Omega_n), \dots, \lambda_k(\Omega_n)),$$

thus Ω is an optimal set for (1.3).

Remark 2.3 In the hypotheses of Theorem 2.1, it is sufficient to suppose that $D \subset \mathbb{R}^N$ has finite measure, so that the embedding $H^1(D) \hookrightarrow L^2(D)$ remains compact (see [9]).

3 Concentration Compactness and Subsolutions

The main problem in extending the result by Buttazzo and Dal Maso to (quasi-)open sets of $\mathbb{R}^N$ is the lack of compactness of the embedding $H^1(\mathbb{R}^N) \hookrightarrow L^2(\mathbb{R}^N)$. The concentration-compactness principle by P.L. Lions (see [25]) tries to focus on "how" the embedding $H^1(\mathbb{R}^N) \hookrightarrow L^2(\mathbb{R}^N)$ can be non compact. In the case of sets Bucur (see [5]) rearranged the principle in the following way, ruling out the *vanishing* case.

Theorem 3.1 (Lions, Bucur) *Let $(\Omega_n)_n \subset \mathbb{R}^N$ be a sequence of quasi-open sets with $|\Omega_n| \le 1$ for all $n \ge 1$. Then there exists a subsequence (not relabeled) such that one of the following situations occur:*

(1) **Compactness**. *There exist $y_n \in \mathbb{R}^N$ and a capacitary measure μ such that $R_{y_n+\Omega_n} \to R_\mu$ in $\mathcal{L}(L^2(\mathbb{R}^N))$.*
(2) **Dichotomy**. *There exist Ω_n^i, $i = 1, 2$ such that $|\Omega_n^i| > 0$, $d(\Omega_n^1, \Omega_n^2) \to \infty$ and $\|R_{\Omega_n} - R_{\Omega_n^1 \cup \Omega_n^2}\|_{L^2(\mathbb{R}^N)} \to 0$ as $n \to \infty$.*

Thanks to the concentration compactness argument above, it is easy to prove the following partial existence result (see [11]) for the unbounded case.

Theorem 3.2 (Bucur–Henrot) *For $k \geq 2$ if there exists a bounded minimizer for $\lambda_1, \dots, \lambda_k$ in the class $\mathcal{A}(\mathbb{R}^N)$, then there exists at least a minimizer for λ_{k+1} in $\mathcal{A}(\mathbb{R}^N)$.*

In particular this provides existence of a minimizer for the problem:

$$\min \left\{ \lambda_3(\Omega) : \Omega \subset \mathbb{R}^N, \text{ quasi-open}, \ |\Omega| \leq 1 \right\}, \tag{3.1}$$

since the minimizers for λ_1 and λ_2 are respectively a ball and two balls, which are bounded. The idea of the proof of Theorem 3.2 is quite simple. Given a minimizing sequence for λ_{k+1} in $\mathcal{A}(\mathbb{R}^N)$, made of bounded sets Ω_n, if compactness occur, existence follows directly considering the regular set[3] Ω_μ of the limit measure (see [21, Theorem 5.3.3]). On the other hand, if dichotomy happens, then $\Omega_n^1 \cup \Omega_n^2$ is also a minimizing sequence. But it is thus possible to see that the sequence $(\Omega_n^i)_n$ must be minimizing for some lower eigenvalue in the class $\mathcal{A}(\mathbb{R}^N)$, with different measure constraints: $c_1, c_2 > 0$ such that $c_1 + c_2 \leq 1$. Hence, up to translations, a minimizer for λ_{k+1} will be the union of the two minimizers corresponding to some lower eigenvalues. Note that if we do not know that there exists a bounded minimizer for every lower eigenvalue, it is not possible to consider the union of two of them, since in principle one can be dense in $\mathbb{R}^N$.

Since not even the boundedness of a minimizer for λ_3 was known, Bucur studied the link between this kind of shape optimization problems and free boundary problems, in order to be able to apply also in this framework the powerful techniques developed by Alt and Caffarelli (see [1]) and later implemented in the study of the energy of the Dirichlet Laplacian by Briançon, Hayouni and Pierre (see [4]).

First of all we need to be able to deal with measurable sets A, with $|A| < \infty$ (we call $\mathcal{M}$ the class of such sets), so we define the *Sobolev-like* space

$$\tilde{H}_0^1(A) := \left\{ u \in H^1(\mathbb{R}^N) : u = 0 \ \text{ a.e. on } \mathbb{R}^N \setminus A \right\}. \tag{3.2}$$

It is well known (see [18] for a more detailed discussion of those spaces) that there exists a quasi-open set $\omega_A \subseteq A$ such that

$$H_0^1(\omega_A) = \tilde{H}_0^1(A),$$

hence for functionals decreasing with respect to set inclusion (e.g. single eigenvalues) it is equivalent to solve problem (1.4) in the class of quasi-open sets with the classical definition of Sobolev space (2.1), or in the family of measurable sets associated to $\tilde{H}_0^1$.

Then it is possible to endow the family of measurable sets with a distance induced by γ-convergence:

[3] The regular set Ω_μ of a measure μ is the largest (in the sense of inclusion q.e.) countable union of sets of finite (μ-)measure.

$$d_\gamma(A, B) := \int_{\mathbb{R}^N} |w_A - w_B|, \qquad A, B \in \mathcal{M},$$

where we considered the torsion functions in $H^1(\mathbb{R}^N)$ extended to zero: $w_\Omega = 0$ in $\mathbb{R}^N \setminus \Omega$.

The most important notion in order to link shape optimization problems with free boundary problems is the one of *shape subsolution*.

Definition 3.3 We say that a set $A \in \mathcal{M}$ is a *local shape subsolution* for a functional $\mathcal{F}: \mathcal{M} \to \mathbb{R}$ if there exist $\delta > 0$ and $\Lambda > 0$ such that

$$\mathcal{F}(A) + \Lambda|A| \le \mathcal{F}(\tilde{A}) + \Lambda|\tilde{A}|, \qquad \forall\, \tilde{A} \subset A,\ d_\gamma(A, \tilde{A}) < \delta.$$

Roughly speaking, a shape subsolution is a set that is optimal with respect to internal perturbations. Bucur (see [7]) proved a very powerful regularity result for shape subsolution of the *torsion energy* functional

$$E(A) := \min_{u \in \tilde{H}^1_0(A)} \left\{ \frac{1}{2} \int_{\mathbb{R}^N} |Du|^2 - \int_{\mathbb{R}^N} u \right\}.$$

Theorem 3.4 *Let A be a local shape subsolution (with constants δ, Λ) for the torsion energy E. Then it is bounded, with $diam(A) \le C(|A|, \delta, \Lambda)$, has finite perimeter and its fine interior has the same measure of A.*

The proof of the theorem for the finite perimeter part is based on controlling the term $\int_{\{0 \le w_A \le \varepsilon\}} |Dw_A|^2$, while the boundedness and the inner density estimate come from the following Alt–Caffarelli type estimate: there exist $r_0, C_0 > 0$ such that for all $r \le r_0$

$$\sup_{B_{2r}(x)} w_A \le C_0 r \qquad \text{implies} \qquad u = 0 \text{ in } B_r(x).$$

The next key point in Bucur's argument consists in linking the minimizers of eigenvalues of Dirichlet Laplacian with shape subsolution of the energy. We consider the minimization problem, equivalent to (1.4) up to choose $\Lambda > 0$ small enough (for a detailed discussion about this equivalence, see [4]),

$$\min \left\{ F(\lambda_1(A), \dots, \lambda_k(A)) + \Lambda|A| : A \subset \mathbb{R}^N,\ \text{quasi-open} \right\}, \tag{3.3}$$

for a functional $F: \mathbb{R}^k \to \mathbb{R}$ which satisfies the following Lipschitz-like condition for some positive α_i, $i = 1, \dots, k$:

$$F(x_1, \dots, x_k) - F(y_1, \dots, y_k) \le \sum_{i=1}^{k} \alpha_i (x_i - y_i), \qquad \forall x_i \ge y_i,\ i = 1, \dots, k. \tag{3.4}$$

Theorem 3.5 *Assume that A is a solution of (3.3), then it is a local shape subsolution for the energy problem.*

The proof is based on [6, Theorem 3.4], which assures, for all $k \in \mathbb{N}$, the existence of a constant $c_k(A)$ such that:

$$\left| \frac{1}{\lambda_k(\tilde{A})} - \frac{1}{\lambda_k(A)} \right| \leq c_k(A) d_\gamma(A, \tilde{A}).$$

Then, up to choose δ small enough and $\tilde{A} \subseteq A$ with $d_\gamma(\tilde{A}, A) < \delta$, it follows

$$\begin{aligned} \Lambda(|A| - |\tilde{A}|) &\leq F(\lambda_1(\tilde{A}), \dots, \lambda_k(\tilde{A})) - F(\lambda_1(A), \dots, \lambda_k(A)) \\ &\leq \sum_i \alpha_i (\lambda_i(\tilde{A}) - \lambda_i(A)) \\ &\leq \sum_i \alpha_i c_i'(E(\tilde{A}) - E(A)) \leq K(E(\tilde{A}) - E(A)), \end{aligned}$$

with a constant K depending on $c_i' = c_i'(A, \delta, i)$ and α_i, for $i = 1, \dots, k$.

Now a straightforward application of Theorem 3.4, together with Theorem 3.2, gives the main existence result.

Theorem 3.6 (Bucur) *If the functional F satisfies the Lipschitz-like condition (3.4), then problem (3.3) has at least a solution for every $k \in \mathbb{N}$. Moreover every optimal set is bounded and has finite perimeter.*

In particular there exists a solution for the problem

$$\min \left\{ \lambda_k(\Omega) : \Omega \subset \mathbb{R}^N \text{ quasi-open}, |\Omega| \leq 1 \right\},$$

for all $k \in \mathbb{N}$. We highlight here that, to our knowledge, it is not known yet if the above problem admits solutions in the class of *open* sets. It is possible to give slightly different proof of Theorem 3.6 that does not use the concentration-compactness principle, but only the regularity of energy shape subsolutions. This proof is due to Bozhidar Velichkov and it has never appeared on a published paper, to our knowledge.

Remark 3.7 (Velichkov) Let $(\Omega_n)_{n \geq 1}$ be a minimizing sequence for problem (3.3), with $|\Omega_n| < \infty$ for all $n \in \mathbb{N}$, and then we consider, for all $n \in \mathbb{N}$, the minimum problem

$$\min \{ F(\lambda_1(\Omega), \dots, \lambda_k(\Omega)) + \Lambda |\Omega| : \Omega \subset \Omega_n \},$$

for some $\Lambda > 0$ sufficiently small. Theorem 2.1 by Buttazzo and Dal Maso assures that there exists a solution Ω_n^*, but this is also a subsolution and hence by Theorem 3.4 it has diameter uniformly bounded, depending only on k, N. Hence we have a new minimizing sequence Ω_n^* uniformly bounded to which it is possible to apply again Theorem 2.1, thus obtaining existence for problem (3.3).

4 How to Choose an Uniformly Bounded Minimizing Sequence

In this section we aim to provide the main ideas of the proof of the existence theorem presented by Mazzoleni and Pratelli in [27], which uses an "elementary" method that requires neither a concentration-compactness argument nor regularity of shape subsolutions.

Theorem 4.1 *Let $k, N \in \mathbb{N}$ and $F : \mathbb{R}^k \to \mathbb{R}$ be a functional increasing in each variable and l.s.c., then there exists a (bounded) minimizer for the problem*

$$\min \left\{ F(\lambda_1(\Omega), \dots, \lambda_k(\Omega)) : \Omega \subset \mathbb{R}^N, \ \textit{quasi-open}, \ |\Omega| \leq 1 \right\}. \tag{4.1}$$

More precisely the diameter of the optimal set is controlled by a constant depending only on k, N (and not on the particular functional F).

The proof is based on the following Proposition, which gives the possibility to consider minimizing sequences for (4.1) with uniformly bounded diameters, which means that we can employ Buttazzo–Dal Maso Theorem.

Proposition 4.2 *If $\Omega \subset \mathbb{R}^N$ is an open set with unit volume, there exists another open set of unit volume, $\widehat{\Omega}$, contained in cube of side $R = R(k, N)$ and such that*

$$\lambda_i(\widehat{\Omega}) \leq \lambda_i(\Omega), \qquad \forall\, i = 1, \dots, k.$$

From Proposition 4.2, Theorem 4.1 follows easily: in fact, given a minimizing sequence $(\Omega_n)_{n\in\mathbb{N}}$ made of open sets with unit volume, it is sufficient to take the corresponding sequence $(\widehat{\Omega}_n)_{n\in\mathbb{N}}$, which is again minimizing and then to apply Theorem 2.1 by Buttazzo and Dal Maso to it.

On the other hand the proof of Proposition 4.2 is quite delicate: we give here below the main ideas of how it is carried on. In particular, given Ω open and with unit volume, we focus on its left "tail", that is, the set

$$\Omega^l_{\bar{t}} := \left\{ x \in \Omega : x_1 < \bar{t} \right\},$$

for a $\bar{t}$ such that $\left|\Omega^l_{\bar{t}}\right| = \widehat{m}$, for a suitably choosen $\widehat{m}$, very small but fixed (depending only on k, N). Then it is possible to find a new set $\widehat{\Omega}$ with bounded tail and the first k eigenvalues lowered. We need some notations: for all $t \leq \bar{t}$ we define:

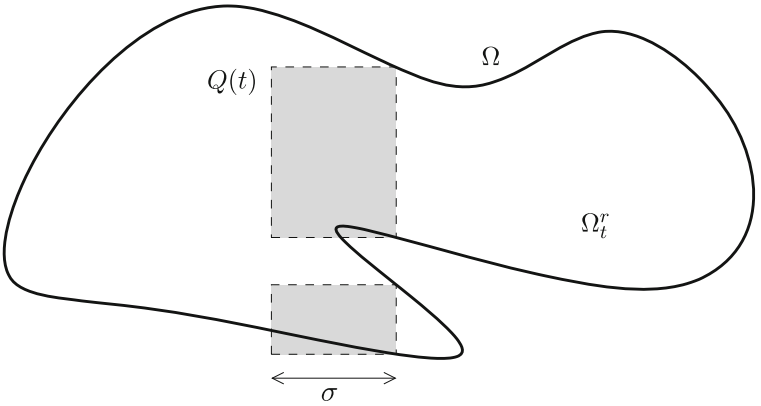


Fig. 1 A set Ω with the cylinder $Q(t)$ (*shaded*)

$$\begin{aligned}
\Omega_t^r &:= \{x \in \Omega : x_1 > t\}, \\
\Omega_t &:= \Big\{(x_2, \dots, x_N) \in \mathbb{R}^{N-1} : (t, x_2, \dots, x_N) \in \Omega\Big\}, \\
\varepsilon(t) &:= \mathcal{H}^{N-1}(\Omega_t), \quad m(t) = \int_{-\infty}^{t} \varepsilon(s)\, ds, \\
\delta(t) &:= \sum_{i=1}^{k} \int_{\Omega_t} |Du_i(t, x_2, \dots, x_N)|^2 \, d\mathcal{H}^{N-1}.
\end{aligned}$$

For all $t \le \bar{t}$ it is possible to compare the first k eigenvalues of Ω with those of $\widetilde{\Omega}(t) := \Omega_t^r \cup Q(t)$, which is obtained by "cutting" the "tail" at level t and adding a suitable small cylinder $Q(t)$ of height $\sigma(t) = \varepsilon(t)^{\frac{1}{N-1}}$ (see Fig. 1).

Using the min-max principle for eigenvalues one obtains

$$\lambda_i(\widetilde{\Omega}(t)) \le \lambda_i(\Omega) + C(k, N)\varepsilon(t)^{\frac{1}{N-1}}\delta(t), \qquad \forall i = 1, \dots, k,$$

if $\varepsilon(t), \delta(t) \le \nu$, for some constant $\nu = \nu(k, N)$. After rescaling $\widetilde{\Omega}(t)$ to unit volume, being $\widehat{\Omega}(t) := |\widetilde{\Omega}(t)|^{-\frac{1}{N}}\widetilde{\Omega}(t)$, it is possible to prove that for a suitable constant $\overline{C} = \overline{C}(k, N)$, exactly one of the following conditions hold.

(1) $\max\{\varepsilon(t), \delta(t)\} > \nu$.
(2) (1) does not hold and $m(t) \le \overline{C}(\varepsilon(t) + \delta(t))\varepsilon(t)^{\frac{1}{N-1}}$.
(3) (1) and (2) do not hold and for all $i = 1, \dots, k$, $\lambda_i(\widehat{\Omega}(t)) < \lambda_i(\Omega)$. Moreover if $m(t) \ge \widehat{m}$, then there exist $\eta = \eta(k, N)$ such that $\lambda_i(\widehat{\Omega}(t)) < \lambda_i(\Omega) - \eta$ for all $i = 1, \dots, k$.

In order to conclude the boundedness of the "tail", we define

$$\hat{t} := \sup\Big\{t \le \bar{t} : \text{condition (3) holds for } t\Big\},$$

with the usual convention that $\hat{t} = -\infty$ if condition (3) is false for every $t \leq \bar{t}$. We consider the following subsets of $(\hat{t}, \bar{t})$

$$A := \left\{ t \in (\hat{t}, \bar{t}) : \text{ condition (1) holds for } t \right\},$$
$$B := \left\{ t \in (\hat{t}, \bar{t}) : \text{ condition (2) holds for } t \text{ and } m(t) > 0 \right\},$$

and it is possible to prove that $|A| + |B| \leq C(k, N)$, since in this case we obtain a differential equation, regarding the measure of the "tail", since for a.e. $t \in \mathbb{R}$ $m'(t) = \varepsilon(t)$,

$$m'(t) \geq \frac{1}{C} m(t)^{\frac{N-1}{N}},$$

and an analogous one about $\int_{-\infty}^{t} \delta(s)\, ds$.

Then, if $\hat{t} = -\infty$, that is, only case (1) or (2) happen, the set Ω has itself a bounded "tail".

On the other hand, if $\hat{t} > -\infty$, we pick a $t^\star \in [\hat{t} - 1, \hat{t}]$ for which condition (3) holds and consider $U_1 := \widehat{\Omega}(t^\star)$.

If $m(t^\star) < \widehat{m}$, then we have that $\lambda_i(U_1) < \lambda_i(\Omega)$ for all $i = 1, \ldots, k$ and U_1 has a bounded "tail", so we have concluded. Instead, if $m(t^\star) \geq \widehat{m}$, the stronger estimate $\lambda_i(U_1) < \lambda_i(\Omega) - \eta$ holds for all $i = 1, \ldots, k$, but possibly U_1 has not bounded "tail". Hence we iterate the procedure, by applying the whole construction to U_1 and thus finding U_2 which either has bounded "tail", or it satisfies $\lambda_i(U_2) < \lambda_i(\Omega) - 2\eta$ for all $i = 1, \ldots, k$. After l steps, if we have not concluded yet, there is U_l such that

$$\lambda_i(U_l) < \lambda_i(\Omega) - l\eta, \qquad \forall\, i = 1, \ldots, k.$$

Since we can reduce to consider sets with $\lambda_k(\Omega) \leq M$ (see [27, Appendix]), the inequality above is impossible if $l\eta \geq M$: as a consequence, the iteration must stop after less than M/η steps, thus finding a set with bounded "tail" and with the first k eigenvalues lowered.

The same procedure can be performed with small changes also for the "inner" part of the set, that is, $\Omega^i := \left\{ (x, y) \in \Omega : \widehat{m} \leq |\Omega^-(x)| \leq 1 - \widehat{m} \right\}$ and to the right tail. Then one can apply the same arguments in all the other coordinate directions. This concludes Proposition 4.2.

At this point, Theorem 4.1 does not guarantee that *every* minimizer is bounded, in fact a constant functional satisfies the hypothesis of the Theorem, but it can not have minimizers uniformly bounded! With a necessary additional assumption on the functional F, in [26] was proved the following.

Theorem 4.3 *In the hypotheses of Theorem 4.1, if the functional F is also* weakly strictly increasing*, that is,*

$$\forall x_i < y_i,\ \forall i = 1, \ldots, k, \qquad F(x_1, \ldots, x_k) < F(y_1, \ldots, y_k),$$

then all the minimizers for problem (4.1) *have diameter bounded by a constant depending only on* k, N.

The proof is carried out improving Proposition 4.2. More precisely, given a sequence of open sets with unit measure that γ-converge to a minimum Ω for problem (4.1), then either, up to pass to a subsequence, $diam(\Omega_n) \leq C(k, N)$, or there exist $(\widehat{\Omega}(t_n))_n$ open sets with unit measure (obtained with a "cutting" procedure as above) such that

$$\lambda_i(\widehat{\Omega}(t_n)) < \lambda_i(\Omega) - \eta(k, N), \qquad \forall\, i = 1, \dots, k.$$

Hence in this last case,

$$\inf_n \left\{ F\Big(\lambda_1(\Omega(t_n)), \dots, \lambda_k(\Omega(t_n))\Big)\right\} < F(\lambda_1(\Omega), \dots, \lambda_k(\Omega)),$$

which is absurd.

Remark 4.4 The main differences in the existence results in $\mathbb{R}^N$ described in this section and in the previous one are the following:

- Bucur's proof gives the important information that all minimizers have finite perimeter, while this property can not be easily deduced from the approach by Mazzoleni and Pratelli;
- On the other hand, the result by Bucur applies to "Lipschitz" functionals of the first k eigenvalues (more precisely satisfying condition (3.4)), while the method by Mazzoleni and Pratelli requires the functionals only to be increasing in each variables and l.s.c.

Remark 4.5 As we have already highlighted, the regularity issue for problem (4.1) is a difficult one and it is not completely understood yet, to our knowledge. In the recent work [12] it is proved that optimal sets are open for a very special class of functional, among which $\lambda_1(\cdot) + \cdots + \lambda_k(\cdot)$. Moreover it is shown that an optimal set for $\lambda_k(\cdot)$ admits an eigenfunction, corresponding to the k th eigenvalue, which is Lipschitz continuous in $\mathbb{R}^N$, but this does not assure the openness.

5 The Case of Internal Constraint

In this section we present the approach used in [10] in order to give some existence results for problem (1.5) with an internal constraint, where D is a quasi-open set with $|D| \leq 1$, possibly unbounded. The main point in order to prove existence results for such problems is the following concentration-compactness principle, inspired by Theorem 3.1, in the case of inner constraints. We remark that the main changes are in the "compactness" case, where there are no more translations.

Theorem 5.1 *Let $(\Omega_n)_n$ be a sequence of quasi-open sets in $\mathbb{R}^N$, each of them containing a given quasi-open set D, with $|\Omega_n| \le 1$ for all $n \ge 1$. Then there exists a subsequence (not relabeled) such that one of the following situations occur:*

(1) **Compactness**. *There exists a capacitary measure μ such that $R_{\Omega_n} \to R_\mu$ in $\mathcal{L}(L^2(\mathbb{R}^N))$ and moreover $D \subset \Omega_\mu$.*

(2) **Dichotomy**. *There exist Ω_n^i, $i = 1, 2$ such that $\liminf_{n\to\infty} |\Omega_n^i| > 0$, $d(\Omega_n^1, \Omega_n^2) \to \infty$ and $\|R_{\Omega_n} - R_{\Omega_n^1 \cup \Omega_n^2}\|_{L^2(\mathbb{R}^N)} \to 0$ as $n \to \infty$. Moreover $\limsup_{n\to\infty} |\Omega_n^1 \cap D| = 0$ or $\limsup_{n\to\infty} |\Omega_n^2 \cap D| = 0$.*

From the above concentration-compactness principle it is possible to prove the following existence result (see [10]). First of all we need to introduce, for $m \ge 0$, the value

$$\lambda_k^*(m) := \inf \{\lambda_k(\Omega) : \Omega \text{ quasi-open}, |\Omega| \le m\}.$$

Theorem 5.2 *Let D be a quasi-open set with $|D| \le 1$. For $k \in \mathbb{N}$ we define*

$$\alpha_k := \inf \left\{\lambda_k(\Omega) : D \subset \Omega \subset \mathbb{R}^N, \ \textit{quasi-open}, |\Omega| \le 1\right\}. \tag{5.1}$$

If $k = 1$ the problem has at least a solution.
For $k \ge 2$ one of the following assertions holds:

(a) Problem (1.5) has a solution;
(b) There exists $l \in \{1, \dots, k-1\}$ and an admissible quasi-open set Ω such that $\alpha_k = \lambda_{k-l}(\Omega) = \lambda_l^*(1 - |\Omega|)$;
(c) There exists $l \in \{1, \dots, k-1\}$ such that $\alpha_k = \lambda_l^*(1 - |D|) > \lambda_{k-l}(D)$.

Clearly in case (b) and (c) we do not have existence of a solution in general. Something more can be said with stronger hypotheses on D and it will be stated later. Now we sketch the proof of Theorem 5.2 for the case $k = 1$. Let $(\Omega_n)_{n\ge1}$ be a minimizing sequence such that $\liminf_{n\to\infty} |\Omega_n|$ is minimal (clearly the value must be strictly positive). Following Theorem 5.1, if we are in the compactness situation, there is a subsequence (not relabeled) that γ-converges to a capacitary measure μ. The set $\Omega_\mu := \left\{R_\mu(1) > 0\right\}$ is admissible and thus it is a solution.

On the other hand, if dichotomy occurs, we get a contradiction. We may assume that $\lambda_1(\Omega_n^1 \cup \Omega_n^2) = \lambda_1(\Omega_n^1)$, since the two sets have positive distance, and clearly the sequence $(\Omega_n^1 \cup D)_n$ is also minimizing. Then only two situations can happen:

1. Either $\liminf_{n\to\infty} |\Omega_n^1 \cup D| < \liminf_{n\to\infty} |\Omega_n|$;
2. Or $\lim_{n\to\infty} |\Omega_n^2 \setminus D| = 0$.

Case (1) contradicts the fact that $(\Omega_n)_n$ is the minimal minimizing sequence. Also case (2) is impossible, since it implies $d(\Omega_n^1, \{0\}) \to \infty$, otherwise the measure of D would be infinite. Hence $|\Omega_n^1 \cap D| \to 0$ and consider the ball B with measure equal to $\limsup_n |\Omega_n^1|$: $B \cup D$ is a solution for every position of B, and when B intersects (but not cover) some connected component of D we have the contradiction.

The proof of the case $k \geq 2$ follows from similar ideas. One takes again $(\Omega_n)_n$ a minimizing sequence with minimal lim inf $|\Omega_n|$ and if there is compactness one gets immediately the existence of a solution. If dichotomy happens, then we can suppose

$$|\Omega_n^1| \to \alpha^1, \qquad |\Omega_n^2| \to \alpha^2, \qquad |\Omega_n^1 \cap D| \to 0,$$

and (up to subsequences) we can take the maximal $l \in \{1, \dots, k-1\}$ for which one of the following holds:

- $|\lambda_k(\Omega_n) - \lambda_{k-l}(\Omega_n^2)| \to 0$ and $\lambda_l(\Omega_n^1) \leq \lambda_{k-l}(\Omega_n^2) \leq \lambda_{l+1}(\Omega_n^1)$,
- $|\lambda_k(\Omega_n) - \lambda_l(\Omega_n^1)| \to 0$ and $\lambda_{k-l}(\Omega_n^2) \leq \lambda_l(\Omega_n^1) \leq \lambda_{k-l+1}(\Omega_n^2)$.

It is clear that case (b) of the thesis follows from the first one and case (c) follows from the second one. With an easy induction argument one can now conclude.

The next result highlight that stronger hypotheses lead to a good improvement.

Theorem 5.3 *In the hypotheses of Theorem 5.2, if moreover we ask the set D to be bounded, then also the cases (b) and (c) of Theorem 5.1 lead to the existence of a solution.*

Moreover in [10] are proved also some regularity properties of solutions of (5.1). In particular, if $k = 1$, $|D| < 1$ and D is *quasi-connected*,[4] all the minimizers are open sets even if D is only quasi-open.

Acknowledgments The author wishes to thank Giovanni Franzina for some discussions about the paper.

References

1. H.W. Alt, L.A. Caffarelli, Existence and regularity for a minimum problem with free boundary. J. Reine Angew. Math. **325**, 105–144 (1981)
2. M.S. Ashbaugh, R. Benguria, Proof of the Payne-Pölya-Weinberger conjecture. Bull. Amer. Math. Soc. **25**(1), 19–29 (1991)
3. T. Briançon, J. Lamboley, Regularity of the optimal shape for the first eigenvalue of the Laplacian with volume and inclusion constraints. Ann. I. H. Poincaré - AN **26**(4), 1149–1163 (2009)
4. T. Briançon, M. Hayouni, M. Pierre, Lipschitz continuity of state functions in some optimal shaping. Calc. Var. PDE **23**(1), 13–32 (2005)
5. D. Bucur, Uniform concentration-compactness for Sobolev spaces on variable domains. J. Differ. Equ. **162**, 427–450 (2000)
6. D. Bucur, Regularity of optimal convex shapes. J. Convex Anal. **10**, 501–516 (2003)
7. D. Bucur, Minimiziation of the k-th eigenvalue of the Dirichlet Laplacian. Arch. Ration. Mech. Anal. **206**(3), 1073–1083 (2012)
8. D. Bucur, G. Buttazzo, *Variational Methods in Shape Optimization Problem*. Progress in Nonlinear Differential Equations and Their Applications (Birkhlauser Verlag, Boston, 2005)
9. D. Bucur, G. Buttazzo, On the characterization of the compact embedding of Sobolev spaces. Calc. Var. PDE **44**(3–4), 455–475 (2012)

[4] A quasi-open set D is called quasi-connected if for all open and nonempty set A_1, A_2 such that cap $(A_i \cap D) > 0$ for $i = 1, 2$ and with $D \subset A_1 \cup A_2$, we have cap$(A_1 \cap A_2) > 0$.

10. D. Bucur, G. Buttazzo, B. Velichkov, Spectral optimization problems with internal constraint. Ann. I. H. Poincaré - AN **30**(3), 477–495 (2013)
11. D. Bucur, A. Henrot, Minimization of the third eigenvalue of the Dirichlet Laplacian. Proc. R. Soc. Lond. **456**, 985–996 (2000)
12. D. Bucur, D. Mazzoleni, A. Pratelli, B. Velichkov, Lipschitz regularity of the eigenfunctions on optimal domains, to appear on Arch. Ration. Mech. Anal., doi:10.1007/s00205-014-0801-6. Preprint available at http://cvgmt.sns.it/person/977
13. G. Buttazzo, Spectral optimization problems. Rev. Mat. Complut. **24**(2), 277–322 (2011)
14. G. Buttazzo, G. Dal Maso, An existence result for a class of shape optimization problems. Arch. Ration. Mech. Anal. **122**, 183–195 (1993)
15. D. Cioranescu, F. Murat, A strange term coming from nowhere. Top. Math. Model. Compos. Mater. Prog. Nonlinear Differ. Equ. Appl. **31**, 45–93 (1997)
16. G. Dal, U. Maso, Wiener criteria and energy decay for relaxed Dirichlet problems. Arch. Ration. Mech. Anal. **95**, 345–387 (1986)
17. G. Dal, U. Masco, Wiener's criterion and Γ-convergence. Appl. Math. Optim. **15**, 15–63 (1987)
18. G. De Philippis, B. Velichkov, Existence and regularity of minimizers for some spectral optimization problems with perimeter constraint. Appl. Math. Optim. **69**, 199–231 (2014)
19. G. Faber, Beweiss, dass unter alles homogenen Membranen von gleicher Flache und gleicher Spannung die kreisformige den tiefsten Grundton gibt. Sitz. Ber. Bayer. Akad. Wiss. 169–172 (1923)
20. A. Henrot, Minimization problems for eigenvalues of the Laplacian. J. Evol. Equ. **3**, 443–461 (2003)
21. A. Henrot, *Extremum Problems for Eigenvalues of Elliptic Operators*. Frontiers in Mathematics (Springer, 2006)
22. A. Henrot, M. Pierre, *Mathématiques et Applications*, Variation et optimisation de formes (Springer, New York, 2005)
23. E. Krahn, Über eine von Rayleigh formulierte Minimaleigenschaft des Kreises. Math. Ann. **94**, 97–100 (1924)
24. E. Krahn, Über Minimaleigenschaften der Kugel in drei und mehr Dimensionen. Acta Comm. Univ. Dorpat. **A9**, 1–44 (1926)
25. P.L. Lions, The concentration-compactness principle it the calculus of variation. The locally compact case, part 1. Ann. I. H. Poincaré – AN **1**(2), 109–145 (1984)
26. D. Mazzoleni, Boundedness of minimizers for spectral problems in $\mathbb{R}^N$, to appear on Rend. Sem. Mat. Univ. Padova, preprint http://cvgmt.sns.it/person/977
27. D. Mazzoleni, A. Pratelli, Existence of minimizers for spectral problems. J. Math. Pures Appl. **100**(3), 433–453 (2013)
28. L. Rayleigh, *The Theory of Sound*, 1st edn. (Macmillan, London, 1877)
29. G. Szegö, Inequalities for certain eigenvalues of a membrane of given area. J. Ration. Mech. Anal. **3**, 343–356 (1954)

Approximate Shape Gradients for Interface Problems

A. Paganini

Abstract Shape gradients of shape differentiable shape functionals constrained to an interface problem (IP) can be formulated in two equivalent ways. Both formulations rely on the solution of two IPs, and their equivalence breaks down when these IPs are solved approximatively. We establish which expression for the shape gradient offers better accuracy for approximations by means of finite elements. Great effort is devoted to provide numerical evidence of the theoretical considerations.

Keywords Shape gradients · Finite element approximations · Interface problems

Mathematics Subject Classification (2010) 65D15 · 65N30 · 49Q12

1 Introduction

Optimal control of mathematical models is a core activity of applied mathematics. The goal is to optimize model parameters with respect to target functionals: real mappings on the set of all admissible configurations. In many practical cases the control parameter is the shape of a structure [1, 2]. In this case we speak of *shape functionals* and, in particular, of PDE constrained shape functionals, when the mapping involves the solution of a PDE, the so-called *state problem*.

The sensitivity of shape functionals with respect to perturbations of shapes is expressed by the *shape gradient*: a linear bounded operator on the space of perturbation directions. The knowledge of this mapping is the starting point for gradient based shape optimization [1–6].

Shape gradients of shape differentiable shape functionals can be stated equivalently as an integration over the volume and as an integration on the boundary

The work of A. Paganini was partly supported by ETH Grant CH1-02 11-1.

A. Paganini (✉)
Seminar for Applied Mathematics, ETH Zurich, 8092 Zurich, Switzerland
e-mail: alberto.paganini@sam.math.ethz.ch

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_9

[7, Chap. 9, Theorem 3.6]. In the case of PDE constrained shape functionals, shape gradients depend on the solution of the state problem and, in general, on the solution of an additional PDE, the so-called *adjoint problem*. When the state and the adjoint solutions are replaced with numerical approximations, the equivalence of the two representations of the shape gradient breaks down [8].

Several authors suggested that the volume based formulation is better suited, when discretizations by means of finite elements are considered, cf. [7, Chap. 10, Remark 2.3], [8] and [9, Chap. 3.3.7]. However, to our knowledge, thorough convergence analysis and numerical evidence have not been provided. For the case of elliptic boundary value problem constraints, a first theoretical investigation was conducted in [10]. The aim of this work is to extend these results to the case of elliptic interface value problems. In particular, we devote great effort to provide numerical evidence through numerical experiments. For the sake of simplicity, we restrict our considerations to a class of shape functionals and interface problems. Nevertheless, we believe that our test case is representative and that no important aspect is missing.

2 Shape Gradients

A *shape functional* is a real valued map $\mathcal{J} : \mathcal{A} \to \mathbb{R}$ defined on a set of admissible domains $\mathcal{A}$, which is usually constructed starting from an initial open bounded domain Ω. In the general approach by Delfour–Zolesio [7, Chap. 4], $\mathcal{A}$ comprises all domains $T_s(\Omega)$ that are generated through the evolution $T_s(\cdot)$ of the flow of a non-autonomous vector field $\mathcal{V}$.

For a fixed perturbation direction $\mathcal{V}$, the *Eulerian derivative*

$$d\mathcal{J}(\Omega; \mathcal{V}) := \lim_{s \searrow 0} \frac{J\left(T_s(\Omega)\right) - J(\Omega)}{s} \tag{1}$$

expresses the sensitivity of the shape functional $\mathcal{J}$ with respect to the perturbation direction $\mathcal{V}$. Without loss of generality, the vector field $\mathcal{V}$ can be assumed to be autonomous [7, Chap. 9, Sect. 3.1]. The shape functional $\mathcal{J}$ is said to be *shape differentiable* at Ω if (1) defines a linear bounded mapping

$$d\mathcal{J}(\Omega; \cdot) : W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d) \to \mathbb{R}, \quad \mathcal{V} \mapsto d\mathcal{J}(\Omega; \mathcal{V}), \tag{2}$$

which is called the *shape gradient* of $\mathcal{J}$ at Ω. As already mentioned in the Introduction, shape gradients play a key role in shape optimization.

Shape optimization literature mostly deals with PDE constrained shape functionals that can be expressed as an integral on a subdomain $D \subset \Omega$ [1–8]. Here we consider

$$\mathcal{J}(\Omega) = \int_D j(u)\, \mathrm{d}\mathbf{x}, \tag{3}$$

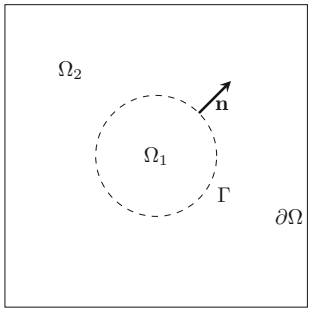


Fig. 1 Computational domain Ω of (4)

where $j : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuos function and u is the solution the scalar interface problem

$$\begin{cases} -\operatorname{div}(\sigma(\mathbf{x})\nabla u) = f & \text{in } \Omega = \Omega_1 \cup \Omega_2, \\ [\![u]\!] = 0 & \text{on } \Gamma, \\ \left[\!\left[\sigma \dfrac{\partial u}{\partial \mathbf{n}}\right]\!\right] = 0 & \text{on } \Gamma, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{4}$$

with real piecewise constant coefficient

$$\sigma(\mathbf{x}) := \sigma_1 \chi_{\Omega_1}(\mathbf{x}) + \sigma_2 \chi_{\Omega_2}(\mathbf{x}).$$

The jump symbol $[\![\cdot]\!]$ denotes discontinuity across the interface Γ. Note that for the Neumann jump the vector $\mathbf{n}$ points outward, see Fig. 1.

The shape gradient of shape differentiable PDE constrained shape functionals can be expressed both as an integration in volume and as an integration on the boundary (the latter as a result of the *Hadamard–Zolésio structure theorem* [7, Chap. 9, Theorem 3.6]). For instance, the shape gradient of (3) under the constraint (4) takes the forms[1]

$$\begin{aligned} d\mathcal{J}(\Omega; \mathcal{V}) = \int_\Omega \Big(& \sigma \nabla u \cdot (D\mathcal{V} + D\mathcal{V}^T)\nabla p + p\nabla f \cdot \mathcal{V} \\ & + \operatorname{div}(\mathcal{V})\,(j(u) - \sigma \nabla u \cdot \nabla p + fp) \Big)\, \mathrm{d}\mathbf{x} \end{aligned} \tag{5}$$

and

$$d\mathcal{J}(\Omega; \mathcal{V}) = \int_\Gamma (\mathcal{V} \cdot \mathbf{n}) \left[\!\!\left[2\sigma \frac{\partial p}{\partial \mathbf{n}} \frac{\partial u}{\partial \mathbf{n}} - \sigma \nabla u \cdot \nabla p \right]\!\!\right] \mathrm{d}S, \tag{6}$$

[1] We tacitly assume that the vector field $\mathcal{V}$ vanishes on $\partial\Omega$ because we are mostly interested in the contribution of the interface.

where p is the solution of the adjoint problem

$$\begin{cases} -\operatorname{div}(\sigma(\mathbf{x})\nabla p) = j'(u)\chi_D & \text{in } \Omega, \\ [\![p]\!] = 0 & \text{on } \Gamma, \\ \left[\!\left[\sigma \dfrac{\partial p}{\partial \mathbf{n}}\right]\!\right] = 0 & \text{on } \Gamma, \\ p = 0 & \text{on } \partial\Omega. \end{cases} \tag{7}$$

Remark 1 Deriving explicit formulas of shape gradients is a delicate and error prone task. Among the several techniques available in literature, the so-called "fast derivation" method of Céa provides a formal shortcut to find the boundary based formulation, cf. [1, Chap. 6.4.3] and [11]. However, great care has to be taken with interface problems. In this case it is worth working out the details in order to overcome the subtle issues induced by the presence of the interface. A thorough derivation of (5) and (6) can be found in [5].

3 Approximation of Shape Gradients

The shape gradient $d\mathcal{J}(\Omega; \mathcal{V})$ of (3) depends on the solution of the two IPs (4) and (7). To better stress this dependency, as well as to distinguish between Formulas (5) and (6), we refer to them with the notation $d\mathcal{J}(\Omega, u, p; \mathcal{V})^{\text{Vol}}$ and $d\mathcal{J}(\Omega, u, p; \mathcal{V})^{\text{Bdry}}$, respectively.

Lemma 1 *Let u and p be exact solutions of (4) and (7), respectively. Then, the following equality holds*

$$d\mathcal{J}(\Omega, u, p; \mathcal{V})^{\text{Vol}} = d\mathcal{J}(\Omega, u, p; \mathcal{V})^{\text{Bdry}}. \tag{8}$$

Proof Integration by parts on Formula (5) yields

$$\begin{aligned} d\mathcal{J}(\Omega; \mathcal{V}) = \int_\Omega \Big(& \sigma \nabla u \cdot (D\mathcal{V} + D\mathcal{V}^T)\nabla p \\ & - \mathcal{V} \cdot \left(j'(u)\nabla u - \sigma \nabla(\nabla u \cdot \nabla p) + f \nabla p \right) \Big) \, \mathrm{d}\mathbf{x} \\ + \int_\Gamma & [\![\mathcal{V} \cdot \mathbf{n} \, (j(u) - \sigma \nabla u \cdot \nabla p + fp)]\!] \, \mathrm{d}S. \end{aligned} \tag{9}$$

With the vector calculus identity [8, Eq. (44)]

$$\nabla u \cdot (D\mathcal{V} + D\mathcal{V}^T)\nabla p + \mathcal{V} \cdot \nabla(\nabla u \cdot \nabla p) = \nabla p \cdot \nabla(\mathcal{V} \cdot \nabla u) + \nabla u \cdot \nabla(\mathcal{V} \cdot \nabla p), \tag{10}$$

Formula (9) can be rewritten as

$$d\mathcal{J}(\Omega;\mathcal{V}) = \int_\Omega \Big(\sigma\nabla p\cdot\nabla(\mathcal{V}\cdot\nabla u) + \sigma\nabla u\cdot\nabla(\mathcal{V}\cdot\nabla p) - j'(u)\mathcal{V}\cdot\nabla u - f\mathcal{V}\cdot\nabla p\Big)\,\mathrm{d}\mathbf{x} + \int_\Gamma [\![\mathcal{V}\cdot\mathbf{n}\,(j(u)-\sigma\nabla u\cdot\nabla p + fp)]\!]\,\mathrm{d}S. \tag{11}$$

Then, integration by parts yields

$$d\mathcal{J}(\Omega;\mathcal{V}) = \int_\Gamma \left[\!\!\left[\sigma\frac{\partial p}{\partial\mathbf{n}}\mathcal{V}\cdot\nabla u\right]\!\!\right] - \int_\Omega \mathrm{div}(\sigma\nabla p)(\mathcal{V}\cdot\nabla u) + j'(u)(\mathcal{V}\cdot\nabla u)\,\mathrm{d}\mathbf{x} + \int_\Gamma \left[\!\!\left[\sigma\frac{\partial u}{\partial\mathbf{n}}\mathcal{V}\cdot\nabla p\right]\!\!\right] - \int_\Omega \mathrm{div}(\sigma\nabla u)(\mathcal{V}\cdot\nabla p) + f(\mathcal{V}\cdot\nabla p)\,\mathrm{d}\mathbf{x} + \int_\Gamma [\![\mathcal{V}\cdot\mathbf{n}\,(j(u)-\sigma\nabla u\cdot\nabla p + fp)]\!]\,\mathrm{d}S. \tag{12}$$

The two domain integrals in (12) vanish because of (4) and (7). Moreover, since $[\![u]\!] = 0$ on Γ,

$$\left[\!\!\left[\sigma\frac{\partial p}{\partial\mathbf{n}}\mathcal{V}\cdot\nabla u\right]\!\!\right] = \mathcal{V}\cdot\mathbf{n}\left[\!\!\left[\sigma\frac{\partial p}{\partial\mathbf{n}}\frac{\partial u}{\partial\mathbf{n}}\right]\!\!\right] \quad\text{and}\quad [\![\mathcal{V}\cdot\mathbf{n}j(u)]\!] = 0,$$

and since $[\![p]\!] = 0$, $[\![\mathcal{V}\cdot\mathbf{n}fp]\!] = 0$, so that we retrieve

$$d\mathcal{J}(\Omega;\mathcal{V}) = \int_\Gamma \mathcal{V}\cdot\mathbf{n}\left[\!\!\left[2\sigma\frac{\partial p}{\partial\mathbf{n}}\frac{\partial u}{\partial\mathbf{n}} - \sigma\nabla u\cdot\nabla p\right]\!\!\right]\mathrm{d}S. \tag{6}$$

□

Remark 2 For $d\mathcal{J}(\Omega,u,p;\mathcal{V})^{\mathrm{Vol}}$ to be well-defined, it is sufficient to assume that $u,\ p\in H^1(\Omega)$. On the other hand, higher regularity of u and p is required for $d\mathcal{J}(\Omega,u,p;\mathcal{V})^{\mathrm{Bdry}}$ to be well-defined because the latter is not continuous on $H^1(\Omega)$.

Usually, exact solutions of IPs are not available, and one has to rely on numerical approximations $u_h,\ p_h\in W^{1,\infty}(\Omega)$. Equality (8) breaks down when u and p are replaced with their approximate counterparts [8], and both formulas (5) and (6) become approximations

$$d\mathcal{J}(\Omega,u_h,p_h;\mathcal{V})^{\mathrm{Vol}} \approx d\mathcal{J}(\Omega;\mathcal{V}) \approx d\mathcal{J}(\Omega,u_h,p_h;\mathcal{V})^{\mathrm{Bdry}} \tag{13}$$

of the exact value $d\mathcal{J}(\Omega;\mathcal{V})$. The natural question is then which among $d\mathcal{J}(\Omega,u_h,p_h;\cdot)^{\mathrm{Vol}}$ and $d\mathcal{J}(\Omega,u_h,p_h;\cdot)^{\mathrm{Bdry}}$ is closer to $d\mathcal{J}(\Omega;\cdot)$.

The answer may depend on the underlying discretization scheme. Although discretization by boundary element method is also possible [2, 4], we focus on discretizations by means of finite elements. This is the most popular choice in shape optimization because of its flexibility, which is much appreciated among engineers.

In applied mathematics several operators that depend on the solution of boundary value problems have equivalent volume and boundary based representations. For instance, this is the case for lift functionals for potential flow [12] and for far field functionals in electromagnetism [13, 14]. When used in the context of finite element approximations, volume based formulations tend to exhibit faster convergence and superior accuracy than their counterparts formulated on the boundary. This can be motivated by volume integrals being continuous in energy norm, whilst boundary integrals involve traces that are not well-defined on the natural variational space. This difference determines whether the formulation displays the superconvergence that holds for the evaluation of continuous functionals on Galerkin solutions [15, Sect. 2].

On account of Remark 2, we heuristically expect the same trend in (13). A rigorous statement can be made in case of smooth interfaces and sufficient regular source function in (4). Following the same lines as for the proofs of Theorems 3.1 and 3.2 in [10], it can be shown that[2]

$$|d\mathcal{J}(\Omega;\mathcal{V}) - d\mathcal{J}(\Omega,u_h,p_h;\mathcal{V})^{\mathrm{Vol}}| = Ch^2\|\mathcal{V}\|_{W^{2,4}(\mathbb{R}^d;\mathbb{R}^d)} \tag{14}$$

and that

$$|d\mathcal{J}(\Omega;\mathcal{V}) - d\mathcal{J}(\Omega,u_h,p_h;\mathcal{V})^{\mathrm{Bdry}}| = Ch\|\mathcal{V}\|_{L^\infty(\mathbb{R}^d;\mathbb{R}^d)}, \tag{15}$$

when u_h and p_h are Ritz-Galerkin solutions computed with piecewise linear Lagrangian finite elements on a family of quasi-uniform triangular meshes with nodal basis functions.

Remark 3 The result (14) is restricted to vector fields in $W^{2,4}(\mathbb{R}^d;\mathbb{R}^d)$ because the proof relies on finite element duality techniques [16, Chap. 5.7]. However, the volume based formulation (5) is a continuous linear operator with respect to $W^{1,\infty}(\mathbb{R}^d;\mathbb{R}^d)$, and it can easily be shown that

$$|d\mathcal{J}(\Omega;\mathcal{V}) - d\mathcal{J}(\Omega,u_h,p_h;\mathcal{V})^{\mathrm{Vol}}| = Ch\|\mathcal{V}\|_{W^{1,\infty}(\mathbb{R}^d;\mathbb{R}^d)}. \tag{16}$$

On the other hand, the estimate (15) relies on the nontrivial approximation properties of finite element solutions in $W^{1,\infty}(\Omega)$ [16, Corollary 8.1.12]. We are not aware of a technique to improve the rate in (15) by restricting the space of vector fields.

[2] We denote by C a generic constant, which may depend on Ω, its discretization, the source function f, and the coefficient σ. Its value may differ between different occurrences.

4 Numerical Experiments

We consider the quadratic shape functional

$$\mathcal{J}(\Omega) = \int_{\Omega} u^2 \, d\mathbf{x}.$$

The shape gradient is a linear bounded operator on $W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$. Hence, the quality of the approximation in (13) should be investigated in the operator norm. Numerically, this is an extremely challenging task, if not impossible. Therefore we have to content ourself with considering convergence with respect to a more tractable operator norm over a finite dimensional space of vector fields.

Since we are mainly interested in contributions of the interface, we select vector fields that vanish on $\partial\Omega$. We set $\Omega =]-2, 2[^2$ (a square centered in the origin and with side equal 4), and we restrict ourself to the finite dimensional space of vector fields of the form[3]

$$\mathcal{V}(x, y) = \sum_{\substack{m_1+n_1\leq 5 \\ m_2+n_2\leq 5 \\ m_1,m_2,n_1,n_2\geq 1}} \lambda_{m_1,n_1} \begin{pmatrix} v(x, y, m_1, n_1) \\ 0 \end{pmatrix} + \lambda_{m_2,n_2} \begin{pmatrix} 0 \\ v(x, y, m_2, n_2) \end{pmatrix}$$

with $v(x, y, m, n) = \sin(mx\pi/2)\sin(ny\pi/2)$ and $\lambda_{m_i,n_i} \in \mathbb{R}$. Moreover, we replace the $W^{1,\infty}$-norm with the more manageable H^1-norm.

To investigate the convergence, we monitor the approximate dual norms

$$\mathrm{err}^{\mathrm{Vol}} := \left(\max_{\mathcal{V}} \frac{1}{\|\mathcal{V}\|^2_{H^1(\Omega)}} |d\mathcal{J}(\Omega; \mathcal{V}) - d\mathcal{J}(\Omega, u_h, p_h; \mathcal{V})^{\mathrm{Vol}}|^2 \right)^{1/2} \tag{17}$$

and

$$\mathrm{err}^{\mathrm{Bdry}} := \left(\max_{\mathcal{V}} \frac{1}{\|\mathcal{V}\|^2_{H^1(\Omega)}} |d\mathcal{J}(\Omega; \mathcal{V}) - d\mathcal{J}(\Omega, u_h, p_h; \mathcal{V})^{\mathrm{Bdry}}|^2 \right)^{1/2} \tag{18}$$

on different meshes generated through uniform refinement.[4] The reference value $d\mathcal{J}(\Omega; \mathcal{V})$ is approximated by evaluating both $d\mathcal{J}(\Omega, u_h, p_h; \mathcal{V})^{\mathrm{Vol}}$ and $d\mathcal{J}(\Omega, u_h, p_h; \mathcal{V})^{\mathrm{Bdry}}$ on a mesh with an extra level of refinement. To avoid biased results we display convergence history both with self- and cross-comparison.

[3] Repeating the experiments for $m_i + n_i \leq 3$ produces results in agreement with the observations made for $m_i + n_i \leq 5$. Therefore, the arbitrary choice of restricting the sum of the indices to 5 does not seem to compromise our observations.

[4] In experiment 1 new meshes are always adjusted to fit the curved interface.

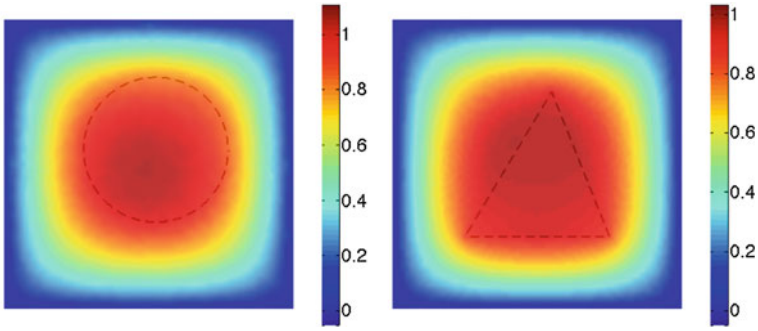


Fig. 2 Plot of the solution u of the state problem in the computational domain Ω for the **first** (*left*) and the **second** (*right*) **numerical experiment**. The interface is drawn with a *dashed line*

As in [10], we consider finite element discretizations based on linear Lagrangian finite elements on quasi-uniform triangular meshes with nodal basis functions.[5] Integrals in the domain are computed by 7 point quadrature rule in each triangle, while line integrals by 6 point Gauss quadrature on each segment. In experiment 1, the interface is approximated by a polygon. Nevertheless, the convergence of linear finite elements is not affected by this discretization [17].

In the **first numerical experiment** the interface Γ is a circle centered in $(0.1, 0.2)$ and with radius equal 1, see Fig. 2 (left). The problem data are

$$f(\mathbf{x}) = 1 \quad \text{and} \quad \sigma(\mathbf{x}) = 2\chi_{\Omega_1}(\mathbf{x}) + 1\chi_{\Omega_2}(\mathbf{x}). \tag{19}$$

The numerical results are displayed in Fig. 3 (left column). We clearly see that the volume based formulation converges faster and is more accurate than its boundary based counterpart. The convergence rates agree with what has been predicted by (14) and (15). In the cross-comparison plot $d\mathcal{J}(\Omega, u_h, p_h; \mathcal{V})^{\text{Vol}}$ saturates due to insufficient accuracy of the reference solution computed with $d\mathcal{J}(\Omega, u_h, p_h; \mathcal{V})^{\text{Bdry}}$, whereas the boundary based formulation converges with the same rate as for the self-comparison.

In the **second numerical experiment** the interface Γ is a triangle with corners located at $(-1, -1)$, $(1, -1)$ and $(0.2, 1)$, see Fig. 2 (right). Interface corners are known to affect the regularity of the solution of interface problems [18]. Therefore, the estimates (14) and (15) can not be proved in this case, and we expect to observe lower convergence rates. To better stress the impact of the corners we increase the contrast of the diffusion coefficient by setting

$$\sigma(\mathbf{x}) = 10\chi_{\Omega_1}(\mathbf{x}) + 1\chi_{\Omega_2}(\mathbf{x}).$$

[5] The experiments are performed in MATLAB and are based on the library LehrFEM developed at ETHZ. Mesh generation and uniform refinement are performed with the functions `initmesh` and `refinemesh` of the MATLAB PDE Toolbox.

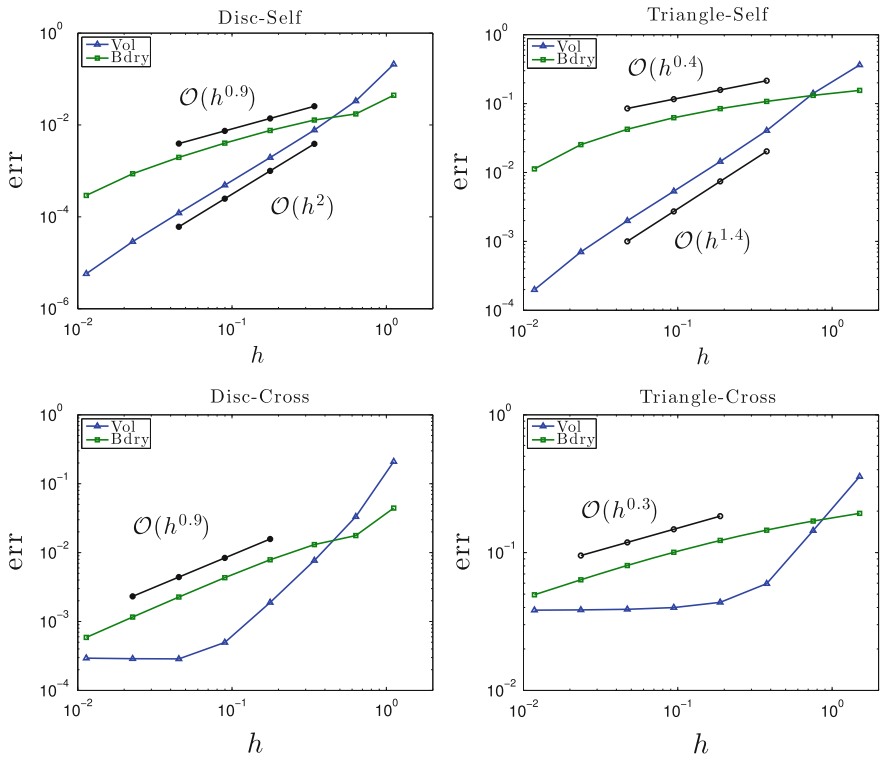


Fig. 3 Convergence history for the **first** (*left column*) and the **second** (*right column*) **numerical experiment**. In the first row the reference value $d\mathcal{J}(\Omega;\mathcal{V})$ is computed with an extra level of refinement. The second row displays cross-comparisons

The source function is the same as in (19). From the results displayed in Fig. 3 (right column) we observe that the volume based formulation converges faster and is more accurate then its boundary based counterpart. Again, in the cross-comparison the convergence history of the volume based formulation saturates due to an insufficient accuracy of the reference solution computed with $d\mathcal{J}(\Omega, u_h, p_h;\mathcal{V})^{\mathrm{Vol}}$. We suspect that this inaccuracy gives rise to the difference in the convergence rates of the boundary based formulation between self- and cross-comparisons.

In the **third numerical experiment** we investigate the impact of the choice of the diffusion coefficient σ on the results obtained in the **first** and in the **second numerical experiment**. For $\sigma_2 = 1$ fixed and $\sigma_1 = 0.1, 0.5, 0.8, 1.25, 2, 10$, we monitor the approximate relative error constructed by dividing the approximate dual norms (17) and (18) by

$$\max_{\mathcal{V}} \frac{|d\mathcal{J}(\Omega;\mathcal{V})|}{\|\mathcal{V}\|_{H^1(\Omega)}}\,.$$

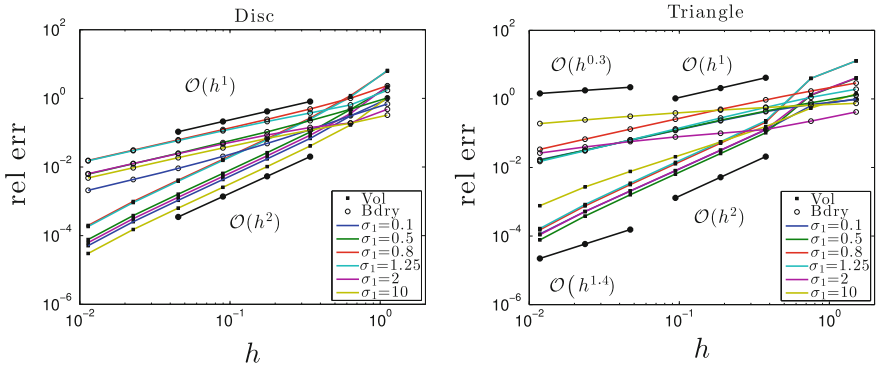


Fig. 4 Convergence history for the **third numerical experiment**. The choice of the diffusion coefficient has no influence on the convergence rates in case of a circular interface (*left*). On the other hand, for a triangular interface (*right*), the effect of the singularity in the functions u and p is visible only for high contrasts σ_1/σ_2

The reference solution is computed evaluating $d\mathcal{J}(\Omega, u, p; \mathcal{V})^{\mathrm{Vol}}$ on a mesh with an extra level of refinement. In Fig. 4 (left), we see that the choice of the diffusion coefficient σ has no influence on the convergence rates in case of a circular interface. On the other hand, for non-smooth interfaces, the effect of the singularity in the functions u and p is visible only for high contrasts σ_1/σ_2.

5 Conclusion

The shape gradient of shape differentiable PDE constrained shape functionals is a linear bounded operator on $W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$, and its knowledge is the starting point for gradient based shape optimization. The shape gradient can be stated both as an integration in volume and as an integration on the boundary, both of which depend on the solution of boundary value problems. When used with discrete solutions, these two representations lose their equivalence and become approximations of $d\mathcal{J}(\Omega; \cdot)$. Theoretical considerations in Sect. 3 and numerical experiments in Sect. 4 convey that volume based approximations of the shape gradient are better suited in the context of finite element discretizations. Although our investigations are conducted on a chosen class of scalar interface problems, we believe that similar conclusions can be drawn for the case of more general PDE constraints stemming from electromagnetism and continuum mechanics.

References

1. G. Allaire, *Conception optimale de structures* (Springer, Berlin, 2007)
2. R. Udawalpola, E. Wadbro, M. Berggren, Optimization of a variable mouth acoustic horn. Int. J. Numer. Methods Eng. **85**(5), 591–606 (2011)
3. E. Bängtsson, D. Noreland, M. Berggren, Shape optimization of an acoustic horn. Comput. Methods Appl. Mech. Eng. **192**(11–12), 1533–1571 (2003)
4. K. Eppler, H. Harbrecht, Coupling of FEM and BEM in shape optimization. Numer. Math. **104**(1), 47–68 (2006)
5. A. Laurain, K, Sturm, Domain expression of the shape derivative and application to electrical impedance tomography. Technical report 1863, Weierstrass Institute for Applied Analysis and Stochastics (2013)
6. O. Pironneau, *Optimal Shape Design for Elliptic Systems* (Springer, Berlin, 1984)
7. M.C. Delfour, J.-P. Zolésio, *Shapes and Geometries. Metrics, Analysis, Differential Calculus, and Optimization*. 2nd edn., Society for Industrial and Applied Mathematics (SIAM) (2011)
8. M. Berggren, A unified discrete-continuous sensitivity analysis method for shape optimization. *Applied and Numerical Partial Differential Equations* (Springer, Berlin, 2010), pp. 25–39
9. E.J. Haug, K.K. Choi, V. Komkov, *Design Sensitivity Analysis of Structural Systems* (Academic Press Inc., 1986)
10. R. Hiptmair, A. Paganini, S. Sargheini, Comparison of approximate shape gradients. BIT Numer. Math. **55**(2), 459–485 (2015)
11. J. Céa, Conception optimale ou identification de formes: calcul rapide de la dérivée directionnelle de la fonction coût. RAIRO Modél. Math. Anal. Numér. **20**(3), 371–402 (1986)
12. H. Harbrecht, On output functionals of boundary value problems on stochastic domains. Math. Methods Appl. Sci. **33**(1), 91–102 (2010)
13. P. Monk, *Finite Element Methods for Maxwell's Equations* (Clarendon Press, 2003)
14. P. Monk, E. Süli, The adaptive computation of far-field patterns by a posteriori error estimation of linear functionals. SIAM J. Numer. Anal. **36**(1), 251–274 (1999)
15. R. Becker, R. Rannacher, An optimal control approach to a posteriori error estimation in finite element methods. Acta Numer. **10**, 1–102 (2001)
16. S.C. Brenner, L.R. Scott, *The Mathematical Theory of Finite Element Methods*, 3rd edn. (Springer, Berlin, 2008)
17. J. Li, J.M. Melenk, B. Wohlmuth, J. Zou, Optimal a priori estimates for higher order finite elements for elliptic interface problems. Appl. Numer. Math. **60**(1–2), 19–37 (2010)
18. M. Blumenfeld, The regularity of interface-problems on corner-regions. *Singularities and Constructive Methods for their Treatment (Oberwolfach, 1983)* (Springer, Berlin, 1985), pp. 38–54

Towards a Lagrange–Newton Approach for PDE Constrained Shape Optimization

Volker H. Schulz, Martin Siebenborn and Kathrin Welker

Abstract The novel Riemannian view on shape optimization developed in [27] is extended to a Lagrange–Newton approach for PDE constrained shape optimization problems. The extension is based on optimization on Riemannian vector space bundles and exemplified for a simple numerical example.

Keywords Shape optimization · Riemannian manifold · Newton method

1 Introduction

Shape optimization problems arise frequently in technological processes, which are modeled in the form of partial differential equations as in [2–4, 13, 14, 24–26]. In many practical circumstances, the shape under investigation is parametrized by finitely many parameters, which on the one hand allows the application of standard optimization approaches, but on the other hand limits the space of reachable shapes unnecessarily. Shape calculus, which has been the subject of several monographs [12, 22, 29] presents a way out of that dilemma. However, so far it is mainly applied in the form of gradient descent methods, which can be shown to converge. The major difference between shape optimization and the standard PDE constrained optimization framework is the lack of the linear space structure in shape spaces. If one cannot use a linear space structure, then the next best structure is the Riemannian manifold structure as discussed for shape spaces in [5, 6, 19–21]. The publication [27] makes a link between shape calculus and shape manifolds and thus enables the usage of optimization techniques on manifolds in the context of shape optimization.

V.H. Schulz (✉) · M. Siebenborn · K. Welker
Department of Mathematics, University of Trier, 54296 Trier, Germany
e-mail: volker.schulz@uni-trier.de

M. Siebenborn
e-mail: siebenborn@uni-trier.de

K. Welker
e-mail: welker@uni-trier.de

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_10

PDE constrained shape optimization however, is confronted with function spaces defined on varying domains. The current paper presents a vector bundle framework based on the Riemannian framework established in [27], which enables the discussion of Lagrange–Newton methods within the shape calculus framework for PDE constrained shape optimization.

The paper first presents the novel Riemannian vector bundle framework on Sect. 2, discusses this approach for a specific academic example in Sect. 3 and presents numerical results in Sect. 4.

2 Constrained Riemannian Shape Optimization

The typical set-up of an equality constrained optimization problem is

$$\begin{aligned} &\min_{y,u} J(y,u)\,,\ J\colon Y\times U\to\mathbb{R}\\ &\text{s.t. } c(y,u)=0\,,\ c\colon Y\times U\to Z \end{aligned}$$

where U, Y, Z are linear spaces and c, J sufficiently smooth nonlinear functions [8]. In some situations the constraint c allows to apply the implicit function theorem in order the define a unique control to state mapping $y(u)$ and thus the constrained problem maybe reduced to an unconstrained one of the form

$$\min_u J(y(u),u).$$

However, the constrained formulation is often computationally advantageous, because it allows the usage of pre-existing solver technology for the constraint and it is geared towards an efficient SAND (simultaneous analysis and design) or one–shot approach based on linear KKT systems. So far, shape optimization methods based on the shape calculus, have been mainly considered with the reduced black-box framework above via the implicit function theorem—mainly because the set of all admissible shapes is typically not a linear space—unlike the space U above. The publication [27] has developed a Riemannian framework for shape optimization in the reduced unconstrained paradigm, which enables Newton—like iteration techniques and convergence results. This publication aims at generalizing those results to the constrained perspective—in particular for the case that the constraint is of the form of a set of partial differential equations (PDE).

Within that framework, the space Y for the state variable is a linear (function) space depending explicitly on $u \in U$, e.g., $H^1(\Omega(u))$, where $\Omega(u)$ is the interior of a shape u. This line of thinking leads to vector bundles of function spaces as discussed in detail in [18]. Thus, we now consider a Riemannian manifold $(\mathcal{N}, G)$ of class C^q $(q \geq 0)$, where G is a smooth mapping assigning any point $p \in \mathcal{N}$ an inner product $G_p(\cdot,\cdot)$ on the tangential bundle $T\mathcal{N}$. For each $u \in \mathcal{N}$, there is given a Hilbert space $H(u)$ such that the set

$$E := \{(H(u), u) \mid u \in \mathcal{N}\}$$

is the total space of a vector bundle $(E, \pi, \mathcal{N})$. In particular, there is a bundle—projection $\pi : E \to \mathcal{N}$ and for an open covering $\{U_i\}$ of $\mathcal{N}$ a local C^q isomorphism

$$\tau_i \colon \pi^{-1}(U_i) \to H_0 \times U_i$$

where H_0 is a Hilbert space. In particular, we have an isomorphism on each fiber

$$\tau_i(u) \colon \pi^{-1}(x) = H(u) \to H_0$$

and for $u \in U_i \cap U_j$, the mapping $\tau_i(u) \circ \tau_j(u)^{-1} : H_0 \to H_0$ is a linear isomorphism. The total space E of the vector bundle $(E, \pi, \mathcal{N})$ is by itself a Riemannian manifold, where the tangential bundle TE satisfies

$$T_{(y,u)}E \cong H(u) \times T_y\mathcal{N}.$$

In Riemannian geometry, tangential vectors are considered as first order differential operators acting on germs of scalar valued functions (e.g. [10]). Such a differential operator will be notated by $h(J)(e)$, if $J \colon E \to \mathbb{R}$ is a differentiable function and $e \in E$. We will have to deal with derivatives, where we will always use directional derivatives of scalar valued functions only, but notate them in the usual fashion. Let the derivative of J at e in direction $h \in TE$ be denoted by $DJ(e)h$. Then, we define in this setting

$$DJ(e)h := h(J)(e)\,,\ h \in TE.$$

In particular, we denote

$$\frac{\partial}{\partial y} J(y,u)h_y := h_1(J)(y,u)\,,\ h_1 := (h_y, 0) \in TE$$
$$\frac{\partial}{\partial u} J(y,u)h_u := h_2(J)(y,u)\,,\ h_2 := (0, h_u) \in TE$$

where $h_y \in H(u)$ and $h_u \in T_y\mathcal{N}$, if $h_1, h_2 \in T_{(y,u)}E$.

We consider now the following constrained optimization problem

$$\min_{(y,u)\in E} J(y,u)\,,\ J \colon E \to \mathbb{R} \tag{1}$$
$$\text{s.t. } a_u(y,p) = b_u(p)\,,\ \forall p \in H \tag{2}$$

where $a_u(.,.)$ is a bilinear form and $b_u(.)$ a linear form defined on the fiber H which are C^q with respect to u. The scalar valued function J is assumed to be C^q. Intentionally, the weak formulation of the PDE is chosen for ease of presentation. Now, it will be necessary to define the Lagrangian $\mathscr{L}$ in order to formulate the adjoint and design equation to the constrained optimization problem (1–2).

Definition 1 We define the Lagrangian in the setting above for $(y, u, p) \in F$ as

$$\mathscr{L}(y, u, p) := J(y, u) + a_u(y, p) - b_u(p)$$

where $F := \{(H(u), u, H(u)) \mid u \in \mathcal{N}\}$ with $T_{(y,u,p)}F \cong H(u) \times T_y\mathcal{N} \times H(u)$.

Let $(\hat{y}, \hat{u}) \in E$ solves the optimization problem (1–2). Then, the (adjoint) variational problem which we get by differentiating $\mathscr{L}$ with respect to y is given by

$$a_{\hat{u}}(z, p) = -\frac{\partial}{\partial y} J(\hat{y}, \hat{u})z\,, \ \forall z \in H(\hat{u}) \tag{3}$$

and the design problem which we get by differentiating $\mathscr{L}$ with respect to u is given by

$$\frac{\partial}{\partial u}\bigg|_{u=\hat{u}} \left[J(\hat{y}, u) + a_u(\hat{y}, \hat{p}) - b_u(\hat{p})\right] w = 0\,, \ \forall w \in T_{\hat{u}}\mathcal{N} \tag{4}$$

where $\hat{p} \in H$ solves (3). If we differentiate $\mathscr{L}$ with respect to p, we get the state equation (2). These (KKT) conditions (2–4) could be collected in the following condition:

$$D\mathscr{L}(\hat{y}, \hat{u}, \hat{p})h = 0\,, \ \forall h \in T_{(y,u,p)}F. \tag{5}$$

Remark 1 In a vector space setting, the existence of a solution $p \in H$ of the (adjoint) variational problem (3) is typically guaranteed by so-called constraint qualifications. From this point of view, here, the existence itself can be interpreted as formulation of a constraint qualification.

By using a Riemannian metric G on $T\mathcal{N}$ and a smoothly varying scalar product $\langle ., . \rangle_u$ on the Hilbert space $H(u)$, we can envision $T_{(y,u,p)}F$ as a Hilbert space with a canonical scalar product

$$\left\langle \begin{pmatrix} z_1 \\ w_1 \\ q_1 \end{pmatrix}, \begin{pmatrix} z_2 \\ w_2 \\ q_2 \end{pmatrix} \right\rangle_{T_{(y,u,p)}F} := \langle z_1, z_2 \rangle_u + G_u(w_1, w_2) + \langle q_1, q_2 \rangle_u \tag{6}$$

and thus also $(F, \langle ., . \rangle_{TF})$ as Riemannian manifold. This scalar product can be used to apply the Riesz representation theorem in order to define the gradient of the Lagrangian $\operatorname{grad}\mathscr{L} \in TF$ by the condition

$$\langle \operatorname{grad}\mathscr{L}, h \rangle_{T_{(y,u,p)}F} := D\mathscr{L}(y, u, p)h\,, \ \forall h \in T_{(y,u,p)}F.$$

Now, similar to standard nonlinear programming we can solve the problem of finding $(y, u, p) \in F$ with

$$\operatorname{grad}\mathscr{L}(y, u, p) = 0 \tag{7}$$

as a means to find solutions to the optimization problem (1–2). The nonlinear problem (7) has exactly the form of the root finding problems discussed in [27]. Exploiting the Riemannian structure on TF, we can formulate a Newton iteration involving the Riemannian Hessian which is based on the resulting Riemannian connection:

kth iteration:

(1) compute increment $\Delta\xi$ as solution of

$$Hess\mathscr{L}(\xi^k)\Delta\xi = -\text{grad}\mathscr{L}(\xi^k) \tag{8}$$

(2) increment $\xi^{k+1} := \exp_{\xi^k}(\alpha^k \cdot \Delta\xi)$ for some steplength α^k

This iteration will be detailed out below. However, before that, we have to specify the scalar product on the Hilbert space involved. Since we will use the exponential map based on the Riemannian metric on F, we would like to choose a metric that is in the Hilbert space parts as simple as possible. Therefore we use the metric defined on the Hilbert space $(H_0, \langle ., . \rangle_0)$ and transfer that canonically to the Hilbert spaces $H(u)$. Thus, we assume now that in the sequel we only have to deal with one particular chart (U_i, τ_i) from the covering $\{U_i\}$ and define there

$$\langle z_1, z_2 \rangle_u := \langle \tau_i(u)z_1, \tau_i(u)z_2 \rangle_0 \ , \ \forall u \in U_i \, .$$

Now, geodesics in the Hilbert space parts of F are represented just by straight lines in H_0 and the exponential map can be expressed in the form

$$\begin{aligned} &\exp_{(y,u,p)}(z, w, q) \\ &= \left(\tau_i\left(\exp_u^{\mathcal{N}}(w)\right)^{-1} \circ \tau_i(u)(y+z), \exp_u^{\mathcal{N}}(w), \tau_i\left(\exp_u^{\mathcal{N}}(w)\right)^{-1} \circ \tau_i(u)(p+q)\right) \end{aligned}$$

where $\exp^{\mathcal{N}}$ denotes the exponential map on the manifold $\mathcal{N}$.

Within iteration (8), the Hessian has to be discussed. It is based on the Riemannian connection ∇ on F at $u \in \mathcal{N}$. The expression $\nabla_u^{\mathcal{N}}$ may denote the Riemannian covariant derivative on $T_u\mathcal{N}$. Since the scalar product in H is completely independent from the location $u \in \mathcal{N}$, we observe that mixed covariant derivatives of vectors from H with respect to tangential vectors in $T\mathcal{N}$ are reduced to simple directional derivatives—which is the case for derivatives in linear spaces anyway. Thus:

$$\begin{aligned} \nabla_{(h_y,h_u,h_p)} : T_{(y,u,p)}F &\to T_{(y,u,p)}F \\ \begin{pmatrix} z \\ w \\ q \end{pmatrix} &\mapsto \begin{pmatrix} \frac{\partial}{\partial y}z[h_y] + \frac{\partial}{\partial u}z[h_u] + \frac{\partial}{\partial p}z[h_p] \\ \frac{\partial}{\partial y}w[h_y] + \nabla_u^{\mathcal{N}}w[h_u] + \frac{\partial}{\partial p}w[h_p] \\ \frac{\partial}{\partial y}q[h_y] + \frac{\partial}{\partial u}q[h_u] + \frac{\partial}{\partial p}q[h_p] \end{pmatrix} \end{aligned}$$

From the definition of the Hessian as $Hess\mathscr{L}[h] := \nabla_h \text{grad}\mathscr{L}$ we conclude the following block structure of the Hessian:

$$Hess\mathscr{L} = \begin{pmatrix} D_y \mathrm{grad}_y \mathscr{L} & D_u \mathrm{grad}_y \mathscr{L} & D_p \mathrm{grad}_y \mathscr{L} \\ D_y \mathrm{grad}_u \mathscr{L} & \nabla_u^{\mathcal{N}} \mathrm{grad}_u \mathscr{L} & D_p \mathrm{grad}_u \mathscr{L} \\ D_y \mathrm{grad}_p \mathscr{L} & D_u \mathrm{grad}_p \mathscr{L} & 0 \end{pmatrix} \tag{9}$$

From a practical point of view, it may be advantageous to solve Eq. (8) in a weak formulation as

$$\nabla (D\mathscr{L}(y,u,p)h) \begin{pmatrix} z \\ w \\ q \end{pmatrix} = -D\mathscr{L}(y,u,p)h\,, \quad \forall h \in T_{(y,u,p)}F \tag{10}$$

i.e., in detail, the following equations have to be satisfied for all $h := (\bar{z}, \bar{w}, \bar{q})^T \in T_{(y,u,p)}F$:

$$H_{11}(z,\bar{z}) + H_{12}(w,\bar{z}) + H_{13}(q,\bar{z}) = -a_u(\bar{z},p) - \frac{\partial}{\partial y} J(y,u)\bar{z} \tag{11}$$

$$H_{21}(z,\bar{w}) + H_{22}(w,\bar{w}) + H_{23}(q,\bar{w}) = -\frac{\partial}{\partial u}\left[J(y,u) + a_u(y,p) - b_u(p)\right]\bar{w} \tag{12}$$

$$H_{31}(z,\bar{q}) + H_{32}(w,\bar{q}) + H_{33}(q,\bar{q}) = -a_u(y,\bar{q}) + b_u(\bar{q}) \tag{13}$$

where

$$\begin{aligned}
H_{11}(z,\bar{z}) &= \frac{\partial^2}{\partial y^2} J(y,u) z\bar{z} \\
H_{12}(w,\bar{z}) &= \frac{\partial}{\partial u}\left[a_u(\bar{z},p) + \frac{\partial}{\partial y} J(y,u)\bar{z}\right] w \\
H_{13}(q,\bar{z}) &= a_u(\bar{z},q) \\
H_{21}(z,\bar{w}) &= \frac{\partial}{\partial y}\frac{\partial}{\partial u}\left(\left[J(y,u) + a_u(y,p)\right]\bar{w}\right) z \\
H_{22}(w,\bar{w}) &= G\left(Hess^{\mathcal{N}}\left(J(y,u) + a_u(y,p) - b_u(p)\right) w, \bar{w}\right) \\
H_{23}(q,\bar{w}) &= \frac{\partial}{\partial u}\left[a_u(y,q) - b_u(q)\right]\bar{w} \\
H_{31}(z,\bar{q}) &= a_u(z,\bar{q}) \\
H_{32}(w,\bar{q}) &= \frac{\partial}{\partial u}\left[a_u(y,\bar{q}) - b_u(\bar{q})\right] w \\
H_{33}(q,\bar{q}) &= 0
\end{aligned}$$

One should note that the covariant derivative ∇ reveals natural symmetry properties and thus obvious symmetries can be observed in the components above not involving second shape derivatives. A key observation in [27] is that even the expression $H_{22}(w,\bar{w})$ is symmetric in the solution of the shape optimization problem. This motivates a shape—SQP method as outlined below, where away from the solution

only expressions in $H_{22}(w, \bar{w})$ are used which are nonzero at the solution. Its basis is the following observation:

If the term $H_{22}(w, \bar{w})$ is replaced by an approximation $\hat{H}_{22}(w, \bar{w})$, which omits all terms in $H_{22}(w, \bar{w})$, which are zero at the solution and if the reduced Hessian of (9) built with this approximation is coercive, equation (10) is equivalent to the linear–quadratic problem

$$\min_{(z,w)} \frac{1}{2}\Big(H_{11}(z, z) + 2H_{12}(w, z) + \hat{H}_{22}(w, w)\Big) + \frac{\partial}{\partial y}J(y, u)z + \frac{\partial}{\partial u}J(y, u)w \tag{14}$$

$$\text{s.t. } a_u(z, \bar{q}) + \frac{\partial}{\partial u}\big[a_u(y, \bar{q}) - b_u(\bar{q})\big]w = -a_u(y, \bar{q}) + b_u(\bar{q})\,, \ \forall \bar{q} \in H(u) \tag{15}$$

where the adjoint variable to the constraint (15) is just $p + q$. In the next sections, we also omit terms in H_{11} and H_{12}, which are zero, when evaluated at the solution of the optimization problems. Nevertheless, quadratic convergence of the resulting SQP method is to be expected and indeed observed in Sect. 4.

3 Discussion for a Poisson–type Model Problem

In this section, we apply the theoretical discussion of Sect. 2 to a PDE constrained shape optimization problem, which is inspired by the standard tracking—type elliptic optimal control problem and motivated by electrical impedance tomography. It is very close to the model problem of example 2 in [9] and the inverse interface problem in [17].

Let the domain $\Omega := (0, 1)^2 \subset \mathbb{R}^2$ split into the two subdomains $\Omega_1, \Omega_2 \subset \Omega$ such that $\Omega_1 \cup \Gamma \cup \Omega_2 = \Omega$ and $\partial\Omega_1 \cap \partial\Omega_2 = \Gamma$. The interface Γ is replaced by u and an element of the following manifold

$$B_e^0([0, 1], \mathbb{R}^2) := \mathrm{Emb}^0([0, 1], \mathbb{R}^2)/\mathrm{Diff}^0([0, 1])$$

i.e., an element of the set of all equivalence classes of the set of embeddings

$$\mathrm{Emb}^0([0, 1], \mathbb{R}^2) := \{\phi \in C^\infty([0, 1], \mathbb{R}^2) \,|\, \phi(0) = (0.5, 0), \phi(1) = (0.5, 1), \\ \phi \text{ injective immersion}\}$$

where the equivalence relation is defined by the set of all C^∞ re-parametrizations, i.e., by the set of all diffeomorphisms

$$\mathrm{Diff}^0([0, 1], \mathbb{R}^2) := \{\phi\colon [0, 1] \to [0, 1] \,|\, \phi(0) = (0.5, 0), \phi(1) = (0.5, 1), \\ \phi \text{ diffeomorphism}\}.$$

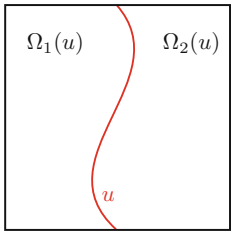


Fig. 1 Example of a domain $\Omega(u) = \Omega_1(u) \cup u \cup \Omega_2(u)$

In Fig. 1 the construction of the domain Ω from the interface $u \in B_e^0([0,1], \mathbb{R}^2)$ is illustrated. Now, we consider Ω dependent on u. Therefore, we denote it by $\Omega(u) = \Omega_1(u) \cup u \cup \Omega_2(u)$.

Remark 2 The manifold $B_e^0([0,1], \mathbb{R}^2)$ is constructed in analogy to the manifold $B_e(S^1, \mathbb{R}^2)$ in [20] as a set of equivalence classes in a set of embeddings with respect to a equivalence relation which is given by a set of diffeomorphisms. Moreover, a particular point on the manifold $B_e^0([0,1], \mathbb{R}^2)$ is represented by a curve $c\colon [0,1] \to \mathbb{R}^2$, $\theta \mapsto c(\theta)$. Because of the equivalence relation Diff([0, 1]), the tangent space is isomorphic to the set of all smooth vector fields along c, i.e.,

$$T_c B_e^0([0,1], \mathbb{R}^2) \cong \{h \,|\, h = \alpha n, \alpha \in C^\infty([0,1], \mathbb{R})\}$$

where n is the unit outer normal to $\Omega_1(u)$ at u. Thus, all considerations of [27] carry easily over to our manifold $B_e^0([0,1], \mathbb{R}^2)$.

The PDE constrained shape optimization problem is given in strong form by

$$\min_u \; J(y,u) \equiv \frac{1}{2}\int_{\Omega(u)} (y - \bar{y})^2 dx + \mu \int_u 1 ds \tag{16}$$

$$\text{s.t.} \; -\triangle y = f \quad \text{in } \Omega(u) \tag{17}$$

$$y = 0 \quad \text{on } \partial\Omega(u) \tag{18}$$

where

$$f \equiv \begin{cases} f_1 = \text{const.} & \text{in } \Omega_1(u) \\ f_2 = \text{const.} & \text{in } \Omega_2(u) \end{cases}. \tag{19}$$

The perimeter regularization with $\mu > 0$ in the objective (16) is a frequently used means to overcome ill—posedness of the optimization problem (e.g. [1]). Let n be the unit outer normal to $\Omega_1(u)$ at u. We observe that the unit outer normal to $\Omega_2(u)$ at u is equal to $-n$, which enables us to use only one normal n for the subsequent discussions. Furthermore, we have interface conditions at the interface u. We formulate explicitly the continuity of the state and of the flux at the boundary u as

$$[\![y]\!] = 0, \quad \left[\!\!\left[\frac{\partial y}{\partial n}\right]\!\!\right] = 0 \quad \text{on } u \tag{20}$$

where the jump symbol $[\![\cdot]\!]$ denotes the discontinuity across the interface u and is defined by $[\![v]\!] := v_1 - v_2$ where $v_1 := v\big|_{\Omega_1}$ and $v_2 := v\big|_{\Omega_2}$.

The boundary value problem (17–20) is written in weak form as

$$a_u(y, p) = b_u(p)\,, \ \forall p \in H_0^1(\Omega(u)) \tag{21}$$

where

$$a_u(y, p) := \int_{\Omega(u)} \nabla y^T \nabla p \, dx - \int_u \left[\!\!\left[\frac{\partial y}{\partial n} p\right]\!\!\right] ds \tag{22}$$

$$b_u(p) := \int_{\Omega(u)} fp \, dx. \tag{23}$$

Now, F from Definition 1 takes the specific form

$$F := \left\{(H_0^1(\Omega(u)), u, H_0^1(\Omega(u))) \mid u \in B_e^0([0, 1], \mathbb{R}^2)\right\}.$$

The metric in the vector space parts is constructed by employing a "mesh deformation". Mesh deformations are often used to deform a computational mesh smoothly in accordance with a deformation of the boundary of the computational domain. Here, we use this in the form of a deformation of the computational domain rather than of the mesh only and assume that there is a bijective C^∞-mapping

$$\Phi_u \colon [0, 1]^2 \to \Omega(u),$$

e.g., Φ_u is the deformation given by the solution of a linear elasticity problem. Thus, we can construct the necessary bijective identification

$$\tau(u) \colon H_0^1(\Omega(u)) \to H_0^1\left((0, 1)^2\right), g \mapsto g \circ \Phi_u.$$

We have to detail the expressions in Eq. (8) or respectively (10). First, the Lagrangian is defined for $(y, u, p) \in F$ as

$$\mathscr{L}(y, u, p) := J(y, u) + a_u(y, p) - b_u(p)$$

where $J(y, u)$ is defined in (16) and a_u, b_u are defined in (22, 23). Now, we focus on the shape derivative of $\mathscr{L}$ in direction of a continuous vector field V. It is defined by

$$\frac{\partial}{\partial u}\mathscr{L}(y, u, p)[V] := \lim_{t \to 0^+} \frac{\mathscr{L}(y, u_t, p) - \mathscr{L}(y, u, p)}{t} \tag{24}$$

if for all directions V this limit exists and the mapping $V \mapsto \frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V]$ is linear and continuous. The perturbed boundaries u_t in (24) are defined by

$$u_t := F_t(u) = \{F_t(x) \colon x \in u\} \text{ with } u_0 = u \tag{25}$$

where $F_t(x) := x + tV(x)$ denotes the perturbation of identity and $t \in [0,T]$ with $T > 0$.

Remark 3 One should note that we get perturbed domains Ω_t given by

$$\Omega_t := F_t(\Omega(u)) = \{F_t(x) \colon x \in \Omega(u)\} \text{ with } \Omega_0 = \Omega(u) \tag{26}$$

due to the perturbed boundaries u_t.

Remark 4 The perturbation of u or respectively $\Omega(u)$ could also be described by the velocity method, i.e., as the flow $F_t(x) := \xi(t,x)$ determined by the initial value problem

$$\begin{aligned} \frac{d\xi(t,x)}{dt} &= V(\xi(t,x)) \\ \xi(0,x) &= x \end{aligned} \tag{27}$$

instead of the perturbation of identity.

We first consider the objective J in (16) without perimeter regularization. Then the shape derivative $\frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V]$ can be expressed as an integral over the domain $\Omega(u)$ as well as an integral over the interface u which is better suited for a finite element implementation as already mentioned for example in [12, Remark 2.3, p. 531]. An important point to note here is that the shape derivative of our $\mathscr{L}$ evaluated in its saddle—point is equal to the one of J due to the theorem of Correa and Seeger [11, theorem 2.1]. Such a saddle—point is given by

$$\frac{\partial \mathscr{L}(\Omega,y,p)}{\partial y} = \frac{\partial \mathscr{L}(\Omega,y,p)}{\partial p} = 0 \tag{28}$$

which leads to the adjoint equation

$$-\triangle p = -(y - \overline{y}) \quad \text{in } \Omega(u) \tag{29}$$

$$p = 0 \quad \text{on } \partial\Omega(u) \tag{30}$$

$$[\![p]\!] = 0 \quad \text{on } u \tag{31}$$

$$\left[\!\left[\frac{\partial p}{\partial n}\right]\!\right] = 0 \quad \text{on } u \tag{32}$$

and to the state equation

$$-\triangle y = f \quad \text{in } \Omega(u). \tag{33}$$

Like in [28] we first deduce a representation of the shape derivative expressed as a domain integral which will later allow us to calculate the boundary expression of the shape derivative by means of integration by parts on the interface u. One should note however, that by the Hadamard structure theorem [29, Theorem 2.27] only the normal part of the continuous vector field has an impact on its value. Applying the following common rule for differentiating domain integrals

$$\frac{d^+}{dt}\left(\int_{\Omega_t}\eta(t)\right)\Big|_{t=0} = \int_{\Omega}(D_m\eta + \operatorname{div}(V)\eta) \tag{34}$$

which was proved in [15, Lemma 3.3] yields

$$\begin{aligned}\frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V] &= \lim_{t\to 0^+}\frac{\mathscr{L}(y,u_t,p)-\mathscr{L}(y,u,p)}{t} = \frac{d^+}{dt}\mathscr{L}(y,u_t,p)\Big|_{t=0}\\ &= \int_{\Omega(u)} D_m\left(\frac{1}{2}(y-\bar{y})^2\right) + D_m\left(\nabla y^T\nabla p\right) - D_m(fp)\\ &\quad + \operatorname{div}(V)\left(\frac{1}{2}(y-\bar{y})^2 + \nabla y^T\nabla p - fp\right)dx\\ &\quad - \int_u D_m\left(\left[\!\left[\frac{\partial y}{\partial n}p\right]\!\right]\right) + \operatorname{div}_u(V)\left[\!\left[\frac{\partial y}{\partial n}p\right]\!\right]ds\end{aligned} \tag{35}$$

where D_m denotes the material derivative with respect to $F_t = id + tV$ which is defined by

$$D_m\left(j(x)\right) := \lim_{t\to 0^+}\frac{\left(j\circ F_t\right)(x) - j(x)}{t} = \frac{d^+}{dt}\left(j\circ F_t\right)(x)\Big|_{t=0} \tag{36}$$

for a generic function $j\colon \Omega_t \to \mathbb{R}$. For the material derivative the product rule holds. Moreover, the following equality was proved in [7]

$$D_m\left(\nabla j\right) = \nabla\left(D_m(j)\right) - \nabla V^T\nabla j. \tag{37}$$

Combining (35), the product rule and (37) we obtain

$$\begin{aligned}\frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V] &= \int_{\Omega(u)}(y-\bar{y})D_m\left(y\right) + \nabla\left(D_m(y)\right)^T\nabla p + \nabla y^T\nabla\left(D_m\left(p\right)\right)\\ &\quad - \nabla y^T\left(\nabla V + \nabla V^T\right)\nabla p - D_m(f)p - fD_m(p)\\ &\quad + \operatorname{div}(V)\left(\frac{1}{2}(y-\bar{y})^2 + \nabla y^T\nabla p - fp\right)dx\\ &\quad - \int_u\left[\!\left[D_m\left(\frac{\partial y}{\partial n}\right)p + \frac{\partial y}{\partial n}D_m(p)\right]\!\right] + \operatorname{div}_u(V)\left[\!\left[\frac{\partial y}{\partial n}p\right]\!\right]ds.\end{aligned}$$

From this we get

$$\begin{aligned}\frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V] = \int_{\Omega(u)} &((y-\bar{y})-\triangle p)\, D_m(y) + (-\triangle y - f)\, D_m(p)\\ &- \nabla y^T\left(\nabla V + \nabla V^T\right)\nabla p - D_m(f)p\\ &+ \operatorname{div}(V)\left(\frac{1}{2}(y-\bar{y})^2 + \nabla y^T \nabla p - fp\right) dx\\ &+ \int_u \left[\!\!\left[\frac{\partial p}{\partial n} D_m(y) - D_m\left(\frac{\partial y}{\partial n}\right) p\right]\!\!\right]\\ &+ \operatorname{div}_u(V)\left[\!\!\left[\frac{\partial y}{\partial n} p\right]\!\!\right] ds. \end{aligned} \tag{38}$$

To deal with the term $D_m(f)p$, we note that the shape derivative of a generic function $j\colon \Omega_t \to \mathbb{R}$ with respect to the vector field V is given by

$$Dj[V] := D_m j - V^T j. \tag{39}$$

Therefore $D_m(f)p$ is equal to $pV^T\nabla f$ in the both subdomains $\Omega_1(u)$, $\Omega_2(u)$. Due to the continuity of the state and of the flux (20) their material derivative is continuous. Thus, we get

$$\left[\!\!\left[\frac{\partial p}{\partial n} D_m(y)\right]\!\!\right] = D_m(y)\left[\!\!\left[\frac{\partial p}{\partial n}\right]\!\!\right] \overset{(32)}{=} 0 \quad \text{on } u \tag{40}$$

$$\left[\!\!\left[D_m\left(\frac{\partial y}{\partial n}\right) p\right]\!\!\right] = D_m\left(\frac{\partial y}{\partial n}\right)[\![p]\!] \overset{(31)}{=} 0 \quad \text{on } u. \tag{41}$$

That

$$\left[\!\!\left[\frac{\partial y}{\partial n} p\right]\!\!\right] = 0 \quad \text{on } u \tag{42}$$

follows from (20), (31) and the identity

$$[\![ab]\!] = [\![a]\!]b_1 + a_2[\![b]\!] = a_1[\![b]\!] + [\![a]\!]b_2 \tag{43}$$

which implies

$$[\![ab]\!] = 0 \text{ if } [\![a]\!] = 0 \wedge [\![b]\!] = 0. \tag{44}$$

By combining (29), (33) and (38–42), we obtain

$$\boxed{\begin{aligned}\frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V] = \int_{\Omega(u)} &-\nabla y^T\left(\nabla V + \nabla V^T\right)\nabla p - pV^T\nabla f \\ &+ \operatorname{div}(V)\left(\frac{1}{2}(y-\overline{y})^2 + \nabla y^T\nabla p - fp\right) dx\end{aligned}} \tag{45}$$

i.e., the shape derivative of $\mathscr{L}$ expressed as domain integral which is equal to the one of J due to the theorem of Correa and Seeger. Now, we convert this domain integral into a boundary integral as mentioned above. Integration by parts in (45) yields

$$\begin{aligned}&\int_{\Omega(u)} \operatorname{div}(V)\left(\frac{1}{2}(y-\overline{y})^2 + \nabla y^T\nabla p - fp\right) dx \\ &= -\int_{\Omega(u)} V^T\left((y-\overline{y})\nabla y + \nabla\left(\nabla y^T\nabla p\right) - \nabla(fp)\right) dx \\ &\quad + \int_u \left[\!\left[\left(\frac{1}{2}(y-\overline{y})^2 + \nabla y^T\nabla p - fp\right)\langle V, n\rangle\right]\!\right] ds \\ &\quad + \int_{\partial\Omega(u)} \left(\frac{1}{2}(y-\overline{y})^2 + \nabla y^T\nabla p - fp\right)\langle V, n\rangle\, ds.\end{aligned} \tag{46}$$

Since the outer boundary $\partial\Omega$ is not variable, we can choose the deformation vector field V equals zero in small neighbourhoods of $\partial\Omega(u)$. Therefore, the outer integral in (46) disappears. Combining (45), (46) and the vector calculus identity

$$\nabla y^T\left(\nabla V + \nabla V^T\right)\nabla p + V^T\nabla\left(\nabla y^T\nabla p\right) = \nabla p^T\nabla\left(V^T\nabla y\right) + \nabla y^T\nabla\left(V^T\nabla p\right)$$

which was proved in [7] gives

$$\begin{aligned}\frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V] = \int_{\Omega(u)} &-\nabla p^T\nabla\left(V^T\nabla y\right) - \nabla y^T\nabla\left(V^T\nabla p\right) \\ &- (y-\overline{y})V^T\nabla y + fV^T\nabla p dx \\ &+ \int_u \left[\!\left[\left(\frac{1}{2}(y-\overline{y})^2 + \nabla y^T\nabla p - fp\right)\langle V, n\rangle\right]\!\right] ds.\end{aligned} \tag{47}$$

Then, applying integration by parts in (47) we get

$$\begin{aligned}&\int_{\Omega(u)} \nabla y^T\nabla\left(V^T\nabla p\right) dx \\ &= -\int_{\Omega(u)} \triangle y V^T\nabla p dx + \int_u \left[\!\left[\frac{\partial y}{\partial n}V^T\nabla p\right]\!\right] ds + \int_{\partial\Omega(u)} \frac{\partial y}{\partial n}V^T\nabla p ds\end{aligned} \tag{48}$$

and analogously

$$\int_{\Omega(u)} \nabla p^T \nabla \left(V^T \nabla y\right) dx$$
$$= -\int_{\Omega(u)} \triangle p V^T \nabla y dx + \int_u \left[\!\left[\frac{\partial p}{\partial n} V^T \nabla y \right]\!\right] ds + \int_{\partial\Omega(u)} \frac{\partial p}{\partial n} V^T \nabla y ds. \tag{49}$$

Like in (46) the outer integral in (48) as well as in (49) vanishes due to the fixed outer boundary $\partial\Omega(u)$. Thus, it follows that

$$\begin{aligned}\frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V] = &\int_{\Omega(u)} V^T \nabla p \left(\triangle y + f\right) + V^T \nabla y \left(\triangle p - (y - \overline{y})\right) dx \\ &+ \int_u \left[\!\left[\left(\frac{1}{2}(y-\overline{y})^2 + \nabla y^T \nabla p - fp \right) \langle V, n\rangle \right]\!\right] \\ &- \left[\!\left[\frac{\partial y}{\partial n} V^T \nabla p \right]\!\right] - \left[\!\left[\frac{\partial p}{\partial n} V^T \nabla y \right]\!\right] ds. \end{aligned} \tag{50}$$

The domain integral in (50) vanishes due to (29) and (33). Moreover, the term $\left[\!\left[\frac{1}{2}(y-\overline{y})^2 \langle V, n\rangle \right]\!\right]$ disappears because of (20) and the term $\left[\!\left[\nabla y^T \nabla p \langle V, n\rangle \right]\!\right]$ because of the continuity of ∇y and ∇p. That

$$\left[\!\left[\frac{\partial y}{\partial n} V^T \nabla p \right]\!\right] = \left[\!\left[\frac{\partial p}{\partial n} V^T \nabla y \right]\!\right] = \langle V, n\rangle \left[\!\left[\frac{\partial y}{\partial n} \frac{\partial p}{\partial n} \right]\!\right] = 0 \tag{51}$$

follows from (20), (32) and (44). Thus, we obtain the shape derivative of $\mathscr{L}$ expressed as interface integral:

$$\boxed{\frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V] = -\int_u [\![f]\!] p \langle V, n\rangle ds} \tag{52}$$

Now, we consider the objective J in (16) with perimeter regularization. Combining (52) with proposition 5.1 in [23] we get

$$\boxed{\frac{\partial}{\partial u}\mathscr{L}(y,u,p)[V] = \int_u \left(-[\![f]\!] p + \mu\kappa\right) \langle V, n\rangle ds} \tag{53}$$

where κ denotes the curvature corresponding to the normal n.

Remark 5 Note that (52) is equal to $\frac{\partial}{\partial u}J(y,u)[V]$ without perimeter regularization and (53) is equal to $\frac{\partial}{\partial u}J(y,u)[V]$ with perimeter regularization due to the theorem of Correa and Seeger as mentioned above.

We focus now on the weak formulation (11–13) and observe the following for the right hand sides in the case of (16–18):

$$-a_u(\bar{z}, p) - \frac{\partial}{\partial y} J(y, u)\bar{z} = -\int_{\Omega(u)} \nabla \bar{z}^T \nabla p + (y - \bar{y})\bar{z} dx \tag{54}$$

$$-\frac{\partial}{\partial u}\left[J(y, u) + a_u(y, p) - b_u(p)\right]\bar{w} = \int_u ([\![f]\!]p - \mu\kappa)\,\langle \bar{w}, n\rangle\, ds \tag{55}$$

$$-a_u(y, \bar{q}) + b_u(\bar{q}) = \int_{\Omega(u)} -\nabla y^T \nabla \bar{q} + f\bar{q} dx \tag{56}$$

These expressions are set to zero, in order to define the necessary conditions of optimality.

Now, we discuss more details about the Hessian operators in the left hand sides of (11–13). We first consider them without the term H_{22} which requires special care. These are at the solution $(y, u, p) \in F$ of the optimization problem (16–18) for all $h := (\bar{z}, \bar{w}, \bar{q})^T \in T_{(y,u,p)}F$ as follows:

$$\begin{aligned}
H_{11}(z, \bar{z}) &= \frac{\partial^2}{\partial y^2} J(y, u) z\bar{z} = \int_{\Omega(u)} z\bar{z} dx \\
H_{12}(w, \bar{z}) &= \frac{\partial}{\partial u}\left[a_u(\bar{z}, p) + \frac{\partial}{\partial y} J(y, u)\bar{z}\right] w = 0 \\
H_{13}(q, \bar{z}) &= a_u(\bar{z}, q) = \int_{\Omega(u)} \nabla \bar{z}^T \nabla q dx \\
H_{21}(z, \bar{w}) &= \frac{\partial}{\partial y}\frac{\partial}{\partial u}(\left[J(y, u) + a_u(y, p)\right]\bar{w})z = 0 \\
H_{23}(q, \bar{w}) &= \frac{\partial}{\partial u}\left[a_u(y, q) - b_u(q)\right]\bar{w} = -\int_u [\![f]\!] q\,\langle \bar{w}, n\rangle\, ds \\
H_{31}(z, \bar{q}) &= a_u(z, \bar{q}) = \int_{\Omega(u)} \nabla z^T \nabla \bar{q} dx \\
H_{32}(w, \bar{q}) &= \frac{\partial}{\partial u}\left[a_u(y, \bar{q}) - b_u(\bar{q})\right] w = -\int_u [\![f]\!]\bar{q}\,\langle w, n\rangle\, ds \\
H_{33}(q, \bar{q}) &= 0
\end{aligned}$$

We compute now the term H_{22}. It will be evaluated at the solution of the optimization problem which means that it consists only of the second shape derivative. In Sect. 4 this solution will be a straight line connection of the points (0.5, 0) and (0.5, 1), i.e., the curvature is equal to zero. Combining proposition 5.1 in [23] with the following rule for differentiating boundary integrals

$$\frac{d^+}{dt}\left(\int_{\Gamma_t} \eta(t)\right)\Big|_{t=0} = \int_\Gamma \left(D\eta[V] + \left(\frac{\partial \eta}{\partial n} + \eta\kappa\right)\langle V, n\rangle\right) \tag{57}$$

which was proved in [16] yields

$$\begin{aligned} &H_{22}(w,\bar{w}) \\ &= G(Hess^{\mathcal{N}}\,(J(y,u)+a_u(y,p)-b_u(p))\,w,\bar{w}) \\ &= \int_u -D\,([\![f]\!]\,p)\,[\bar{w}]\,\langle w,n\rangle - [\![f]\!]\left(\kappa p+\frac{\partial p}{\partial n}\right)\langle\bar{w},n\rangle\,\langle w,n\rangle \\ &\quad + \mu\frac{\partial w}{\partial\tau}\frac{\partial\bar{w}}{\partial\tau}\,\langle\bar{w},n\rangle\,\langle w,n\rangle\,ds \end{aligned} \tag{58}$$

where $\partial/\partial\tau$ denotes the derivative tangential to u. We have to evaluate the shape derivative $D\,([\![f]\!]p)\,[\bar{w}]$ in (58). We observe in our special case

$$p = 0 \quad \text{on } u \tag{59}$$

because of the necessary optimality condition (55). Thus, it holds that

$$Dp[\bar{w}] = -\bar{w}^T\nabla p = -\bar{w}^T\frac{\partial p}{\partial n}n \quad \text{on } u \tag{60}$$

due to (39). Applying the product rule for shape derivatives yields

$$\begin{aligned} D\,([\![f]\!]p)\,[\bar{w}] &= [\![Df[\bar{w}]p]\!] + [\![fDp[\bar{w}]]\!] \overset{(20)}{=} [\![Df[\bar{w}]]\!]p + [\![f]\!]Dp[\bar{w}] \\ &\underset{(60)}{\overset{(59)}{=}} -[\![f]\!]\frac{\partial p}{\partial n}\,\langle\bar{w},n\rangle \quad \text{on } u. \end{aligned} \tag{61}$$

Thus, the Hessian operator H_{22} reduces to

$$\hat{H}_{22}(w,\bar{w}) = \int_u\left(\mu\frac{\partial w}{\partial\tau}\frac{\partial\bar{w}}{\partial\tau} - [\![f]\!]\kappa p\right)\langle w,n\rangle\,\langle\bar{w},n\rangle\,ds. \tag{62}$$

By using the expressions above, we can formulate the QP (14, 15) at the solution in the following form:

$$\min_{(z,w)} F(z,w,y,p) \tag{63}$$

$$\begin{aligned} \text{s.t.} \int_{\Omega(u)} \nabla z^T\nabla\bar{q}dx - \int_u [\![f]\!]\bar{q}w ds \\ = -\int_{\Omega(u)}\nabla y^T\nabla\bar{q}dx + \int_{\Omega(u)} f\bar{q}dx\,,\ \forall\bar{q}\in H_0^1(\Omega(u)) \end{aligned} \tag{64}$$

where the objective function F is given by

$$F(z, w, y, p) = \int_{\Omega(u)} \frac{z^2}{2} + (y - \bar{y})z dx + \int_u \mu\kappa w - [\![f]\!] pw ds + \frac{1}{2}\int_u \mu \left(\frac{\partial w}{\partial \tau}\right)^2 - [\![f]\!]\kappa p w^2 ds. \tag{65}$$

This QP in weak formulation can be rewritten in the more intelligible strong form of an optimal control problem:

$$\min_{(z,w)} F(z, w, y, p) \tag{66}$$

$$\text{s.t.} \quad -\triangle z = \triangle y + f_1 \quad \text{in } \Omega_1(u) \tag{67}$$

$$-\triangle z = \triangle y + f_2 \quad \text{in } \Omega_2(u) \tag{68}$$

$$\frac{\partial z}{\partial n} = f_1 w \quad \text{on } u \tag{69}$$

$$-\frac{\partial z}{\partial n} = f_2 w \quad \text{on } u \tag{70}$$

$$z = 0 \quad \text{on } \partial\Omega(u) \tag{71}$$

The adjoint problem to this optimal control problem is the boundary value problem:

$$-\triangle q = -z - (y - \bar{y}) \quad \text{in } \Omega(u) \tag{72}$$

$$q = 0 \quad \text{on } \partial\Omega(u) \tag{73}$$

The resulting design equation for the optimal control problem (66–71) is

$$0 = -[\![f]\!]\,(p + \kappa p w + q) + \mu\kappa - \mu\frac{\partial^2 w}{\partial \tau^2} \quad \text{on } u. \tag{74}$$

4 Numerical Results

In this section, we use the QP (63, 64) away from the optimal solution as a means to determine the step in the shape normal direction and thus create an iterative solution technique very similar to SQP techniques known from linear spaces. We solve the optimal control problem (66–71) by employing a CG–iteration for the reduced problem (74). I.e., we iterate over the variable w and each time the CG–iteration needs a residual of equation (74) from w^k, we compute the state variable z^k from (67–71) and then the adjoint variable q^k from (72, 73), which enables

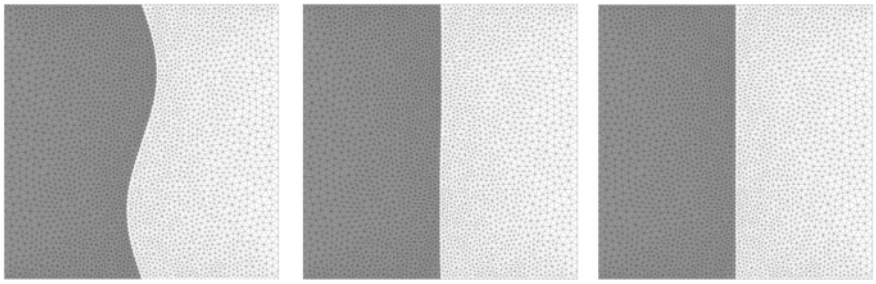


Fig. 2 Iterations 0, 1 and 2 (*left* to *right*) together with deformations of mesh Ω_h^1

Table 1 Performance of shape Lagrange–Newton algorithms: distances $\mathrm{dist}(u^k, u^*)$ from the optimal solution on meshes with varying refinement

It. no.	Ω_h^1	Ω_h^2	Ω_h^3
0	0.0705945	0.070637	0.0706476
1	0.0043115	0.004104	0.0040465
2	0.0003941	0.000104	0.0000645

Quadratic convergence on the finest grid can be observed

the evaluation of the residual

$$r^k := -[\![f]\!]\left(p + \kappa p w^k + q^k\right) + \mu\kappa - \mu\frac{\partial^2 w^k}{\partial \tau^2}. \tag{75}$$

The particular values for the parameters are chosen as $f_1 = 1000, f_2 = 1$ and $\mu = 10$. The data $\bar{y}$ are generated from a solution of the state equation (17) with u being the straight line connection of the points $(0.5, 0)$ and $(0.5, 1)$. The starting point of our iterations is described by a B-spline defined by the two control points $(0.6, 0.7)$ and $(0.4, 0.3)$. We build a coarse unstructured tetrahedral grid Ω_h^1 with roughly 6000 triangles as shown in the leftmost picture of Fig. 2. We also perform computations on uniformly refined grids Ω_h^2, Ω_h^3 with roughly 24000 and 98000 triangles. In Fig. 2 are also shown the next two iterations on the coarsest grid, where table 4 gives the distances of each shape to the solution approximated by

$$\mathrm{dist}(u^k, u^*) := \int_{u^*} \left| \langle u^k, e_1 \rangle - \frac{1}{2} \right| ds$$

where u^* denotes the solution shape and $e_1 = (1, 0)$ is the first unit vector. Similar to [27], the retraction chosen for the shape is just the addition of the $q^k n_1$ to the current shape. In each iteration, the volume mesh is deformed according to the elasticity equation. Table 1 [27] demonstrates that indeed quadratic convergence can be observed on the finest mesh, but also that the mesh resolution has a strong influence on the convergence properties revealed.

The major advantage of the Newton method over a standard shape calculus steepest method based on the (reduced) shape derivative

$$dJ(y,u)[V] = -\int_u (\llbracket f \rrbracket p - \mu\kappa)\,\langle V, n\rangle ds$$

is the natural scaling of the step, which is just 1 near to the solution. When first experimenting with a steepest descent method, we found by trial and error, that one needs a scaling around 10 000 in order to obtain sufficient progress.

5 Conclusions

This paper presents a generalization of the Riemannian shape calculus framework in [27] to Lagrange–Newton approaches for PDE constrained shape optimization problems. It is based on the idea that Riemannian shape Hessians do not differ from classical shape Hessians in the solution of a shape optimization problem and that Newton methods still converge locally quadratically, if Hessian terms are neglected which are zero at the solution anyway. It is shown that this approach is viable and leads to computational methods with superior convergence properties, when compared to only linearly converging standard steepest descent methods. Nevertheless, several issues have to be addressed in future investigations, like:

- More refined retractions have to be developed for large shape deformations.
- As observed during the computations, the shape deformation sometimes leads to shapes, where normal vectors can no longer be reliably evaluated. Provisions for those cases have be developed.
- Full Lagrange–Newton methods may turn out being not very computationally efficient. However, this paper lays the foundation for the construction of appropriate preconditoners for the reduced optimization problem in many practical cases.
- The Riemannian shape space properties including quadratic convergence of the Lagrange–Newton approach seem to materialize only on very fine grids. A logical next development is then to use locally adapted meshes near the shape front to be optimized.

Acknowledgments This research has been partly funded by the DFG within the collaborative project EXASOLVERS as part of the DFG priority program SPP 1648 SPPEXA.

References

1. H.B. Ameur, M. Burger, B. Hackl, Level set methods for geometric inverse problems in linear elasticity. Inverse Probl. **20**, 673–696 (2004)
2. E. Arian, Analysis of the Hessian for aeroelastic optimization. Technical report 95-84, Institute for Computer Applications in Science and Engineering (ICASE) (1995)
3. E. Arian, S. Ta'asan, Analysis of the Hessian for aerodynamic optimization: inviscid flow. Technical report 96-28, Institute for Computer Applications in Science and Engineering (ICASE) (1996)
4. E. Arian, V.N. Vatsa, A preconditioning method for shape optimization governed by the Euler equations. Technical report 98-14, Institute for Computer Applications in Science and Engineering (ICASE) (1998)
5. M. Bauer, P. Harms, P.W. Michor, Sobolev metrics on shape space II: weighted Sobolev metrics and almost local metrics. (2011) arXiv:1109.0404
6. M. Bauer, P. Harms, P.W. Michor, Sobolev metrics on shape space of surfaces. J. Geom. Mech. **3**(4), 389–438 (2011)
7. M. Berggren, A unified discrete–continuous sensitivity analysis method for shape optimization, in *Applied and Numerical Partial Differential Equations*, Computational Methods in Applied Sciences, vol. 15, ed. by W. Fitzgibbon et al. (Springer, 2010), pp. 25–39
8. A. Borzì, V.H. Schulz, *Computational Optimization of Systems Governed by Partial Differential Equations*, vol. 08, SIAM Book Series on Computational Science and Engineering (SIAM Philadelphia, 2012)
9. J. Céa, Conception optimale ou identification de formes calcul rapide de la dérivée directionelle de la fonction coût. RAIRO Modelisation mathématique et analyse numérique **20**(3), 371–402 (1986)
10. L. Conlon, *Differentiable Manifolds* (Birkhäuser, Birkäuser Advanced Texts: Basler Lehrbücher (Birkäuser, 2001)
11. R. Correa, A. Seeger, Directional derivative of a minmax function. Nonlinear Anal. **9**(1), 13–22 (1985)
12. M.C. Delfour, J.-P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Advances in Design and Control (SIAM Philadelphia, 2001)
13. K. Eppler, S. Schmidt, V.H. Schulz, C. Ilic, Preconditioning the pressure tracking in fluid dynamics by shape Hessian information. J. Optim. Theory Appl. **141**(3), 513–531 (2009)
14. N. Gauger, C. Ilic, S. Schmidt, V.H. Schulz, Non-parametric aerodynamic shape optimization, in *Constrained Optimization and Optimal Control for Partial Differential Equations*, ed. by G. Leugering, S. Engell, A. Griewank, M. Hinzea, R. Rannacher, V.H. Schulz, M. Ulbrich, S. Ulbrich, vol. 160, (Birkhäuser, Basel, 2012), pp. 289–300
15. J. Haslinger, M.A.E. Mäkinen, *Introduction to Shape Optimization: Theory, Approximation, and Computation*, Advances in Design and Control, (SIAM Philadelphia, 2008)
16. E.J. Haug, K.K. Choi, V. Komkov, *Design Sensitivity Analysis of Structural Systems* (Academic Press, Orlando, 1986)
17. K. Ito, K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Applications*. Advances in Design and Control (SIAM, 2008)
18. S. Lang, *Fundamentals of Differential Geometry* (Springer, 2001)
19. P.W. Michor, D. Mumford, Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms. Documeta Math. **10**, 217–245 (2005)
20. P.W. Michor, D. Mumford, Riemannian geometries on spaces of plane curves. J. Eur. Math. Soc. (JEMS) **8**, 1–48 (2006)
21. P.W. Michor, D. Mumford, An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach. Appl. Comput. Harm. Anal. **23**, 74–113 (2007)
22. B. Mohammadi, O. Pironneau, *Applied Shape Optimization for Fluids*, Numerical Mathematics and Scientific Computation (Clarendon Press, Oxford, 2001)
23. A. Novruzi, M. Pierre, Structure of shape derivatives. J. Evol. Equ. **2**, 365–382 (2002)

24. C. Schillings, S. Schmidt, V.H. Schulz, Efficient shape optimization for certain and uncertain aerodyamic design. Comput. Fluids **46**(1), 78–87 (2011)
25. S. Schmidt, V.H. Schulz, Impulse response approximations of discrete shape Hessians with application in CFD. SIAM J. Control Optim. **48**(4), 2562–2580 (2009)
26. S. Schmidt, V.H. Schulz, Shape derivatives for general objective functions and the incompressible Navier-Stokes equations. Control Cybern. **39**(3), 677–713 (2010)
27. V.H. Schulz, A Riemannian view on shape optimization. Found. Comput. Math. **14**, 483–501 (2014)
28. V.H. Schulz, M. Siebenborn, K. Welker, Structured inverse modeling in parabolic diffusion problems. submitted to SICON (2014) arXiv:1409.3464
29. J. Sokolowski, J.P. Zolésio, *Introduction to Shape Optimization: Shape Sensitivity Analysis* (Springer, 1992)

Shape Optimization in Electromagnetic Applications

Johannes Semmler, Lukas Pflug, Michael Stingl and Günter Leugering

Abstract We consider shape optimization for objects illuminated by light. More precisely, we focus on time-harmonic solutions of the Maxwell system in **curl-curl**-form scattered by an arbitrary shaped rigid object. Given a class of cost functionals, including the scattered energy and the extinction cross section, we develop an adjoint-based shape optimization scheme which is then applied to two key applications.

1 Introduction

Problems of shape optimization for Maxwell's system have been recently studied in the literature. In particular, for time harmonic solutions, adjoint based gradient descent methods have been considered in [1, 16, 17, 22] and most recently in [13]. The transient problem has been discussed in [8]. Applications range from inverse problems related to the detection of cavities to shape optimization of machinery tools. In this note, we concentrate on an adjoint-based descent scheme using volume representations for the shape derivative. We consider two particular applications: the first concerns the optimization of particle shapes with respect to the extinction cross section as cost functional, whereas the second is related to the geometry optimization of nano-antennas. In both cases the results are realized in close contact with physicists and particle engineers.

1.1 Maxwell's Equation

The first section is a formal derivation of Maxwell's equation and follows the format of standard publications on electrodynamics, such as [19, 29]. In classical elec-

J. Semmler (✉) · L. Pflug · M. Stingl, · G. Leugering
Institute of Applied Mathematics 2, Friedrich-Alexander University Erlangen-Nürnberg,
Cauerstrasse 11, 91058 Erlangen, Germany
e-mail: johannes.semmler@fau.de

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_11

trodynamics, the macroscopic electromagnetic fields are described by four vector functions $\mathcal{E}$, $\mathcal{D}$, $\mathcal{H}$ and $\mathcal{B}$. These three-dimensional real vector functions are functions of position $x \in \mathbb{R}^3$ and time $t \in \mathbb{R}$. The fields $\mathcal{E}$ and $\mathcal{H}$ are called electric and magnetic field intensities, respectively. These fields are sufficient to describe the electromagnetic field, if the electric displacement $\mathcal{D}$ and the magnetic induction $\mathcal{B}$ are related to $\mathcal{E}$ and $\mathcal{H}$ by constitutive equations. The distribution of sources, consisting of static electric charges and directed flow of electric charges, which is called current, creates an electromagnetic field. Static charges are given by a scalar charge density function $\mathcal{R}$, while currents are described by the vector current density function $\mathcal{J}$. The field variables and sources are related by the following Maxwell equations:

$$\begin{aligned} \frac{\partial \mathcal{B}}{\partial t} + \operatorname{curl} \mathcal{E} &= 0 && \text{(Faraday's law)} \\ \operatorname{div} \mathcal{D} &= \mathcal{R} && \text{(Gauss's law)} \\ \frac{\partial \mathcal{D}}{\partial t} - \operatorname{curl} \mathcal{H} &= -\mathcal{J} && \text{(Ampere's law)} \\ \operatorname{div} \mathcal{B} &= 0. \end{aligned}$$

These equations apply in the region of space in $\mathbb{R}^3$ occupied by the electromagnetic field. Faraday's law describes the action of changing the magnetic field on the electric field. Gauss's law relates the charge density to the electric displacement. Ampere's law is modified by James C. Maxwell and the last equation states that the magnetic induction $\mathcal{B}$ is solenoidal. Since charge has to be conserved, the sources are related to each other via

$$\partial_t \mathcal{R} + \operatorname{div} \mathcal{J} = 0 \tag{1.1}$$

which is already included in the divergence equations, but can not be entirely ignored, when solving Maxwell's equation numerically. If we are only interested in electromagnetic fields and sources at a single frequency then the time-dependent Maxwell equations can be reduced to the time-harmonic Maxwell system. With a frequency $\omega > 0$ the time-harmonic electric field are given by

$$\mathcal{E}(x, t) = \operatorname{Re}\left(\exp(-\imath\omega t) E(x)\right) \tag{1.2}$$

where $\imath$ is the imaginary unit and E is a complex-valued vector function of position. The time-harmonic formulation of the electric field (1.2) applies to all field and source quantities as well. Thus with (1.1), the time-harmonic Maxwell system reads as

$$\left.\begin{aligned} -\imath\omega B + \operatorname{curl} E &= 0 \\ \imath\omega \operatorname{div} D &= \operatorname{div} J \\ -\imath\omega D - \operatorname{curl} H &= -J \\ \operatorname{div} B &= 0 \end{aligned}\right\} \tag{1.3}$$

The time-harmonic Maxwell system (1.3) is extended by constitutive laws that relate E and H to D and B respectively. The properties of material occupied by the electromagnetic field affect these laws. Here it is assumed that the material properties are linear, i.e. independent of the electromagnetic field, and isotropic, i.e. uniform in all directions of the medium. Thus the electric displacement D and magnetic induction B can be expressed by

$$D = \varepsilon E \quad \text{and} \quad B = \mu H, \tag{1.4}$$

where ε and μ are called, respectively, electric permittivity and magnetic permeability. In anisotropic media the material constants ε and μ would be second-order tensors. For vacuum these constants are denoted by ε_0 and μ_0 and have the values

$$\varepsilon_0 \approx 8.854 \times 10^{-12}\,\mathrm{F\,m^{-1}} \quad \text{and} \quad \mu_0 = 4\pi 10^{-7}\,\mathrm{H\,m^{-1}}.$$

If the material is conductive, like metals, then the electric field gives rise to currents. We assume that Ohm's law holds, which is satisfied for sufficiently small field strengths, so that

$$J = \sigma E + J_a, \tag{1.5}$$

where σ is called conductivity and is a non-negative number. The vector function J_a describes the applied current density. A nonconducting medium, except vacuum, is termed dielectric. Using the constitutive equations for fields (1.4) and currents (1.5), we arrive at the following time-harmonic Maxwell system

$$\left.\begin{aligned} -\imath\kappa c\mu_0\mu_r H + \operatorname{curl} E &= 0 \\ -\imath\kappa c\varepsilon_0\varepsilon_r E - \operatorname{curl} H &= -J_a, \end{aligned}\right\} \tag{1.6}$$

where the frequency $\omega = \kappa c$ is replaced by the wave number κ times the vacuum speed of light $c = (\varepsilon_0\mu_0)^{-1/2}$. The relative permittivity ε_r and relative permeability μ_r are defined by

$$\varepsilon_r = \frac{1}{\varepsilon_0}\left(\varepsilon + \frac{\imath\sigma}{\omega}\right) \quad \text{and} \quad \mu_r = \frac{\mu}{\mu_0}.$$

Materials are also characterized by their complex refractive index $n = \sqrt{\varepsilon_r\mu_r}$ in the literature. Thus, ε_r is computed by $\varepsilon_r = n^2\mu_r^{-1}$, if μ_r is known. By eliminating H from the first-order Maxwell system (1.6), we obtain the second-order Maxwell system or curl-curl formulation with $F = \imath\kappa c\mu_0 J_a$

$$\operatorname{curl} \mu_r^{-1} \operatorname{curl} E - \kappa^2\varepsilon_r E = F. \tag{1.7}$$

The natural interface conditions between two different materials, where the material constants are discontinuous, are derived from the integral form of Maxwell equations [25]. These conditions state that the tangential components of electric and magnetic fields are continuous over the interface with normal vector n between material 1 and

material 2 with relative permeability μ_1 and μ_2 respectively:

$$E_1 \times n = E_2 \times n \quad \text{and} \quad \mu_1^{-1}\text{curl}\, E_1 \times n = \mu_2^{-1}\text{curl}\, E_2 \times n$$

E_1 denotes the electric field in material 1 and E_2 denotes the electric field in material 2.

1.2 Scattering Problem

A special case of Maxwell's equation in the curl-curl formulation (1.7) is the scattering of a bounded object, see Fig. 1, which is exposed to an incident electric field. Assuming the absence of external sources and propagation of the electric field with wave number κ, the total electric field E_T has to satisfy the curl-curl problem, i.e.

$$\text{curl}\, \mu_r^{-1}\text{curl}\, E_T - \kappa^2 \varepsilon_r E_T = 0 \tag{1.8}$$

where ε_r is continuously differentiable in each part of the scatterer with $\text{Re}\,(\varepsilon_r) > 0$ and $\text{Im}\,(\varepsilon_r) \geq 0$ and $\mu_r > 0$. Obviously, one solution of the problem is a vanishing electric field, but the total electric field has also to contain the incident field E_I. Thus, the total electric field E_T can be expressed by the scattered field E and the incident field E_I

$$E_T = E + E_I. \tag{1.9}$$

The incident electric field must satisfy Maxwell's equation in absence of the scattering object, this means, for instance, for constant relative permittivity ε_b with $\varepsilon_b = \varepsilon_r$ outside the bounded scattering object and constant relative permeability $\mu_b = \mu_r$, i.e.

$$\text{curl}\, \mu_b^{-1}\text{curl}\, E_I - \kappa^2 \varepsilon_b E_I = 0. \tag{1.10}$$

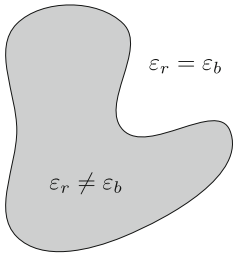


Fig. 1 A bounded scatterer where the relative permittivity ε_r differs from the background permittivity ε_b

Inserting (1.9) and (1.10) in (1.8), we get the curl-curl formulation of the time-harmonic Maxwell system for the scattered field E incident by E_I

$$\operatorname{curl} \mu_r^{-1} \operatorname{curl} E - \kappa^2 \varepsilon_r E = \kappa^2 (\varepsilon_r - \varepsilon_b) E_I. \tag{1.11}$$

To obtain uniqueness of the solution, we have to impose a condition for $\|x\| \to \infty$ [20, Theorem 61], e.g. the Silver-Müller-radiation condition

$$\lim_{\|x\| \to \infty} \left(\operatorname{curl} E \times x - \|x\| \imath \kappa E \right) = 0. \tag{1.12}$$

For numerical simulations the radiation condition can be approximated on a bounded domain by a perfectly matched layer (PML) as derived in [6]. The PML is realized by replacing the relative permeability and permittivity by an anisotropic complex tensor which represents an infinite extended and absorbing material.

2 Shape Optimization

Let the hold-all $D \subset \mathbb{R}^3$ be a bounded domain composed of N disjoint sub-domains Ω_i with boundaries Γ_i for $i \in \{1, \ldots, N\}$. Furthermore Γ_{ij} denotes for $i < j$ the interface between two domains Ω_i and Ω_j, i.e. $\Gamma_{ij} = \Gamma_i \cap \Gamma_j$ (see Fig. 2). Then we formulate the shape optimization problem for electrodynamic applications as

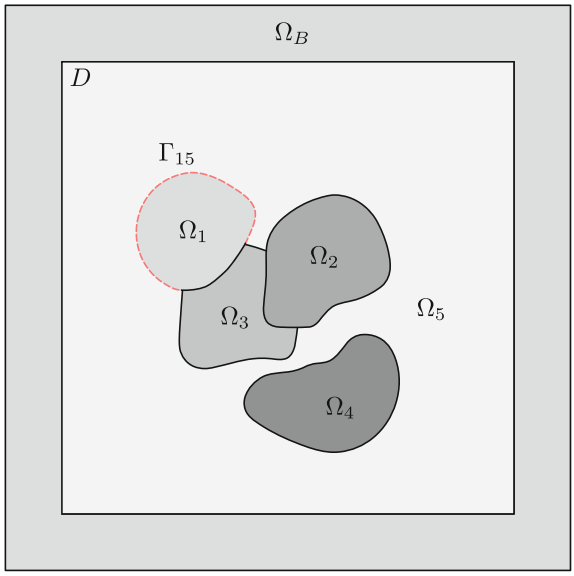


Fig. 2 Exemplary partition of the hold-all D into five sub-domains $\Omega_1, \ldots, \Omega_5$ surrounded by Ω_B, which represents a PML, and the interface Γ_{15} between domain 1 and domain 5 in *red*

$$\min_{\Omega\in\mathcal{O}} J(\Omega, u_\Omega) \quad \text{s.t. } \mathcal{A}_\Omega(u_\Omega) = 0.$$

We summarize the sub-domains Ω_i to a tuple denoted by $\Omega := (\Omega_1, \ldots, \Omega_N)$ with the set $\mathcal{O}$ of admissible Ω. The tuple Ω will be referred to as shape. The real scalar function J is called objective functional and depends on the shape Ω and solution u_Ω of a partial differential equation on Ω denoted by $\mathcal{A}_\Omega$. According to the electrodynamic application, $\mathcal{A}_\Omega$ is one of the formulations of Maxwell's equation.

Since we are interested in shape optimization of scattering problems for a single frequency, $\mathcal{A}_\Omega$ is the curl-curl formulation for the scattered electric field surrounded by an additional perfectly matched layer Ω_{N+1}. The PML $\Omega_B := \Omega_{N+1}$ is a numerical tool to approximate the scattering condition (1.12) and is fixed during optimization. The outer boundary of Ω_B is denoted by $\Gamma_{B\infty}$. Thus we conclude our shape optimization setting

$$\left.\begin{aligned} &\min_{\Omega\in\mathcal{O}} J(\Omega, E) \text{ s.t. } \forall i \in \{1, \ldots, N\} \text{ and } i < j \le N+1 \\ &\operatorname{curl} \mu_r^{-1} \operatorname{curl} E - \kappa^2 \varepsilon_r E = \kappa^2(\varepsilon_r - \varepsilon_b) E_I \quad \text{in } \Omega_i \\ &\operatorname{curl} \hat{\mu}_b^{-1} \operatorname{curl} E - \kappa^2 \hat{\varepsilon}_b E = 0 \quad \text{in } \Omega_B \\ &[E \times n] = 0 \quad \text{on } \Gamma_{ij} \\ &\left[\mu_r^{-1} \operatorname{curl} E \times n\right] = 0 \quad \text{on } \Gamma_{ij} \\ &E \times n = 0 \quad \text{on } \Gamma_{B\infty} \end{aligned}\right\} \tag{2.1}$$

where, on each Ω_i for $i \in \{1, \ldots, N\}$, the relative permittivities ε_r and ε_b are chosen constant and the relative permeability $\mu_r = 1$, due to the application. In Ω_B the coefficients $\hat{\varepsilon}_b = \varepsilon_b M_\varepsilon$ and $\hat{\mu}_b = \mu_b M_\mu$ are complex-valued matrix functions representing the PML. Moreover, on the inner boundary Γ_{iB} of Ω_B the PML matrices M_ε and M_μ are the identity. To simplify the notation, we restrict our-self to use ε_r and μ_r through out, i.e. $\varepsilon_r = \hat{\varepsilon}_b$ and $\mu_r = \hat{\varepsilon}_b$ on Ω_B. Furthermore, the extension of the hold-all D through Ω_B is denoted by $\mathcal{B} = \bar{D} \cup \Omega_B$. The weak formulation of the PDE constraint, further called state or primal problem, reads as follows:

$$\begin{aligned} &\text{Find } E \in H_0(\operatorname{curl}, \mathcal{B}) \text{ s.t.} \\ &\quad (\mu_r^{-1} \operatorname{curl} E, \operatorname{curl} \varphi)_\mathcal{B} - \kappa^2(\varepsilon_r E, \varphi)_\mathcal{B} \quad = (F, \varphi)_D \quad \forall \varphi \in H_0(\operatorname{curl}, \mathcal{B}) \end{aligned} \tag{2.2}$$

where $F := \kappa^2((\varepsilon_r - \varepsilon_b) E_I$. The function space $H(\operatorname{curl}, \mathcal{B})$ with $L^2(\mathcal{B})$ functions with a weak curl in $L^2(\mathcal{B})$ and the function space $H_0(\operatorname{curl}, \mathcal{B})$ of $H(\operatorname{curl}, \mathcal{B})$-functions with vanishing tangential trace are defined as

$$\begin{aligned} H(\operatorname{curl}, \mathcal{B}) &:= \{v \in (L^2(\mathcal{B}))^3 : \operatorname{curl} v \in (L^2(\mathcal{B}))^3\}, \\ H_0(\operatorname{curl}, \mathcal{B}) &:= \{u \in H(\operatorname{curl}, \mathcal{B}) : u \times n = 0 \text{ a.e. on } \partial\mathcal{B}\}. \end{aligned}$$

2.1 Objective Functionals

The objective functional must be properly chosen according to the application. We give three examples of objective functionals, which are used in common applications.

1. **Electric field intensity**
   Let $B \subseteq D$, then the intensity of the scattered electric field E in the sub-region B is measured by

$$J(\Omega, E) = \int_B |E|^2 \, \mathrm{d}x.$$

   For example this functional is used to minimize the scattered field in a specific region. Modifying the intensity function to a tracking type functional, it also models an inverse scattering problem.
2. **Energy flux**
   For optimization of nano-optical circuits, the energy flow through a monitoring surface is of interest. Let ∂B be the boundary of a domain $B \subset D$ with outer unit normal vector n. The energy flow of an electromagnetic field with continuously differentiable weight function η is defined over the time averaged Poynting vector S [9]:

$$J(\Omega, E) = \int_{\partial B} \eta S \cdot n \, \mathrm{d}x = \frac{1}{2} \int_{\partial B} \eta \mathrm{Re}\,(E \times \bar{H}) \cdot n \, \mathrm{d}x \tag{2.3}$$

   Furthermore the energy flow can be transferred to a domain integral in B by extension of η and integration by parts. With a properly chosen region B and weight function η, this functional is used for the maximization of the energy flow inside a waveguide.
3. **Extinction cross section of a particle *P***
   The extinction cross section is the wavelength dependent section of light which is interacting with the particle—either scattered or absorbed. The so-called extinction spectrum of a particle is in most applications the averaged extinction over all possible orientations of the particle. Obviously, the mean value of all particle orientations is equivalent to average over all illumination directions and polarizations. In low particle concentrations ρ multiple scattering can be neglected as the spectra scale linear with ρ.

$$W^{\mathrm{ext}}(d, p, \lambda) = \int_{\partial B} S^{\mathrm{ext}}(d, p, \lambda) \cdot n \, \mathrm{d}\omega \tag{2.4}$$

$$\Theta(\lambda) = \frac{1}{4\pi^2} \int_{S^2} \int_U W^{\mathrm{ext}}(d, p, \lambda) \, \mathrm{d}p \, \mathrm{d}d \tag{2.5}$$

$$J(\Omega) = \min_{\rho \geq 0} \int_\Lambda (\Theta_D(\lambda) - \rho\Theta(\lambda))^2 \, \mathrm{d}\lambda \tag{2.6}$$

The extinction cross section W^{ext} (2.4) of a particle P illuminated by a plane wave with direction d and polarization p is calculated by integration of the extinction part S^{ext} of the time-averaged pointing vector over the surface of a sphere B enclosing the particle $P \subset B$ ([7, 18]). By averaging over all illumination directions and polarizations, the extinction spectrum Θ (2.5) is computed. Finally in (2.6), the squared L^2-error on the wavelength set Λ of the spectrum Θ to the desired spectrum Θ_D is measured for the optimal concentration ρ. Of course, the optimal concentration can be directly computed

$$\rho^*(\Theta, \Theta_D) = \Big(\int_\Lambda (\Theta_D(\lambda)\Theta(\lambda))\, \mathrm{d}\lambda \Big) \Big(\int_\Lambda \Theta(\lambda)^2\, \mathrm{d}\lambda \Big)^{-1}.$$

Thus, (2.6) can be rewritten as

$$J(\Omega) = \int_\Lambda (\Theta_D(\lambda) - \rho^*(\Theta, \Theta_D)\Theta(\lambda))^2\, \mathrm{d}\lambda. \tag{2.7}$$

2.2 Adjoint Calculus

By using an adjoint approach we take the PDE constraint in (2.1) into account. The PDE constraint will also be referred to primal problem. Let the Lagrangian function $\mathcal{L}$ be a function of shape $\Omega := (\Omega_1, \ldots, \Omega_N)$ and two function $u,\ p \in H(\text{curl}\,, \mathbb{R}^3)$. It is specified by

$$\mathcal{L}(\Omega, u, p) = J(\Omega, u) - (\mu_r^{-1}\text{curl}\, u, \text{curl}\, p)_{\mathcal{B}} + \kappa^2(\varepsilon_r u, p)_{\mathcal{B}} - (F, p)_D$$

where $F = \kappa^2(\varepsilon_r - \varepsilon_b)E_I$ is the given right hand side of the scattering problem (1.11). The parentheses denote the real scalar product in L^2 of two complex-valued vector functions over the domain $\mathcal{B}$

$$(u, v)_{\mathcal{B}} = \int_{\mathcal{B}} \text{Re}\, (u \cdot v)\, \mathrm{d}x.$$

As necessary condition for an optimal shape, all partial derivative of the Lagrangian function $\mathcal{L}$ have to vanish. If the adjoint function $P \in H_0(\text{curl}\,, \mathcal{B})$ is a solution of the weak problem

$$(\mu_r^{-1}\text{curl}\, P, \text{curl}\, \varphi)_{\mathcal{B}} - \kappa^2(\varepsilon_r P, \varphi)_{\mathcal{B}} = \partial_u J(\Omega)[\varphi] \quad \forall \varphi \in H_0(\text{curl}\,, \mathcal{B}), \tag{2.8}$$

then the partial derivatives with respect to u and p of the Lagrangian function $\mathcal{L}(\Omega, E, P)$ vanish for $E \in H_0(\text{curl}\,, D)$ which is solution of the primal problem. The partial derivative with respect to u in a direction φ of the objective functional is located on the right hand side of the adjoint problem and denoted by $\partial_u J(\Omega)[\varphi]$.

2.3 Shape Calculus

Let $T_t : \mathcal{M} \to \mathcal{M}_t$ be a one-to-one transformation from the reference domain $\mathcal{M} \subset \mathbb{R}^3$ to a perturbed domain $\mathcal{M}_t \subset \mathbb{R}^3$ by perturbation of identity [2, 11, 12, 28]

$$T_t(x) = x + tV(x)$$

with the so called velocity field $V \in (C^1(\mathcal{M}))^3$ and a small non-negative number t. Furthermore, let u_t be a function in the domain $\mathcal{M}_t$, i.e. the solution of the Maxwell's equation. Then u_t is transported to the reference domain $\mathcal{M}$ by the transformation T_t with

$$\bar{u}(x) = DT_t^T u_t(T_t(x)) \text{ and curl } \bar{u}(x) = J_t DT_t{}^{-1} \text{curl } u_t(T_t(x)) \; \forall x \in \mathcal{M}, \qquad (2.9)$$

where DT_t is the Jacobian matrix of T_t and J_t the determinant of DT_t, see Fig. 3. By a simple computation we see that the transported function $\bar{u}$ is also in $H(\text{curl}\,, \mathcal{M})$ if u_t was in $H(\text{curl}\,, \mathcal{M}_t)$. For shape sensitivity analysis we need the sensitivity of the function $\bar{u}$ on the parameter t, which is equivalent to the material derivative $\dot{u}$ of the transported function $\bar{u}$. Thus, there is the relation

$$\dot{u}(x) = \frac{\mathrm{d}}{\mathrm{d}t}\bar{u}\bigg|_{t=0} = DV^T \cdot u + u' + Du \cdot V$$

where u' is the shape derivative of $u = u_0$ and Du denotes the Jacobian of u. The shape derivative u' vanishes if the function u is independent of the shape $\mathcal{M}$, which will be applied for the incident field of the scattering problem. In the following, useful identities regarding the material derivative are provided:

$$\begin{aligned} &T_t \to V, && DT_t \to DV, && DT_t{}^{-1} \to -DV, \\ &J_t \to \text{div}\,(V), && DT_t^T \to DV^T, && DT_t{}^{-T} \to -DV^T. \end{aligned}$$

In order to obtain a partial differential equation for the material derivative $\dot{E}$ of the electric field, we differentiate the terms of the primal weak form (2.2) with respect to the shape. By testing (2.2) with a function $\varphi_t \in H(\text{curl}\,, \mathcal{M}_t)$ which is constructed

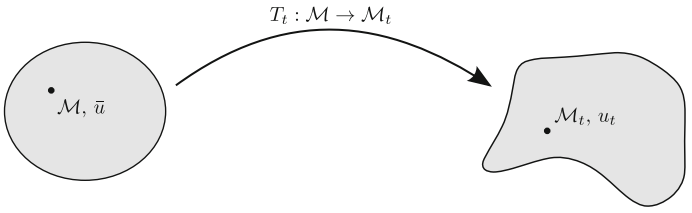


Fig. 3 Transformation from reference domain Ω to perturbed domain Ω_t

by $\varphi_t = DT_t^{-T}\varphi \circ T_t^{-1}$ over a function $\varphi \in H(\text{curl}\,, \mathcal{M})$, we derive

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{M}_t} \text{curl}\, E_t \cdot \text{curl}\, \varphi_t \,\mathrm{d}x_t\Big|_{t=0} &= \int_{\mathcal{M}} (\text{curl}\, \dot{E} + A\text{curl}\, E) \cdot \text{curl}\, \varphi \,\mathrm{d}x,\\
\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{M}_t} E_t \cdot \varphi_t \,\mathrm{d}x_t\Big|_{t=0} &= \int_{\mathcal{M}} (\dot{E} - AE) \cdot \varphi \,\mathrm{d}x,\\
\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{M}_t} F_t \cdot \varphi_t \,\mathrm{d}x_t\Big|_{t=0} &= \int_{\mathcal{M}} (DV^T \cdot F + DF \cdot V - AF) \cdot \varphi \,\mathrm{d}x,
\end{aligned}$$

where $A = DV^T + DV - \text{div}\,(V)\mathbb{1}$. To derive a weak system for the material derivative $\dot{E}$, we apply those rules to each element of the perturbed shape $\Omega_t = (\Omega_1^t, \ldots, \Omega_N^t)$. Thanks to the transformation (2.9), we simply combine the result to Ω, since the tangential continuity conditions in (2.1) are fulfilled. The weak formulation for $\dot{E}$ reads for a deformation field $V|_{\Omega_B} = 0$ as

$$\begin{aligned}
&\text{Find } \dot{E} \in H_0(\text{curl}\,, \mathcal{B}) \text{ s.t. for all } \varphi \in H_0(\text{curl}\,, \mathcal{B})\\
&(\mu_r^{-1}\text{curl}\, \dot{E}, \text{curl}\, \varphi)_{\mathcal{B}} - \kappa^2(\varepsilon_r \dot{E}, \varphi)_{\mathcal{B}} =\\
&\quad -(A\text{curl}\, E, \text{curl}\, \varphi)_D - \kappa^2(\varepsilon_r AE, \varphi)_D + (DV^T \cdot F + DF \cdot V - AF, \varphi)_D.
\end{aligned} \tag{2.10}$$

We do not solve (2.10) explicitly, because we use this result and the adjoint system to write down the volume representation of the shape derivative of the objective function in direction V with $V|_{\Omega_B} = 0$

$$\begin{aligned}
\mathrm{d}J(\Omega, V) = \partial_\Omega J(\Omega) &+ (A\text{curl}\, E, \text{curl}\, P)_D + \kappa^2(\varepsilon_r AE, P)_D\\
&- (DV^T \cdot F + DF \cdot V - AF, P)_D.
\end{aligned}$$

If the objective functional depends explicitly on the shape then the shape derivative consist of a term $\partial_\Omega J(\Omega)$, which has to be determined for each objective function individually. The remaining parts apply to any objective function.

To obtain a continuous velocity field with negative shape derivative, we use an H^1-projection by solving

$$\begin{aligned}
-\Delta V + \eta^2 V &= -\mathrm{d}J \qquad \text{in } D\\
V &= 0 \qquad \text{on } \partial D
\end{aligned} \tag{2.11}$$

with a positive scalar $\eta > 0$. Furthermore, some components V_i of the deformation field V may be set to zero on interfaces $\mathcal{G}_i$, i.e. $V_i = 0$ on $\mathcal{G}_i$, which realize the fixing of boundary parts. It is clear that this projection leads to a descent direction of the objective functional for all $\eta > 0$ as long as there exists descent direction under movement restrictions. Furthermore, it is obvious that there exists a scalar $0 < \tau \leq 1$ such that $\tau V \in \mathcal{V}$ and is a descent direction. Based on this information, a gradient-based algorithm can be obtained.

Algorithm 1 Shape gradient algorithm for electromagnetic applications

Initial domain Ω_0, $i = 0$
repeat
  $E \leftarrow$ solving primal system
  $J \leftarrow$ evaluation objective function
  $P \leftarrow$ solving adjoint system
  $G \leftarrow$ evaluating shape gradient
  $V \leftarrow$ computing velocity field
  $\tau \leftarrow$ step size control
  $\Omega_{i+1} = \Omega_i + \tau V$
  $i = i + 1$
until $\mathrm{d}J < TOL$

2.4 Algorithm

In this subsection we give a brief overview over the gradient-based shape optimization algorithm for electromagnetic applications, see Algorithm 1. Starting with an admissible shape Ω_0 the shape is iteratively updated until a stopping criterion is reached. First the shape is tessellated with suitable mesh generator, since a finite element method is used to approximate Maxwell's equation in the examples of Sect. 3. After the discretization of the domain, the primal problem of (2.1) is solved (E) and the objective functional J can be evaluated. With the solution P of the discretized adjoint problem (2.8), the right hand side of (2.11) is assembled (G). The projection (2.11) and the resulting deformation field V leads to a descent of the objective functional by updating Ω_i to Ω_{i+1} with a properly chosen step-size $\tau \leq 1$.

3 Numerical Experiments

In this section we demonstrate shape optimization in two applications in context of electromagnetic fields. In the first example an extinction spectrum of a particle is prescribed and the distance to this desired spectrum is minimized by changing the shape of the particle, c.f. objective functional (2.7). The maximization of the incoupling efficiency of a nano-optical antenna is considered in the second configuration. Therefore the objective function (2.3) is used as the aim for this subject.

In both applications the electric fields are approximated by the finite element method with Nédélec basis functions (curl-conforming tangential vector basis function [3, 21, 24]) on tetrahedral elements. Those induced function-spaces inherently fulfill the tangential continuity condition of the electric field in (2.1). Algorithm 1 is implemented in an own MATLAB [30] package exploiting the parallel computation capabilities of LAPACK [5]. The finite element mesh is generated by the open-source

tetrahedral mesh generator TETGEN [27] and the arising linear system of equation is solved by HSL_MA87 [26], a direct solver for complex symmetric indefinite sparse matrices. Both software packages were integrated into MATLAB over the C/C++ MEX-file creation API.

3.1 Experiment 1

The first example deals with the tracking type functional of the extinction spectrum of a particle (2.7). To verify the method, we choose as target spectrum Θ_D the extinction spectrum of an ellipsoid and start with a deformed shape.

The transformation of the surface integral (2.4) to a numerically preferable volume integral is done by using Gauss Theorem and the identities of (1.6) as follows

$$\begin{aligned} W^{\text{ext}} &= \int_{\partial B} S_h^{\text{ext}} \cdot n \, \mathrm{d}\omega = \frac{1}{2} \int_{\partial B} \operatorname{Re}\left(E \times \bar{H}_I + E_I \times \bar{H}\right) \cdot n \, \mathrm{d}\omega \\ &= \frac{1}{2} \int_B \operatorname{Re}\left(\bar{H}_I \cdot \operatorname{curl} E - E \cdot \operatorname{curl} \bar{H}_I + \bar{H} \cdot \operatorname{curl} E_I - E_I \cdot \operatorname{curl} \bar{H}\right) \mathrm{d}x \\ &= \frac{\kappa}{2} \int_B \operatorname{Im}\left((\bar{\varepsilon}_r - \varepsilon_b) E_I \cdot \bar{E}_T\right) \mathrm{d}x. \end{aligned} \tag{3.1}$$

The outer integral in equation (2.5) is discretized by choosing 32 distributed directions on one half sphere Fig. 4. The extinction value of a particle is independent of the sign chosen in the exponential term of (1.2) and thus equivalent for the directions d and $-d$. Due to the linearity of Maxwell equations the inner integral over the polarizations can be calculated directly by computing the extinction value for two orthogonal polarizations and taking the mean value. Thus solving Maxwell equations for 32 illumination directions and two orthogonal polarizations each, gives us, up to numerical errors, an exact value of the inner integral for 128 incident waves. Integrating over the wavelength range in equation (2.6) is also done numerically by a linear interpolation between wavelengths. Thus, we define with (3.1) the extinction cross section for discrete illumination directions d_i, polarizations p_j and wavelengths λ_k

$$W^{\text{ext}}_{i,j,k} := \frac{\kappa}{2} \int_B \operatorname{Im}\left((\bar{\varepsilon}_r - \varepsilon_b)\, E_I(d_i, p_j, \lambda_k) \cdot \bar{E}_T(d_i, p_j, \lambda_k)\right) \mathrm{d}x.$$

Finally, the objective functional in (2.1) for the tracking of an extinction spectrum reads as

$$J(\Omega) = 1 - \frac{\langle \Theta_D, \Theta \rangle_\alpha^2}{\langle \Theta, \Theta \rangle_\alpha}$$

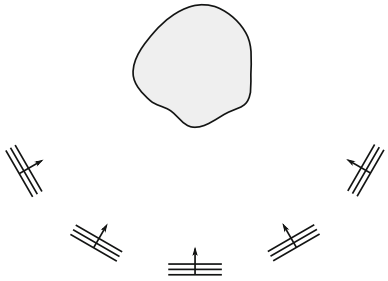


Fig. 4 Setting of example 1. The gray region is the illuminated particle, the *black lines* and *arrows* symbolize the different incident waves

where $\langle \cdot, \cdot \rangle_\alpha$ represents the discretized integration over the wavelength range with quadrature weights α, i.e.

$$\langle a, b \rangle_\alpha := \sum_{k=0}^{M} \alpha_k a_k b_k \quad \text{with} \sum_{k=0}^{M} \alpha_k = |\Lambda|.$$

The averaged extinction cross section at a wavelength λ_k is defined by

$$\Theta_k := \sum_{i=1}^{N} \frac{\nu_i}{2} \left(W_{i,1,k}^{\text{ext}} + W_{i,2,k}^{\text{ext}} \right) \quad \text{with} \sum_{i=0}^{N} \nu_i = 1$$

which includes the weighted sum over all illumination direction with weights ν and the mean between two polarizations. Furthermore, the target extinction spectrum Θ_D is normalized, i.e. $\langle \Theta_D, \Theta_D \rangle_\alpha = 1$.

In this example, the particle is made of gold and is imbedded in water with the two refractive indices given in [14] and [10] respectively. The ellipsoidal target particle has two half axis with 10 nm and one with 20 µm length. To solve the primal and the adjoint equation we use tetrahedral elements and second-order Nédélec finite elements which results in an indefinite system of equations with more then 10^6 unknowns. In each step, this system with 64 right-hand-sides has to be solved twice for each wavelength. For the wavelength range we chose the visible part of the spectrum, i.e. 0.4 µm up to 0.7 µm. This range is discretized in 30 equidistant intervals.

In Fig. 5, the shape of initial particle and the final particle are visualized. The visible difference between the final particle and the desired ellipsoid is mainly due to the investigated wavelength range. Figure 6 depicts the progression of the objective function in logarithmic scale and Fig. 7 shows the extinction spectrum of the ellipsoid (green) and the initial shape (blue). At the end of the optimization, these two lines do not differ with the naked eye.

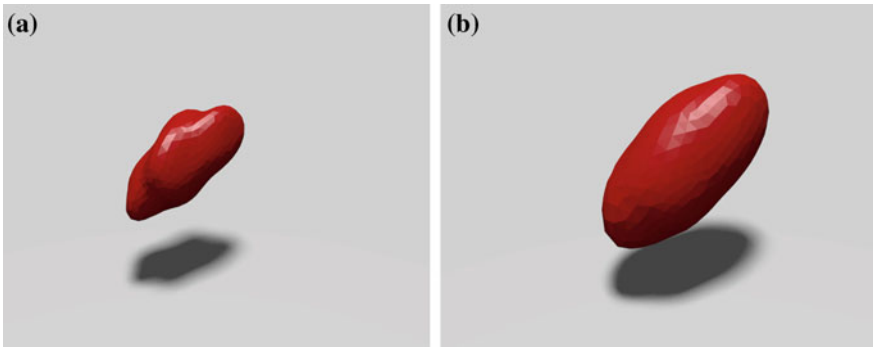


Fig. 5 Visualization of the particle, **a** initial particle, **b** final particle

Fig. 6 Evolution of the objective functional over iteration time

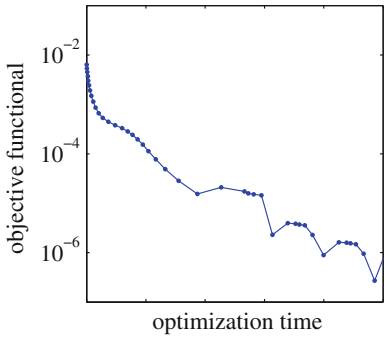
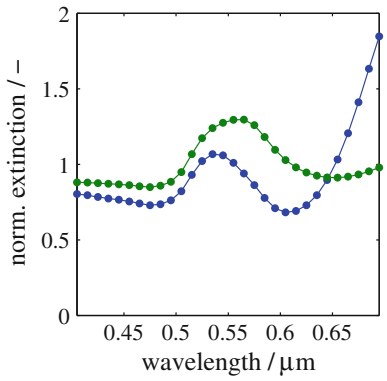


Fig. 7 The target spectrum (*green*) and the initial spectrum (*blue*). The final spectrum has in this scaling no visible difference to the target spectrum



3.2 Experiment 2

Nanoantennas are one component of plasmonic circuits where the efficient use of energy is a necessary step towards enhanced computer chips. Plasmonic based chips will be significantly smaller and faster compared to the current silicon based technology. By numerical and practical experiments it is shown that shape of a nanoantenna influences its efficiency [4, 15].

In this example the incoupling efficiency of a gold (Au) nanoantenna inside a silica (SiO_2) block is studied, see a visualization in Fig. 9. Moreover, the antenna is attached to slot waveguide with a width of 0.3 µm and both are extruded by 0.22 µm. This design was introduced by [15] as Yagi-Uda type antenna. The geometry is associated with a Cartesian grid, where the x-axis is parallel to the waveguide direction and the z-axis is parallel to the extrusion direction, c.f. Fig. 8. A quasi-Gaussian laser beam, propagating in positive z direction, is incoupled by the antenna to a guided wave. The laser emits light with a vacuum wavelength of $\lambda = 1.55\,\mu m$, is linear polarized along y-axis and is highly focused to a waist size of 0.75 µm. The model of the Gaussian beam is described by [31] and slightly modified using Fourier transform for faster numerical evaluation. Furthermore, the laser beam is aligned to the smallest gap of the initial shape of the nanoantenna at $x = 0.6\,\mu m$. The refractive indices at a vacuum wavelength of 1.55 µm of the used materials are determined by [14] and [23] for gold and silica respectively.

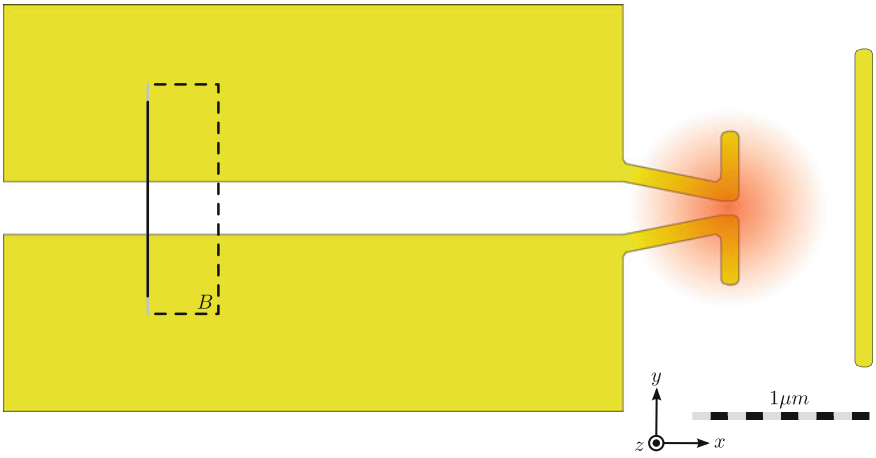


Fig. 8 True to scale geometrical setting of the incoupling efficiency optimization of a nanoantenna (*yellow*) with the incident laser beam position (*red*). The weight function η is equal to one on the *solid line* and transits smoothly on the *gray line* to zero on the *dashed line*

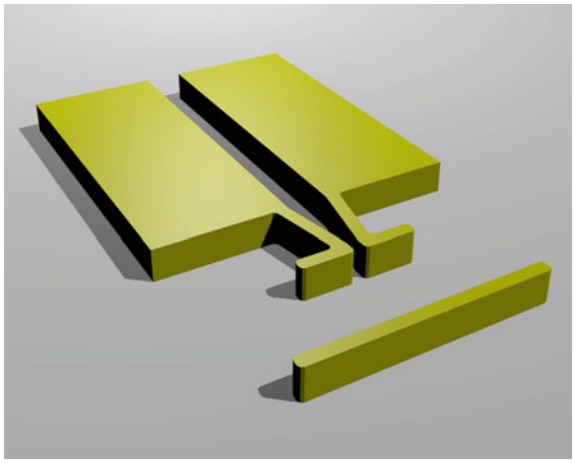


Fig. 9 Three-dimensional illustration of antenna attached to slot waveguide

The incoupling efficiency is the ratio between the energy flow through a rectangular monitor inside the waveguide and the incident beam energy. The weight function $\eta \in C^1(\mathbb{R}^3)$ in (2.3) is chosen to respect the energy flow near the waveguide, i.e. $\eta = 1$, at the position $x = -2.75\,\mu\text{m}$ and neglect the energy apart, i.e. $\eta = 0$, see Fig. 8. Furthermore, the waveguide width, the extrusion thickness and the observation area B are fixed during optimization. After normalization of the incident beam power and switching to a minimization problem by changing the sign, the cost functional in (2.1) reads as

$$J(\Omega, E) = \frac{1}{2\kappa} \int_B \nabla\eta \cdot \text{Im}\,(E_T \times \text{curl}\,\bar{E}_T) + \kappa^2 \text{Im}\,(\varepsilon_r)\eta |E_T|^2 \,\mathrm{d}x.$$

The evolution of the normalized objective functional is depicted in Fig. 10. Note that the efficiency increases by 2.3 times in comparison to the initial value. In Fig. 11, both the y-component of the electric field E_y (lower half) and the magnitude of x-component of the energy flow $|S_x|$ (lower half) in the initial shape (Fig. 11a) are compared to the respective field quantities in the optimized domain (Fig. 11b). The colors show small values in blue and large values in red. So the effect of the shape optimization on the efficiency of the nanoantenna is qualitatively visible inside the waveguide. Since the initial shape and the incident beam are symmetric to the xz-plane, the final shape has the same symmetry as well.

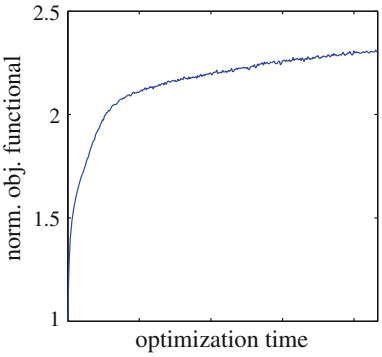


Fig. 10 Normalized incoupling efficiency over optimization time

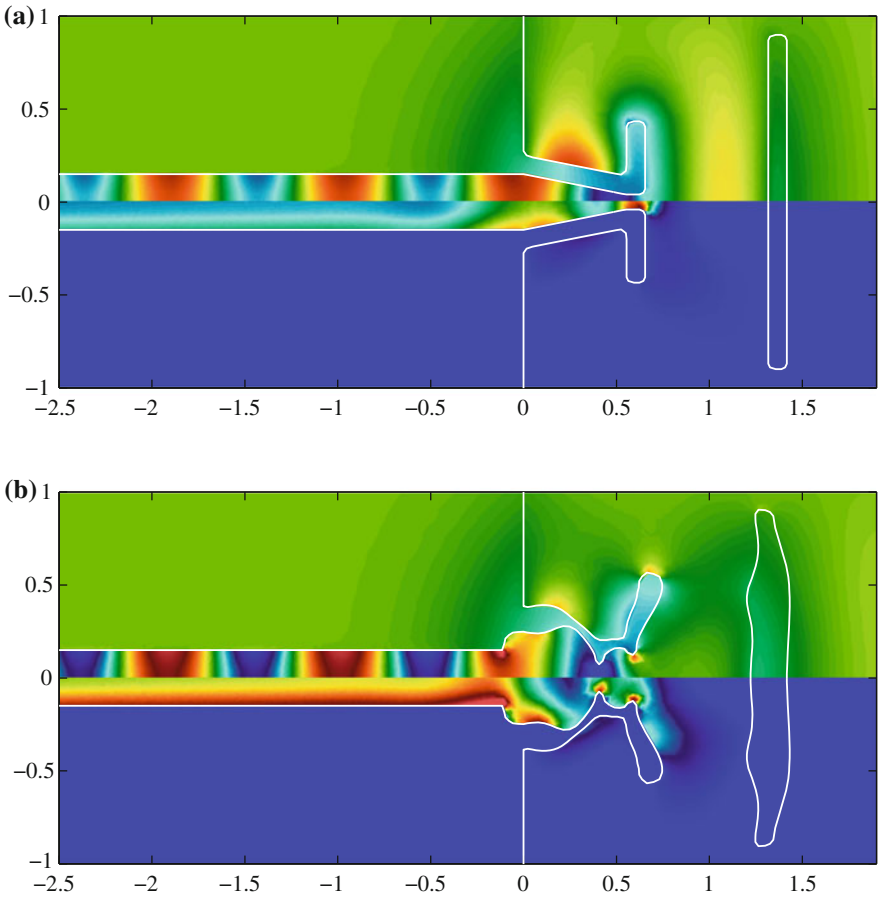


Fig. 11 Horizontal slice though the center of the geometry with combined visualization of the y-component of the scattered electric field E_y (*upper half*) and magnitude of x-component of the Poynting vector $|S_x|$ (*lower half*). **a** Initial domain, **b** Optimized domain

4 Concluding Remarks

We derived a gradient descent algorithm for the optimization of nanooptical objects based on the volume representation of the shape gradient. In our setting, the physical situation is modeled by the time-harmonic formulation of Maxwell's equations. With aid of adjoint calculus and the consequently partial differential equation for the adjoint state, the volume representation of the shape gradient was deduced.

Moreover, two application for the shape optimization method were presented. The results, which show the high sensitivity of scattering phenomena with respect shape changes, give insight into the large potential of shape optimization for the design of nanooptical objects. Providing innovative ideas for the design of nanoantennas or particles, shape optimization supports the current practical and theoretical researches in various research areas like physics and chemical engineering.

References

1. V. Akcelik, G. Biros, O. Ghattas, D. Keyes, K. Ko, L.-Q. Lee, E.G. Ng, Adjoint methods for electromagnetic shape optimization of the low-loss cavity for the international linear collider. J. Phys. Conf. Ser. **16**, 435–445 (2005). ISSN 1742-6588. doi:10.1088/1742-6596/16/1/059
2. G. Allaire, *Conception Optimale de Structures* (Springer, Berlin, 2006)
3. L. Andersen, J. Volakis, Hierarchical tangential vector finite elements for tetrahedra. IEEE Microw. Guid. Wave Lett. **8**(3), 127–129, (1998). ISSN 10518207. doi:10.1109/75.661137
4. A. Andryieuski, R. Malureanu, G. Biagi, T. Holmgaard, A. Lavrinenko, Compact dipole nanoantenna coupler to plasmonic slot waveguide. Opt. Lett. **37**(6), 1124–1126 (2012). ISSN 1539-4794. doi:10.1364/OL.37.001124
5. E. Angerson, Z. Bai, J. Dongarra, A. Greenbaum, A. McKenney, J. Du Croz, S. Hammarling, J. Demmel, C. Bischof, D. Sorensen, LAPACK: a portable linear algebra library for high-performance computers, in *Proceedings SUPERCOMPUTING'90* (IEEE Computer Society Press, 1990), pp. 2–11. ISBN 0-8186-2056-0. doi:10.1109/SUPERC.1990.129995
6. J.-P. Berenger, A perfectly matched layer for the absorption of electromagnetic waves. J. Comput. Phys. **114**(2), 185–200 (1994). ISSN 00219991. doi:10.1006/jcph.1994.1159
7. C. Bohren, *Absorption and Scattering of Light by Small Particles* (Wiley, New York, 1983)
8. J. Cagnol, M. Eller, Boundary regularity for Maxwell's equations with applications to shape optimization. J. Differ. Equ. **250**(2), 1114–1136 (2011). ISSN 00220396. doi:10.1016/j.jde.2010.08.004
9. M. Cessenat, *Mathematical Methods in Electromagnetism: Linear Theory and Applications* (World Scientific, Singapore, 1996)
10. M. Daimon, A. Masumura, Measurement of the refractive index of distilled water from the near-infrared region to the ultraviolet region. Appl. Opt. **46**(18), 3811 (2007). ISSN 0003-6935. doi:10.1364/AO.46.003811
11. M. Delfour, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization* (Society for Industrial and Applied Mathematics, Philadelphia, 2001)
12. J. Haslinger, *Introduction to Shape Optimization: Theory, Approximation, and Computation* (SIAM Society for Industrial and Applied Mathematics, Philadelphia, 2003)
13. M. Hintermüller, A. Laurain, I. Yousept, Shape sensitivities for an inverse problem in magnetic induction tomography based on the Eddy current model. *uni-graz.at* (2014)
14. P.B. Johnson, R.W. Christy, Optical constants of the noble metals. Phys. Rev. B **6**(12), 4370–4379 (1972). ISSN 0556-2805. doi:10.1103/PhysRevB.6.4370

15. A. Kriesch, S.P. Burgos, D. Ploss, H. Pfeifer, H.A. Atwater, U. Peschel, Functional plasmonic nanocircuits with low insertion and propagation losses. Nano Lett. **13**(9), 4539–45, 2013. ISSN 1530-6992. doi:10.1021/nl402580c
16. C.M. Lalau-Keraly, S. Bhargava, O.D. Miller, E. Yablonovitch, Adjoint shape optimization applied to electromagnetic design. Opt. Express **21**(18), 21693–21701 (2013). ISSN 1094-4087. doi:10.1364/OE.21.021693
17. P. Li, An inverse cavity problem for Maxwell's equations. J. Differ. Equ. **252**(4), 3209–3225 (2012). ISSN 00220396. doi:10.1016/j.jde.2011.10.023
18. M. Mishchenko, *Scattering, Absorption, and Emission of Light by Small Particles* (Cambridge University Press, Cambridge, 2002)
19. P. Monk, *Finite Element Methods for Maxwell's Equations* (Clarendon Press, Oxford, 2003)
20. C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves* (Springer, Berlin, 1969). ISBN 978-3-662-11775-0. doi:10.1007/978-3-662-11773-6
21. J. C. Nédélec, A new family of mixed finite elements in $\mathbb{R}^3$. Numer. Math. **50**(1), 57–81 (1986). ISSN 0029-599X. doi:10.1007/BF01389668
22. D.M. Nguyen, A. Evgrafov, J. Gravesen. Isogeometric shape optimization for electromagnetic scattering problems. Prog. Electromagn. Res. B. **45**, 117–146 (2012). ISSN 1937-6472. doi:10.2528/PIERB12091308
23. T. Radhakrishnan. Further studies on the temperature variation of the refractive index of crystals. Proc. Indian Acad. Sci. Sect. A, **33**(1), 22–34 (1951). ISSN 0370-0089. doi:10.1007/BF03172255
24. J. Schöberl, S. Zaglmayr, High order Nédélec elements with local complete sequence properties. COMPEL Int. J. Comput. Math. Electr. Electron. Eng. **24**(2), 374–384 (2005). ISSN 0332-1649. doi:10.1108/03321640510586015
25. J. Schwinger, L.L.J. Deraad, K.A. Milton, *Classical Electrodynamics* (Westview Press, Boston, 1998). ISBN 0813346622
26. Science & Technology Facility Council. HSL, a collection of Fortran codes for large scale scientific computation (2013). www.hsl.rl.ac.uk
27. H. Si, TetGen: a quality tetrahedral mesh generator and 3D delaunay triangulator (2013). www.tetgen.org
28. J. Sokolowski, *Introduction to Shape Optimization: Shape Sensitivity Analysis* (Springer, Berlin, 1992). ISBN 9783540541776
29. J. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company Inc., New York, 1941). ISBN 9780070621503
30. The MathWorks Inc. MATLAB Release 2014a (2014). www.mathworks.com
31. P. Varga, P. Török, The Gaussian wave solution of Maxwell's equations and the validity of scalar wave approximation. Opt. Commun. **152**(1–3), 108–118 (1998). doi:10.1016/S0030-4018(98)00092-3

Shape Differentiability Under Non-linear PDE Constraints

Kevin Sturm

Abstract We review available methods to compute shape sensitivities and apply these methods to a semi-linear model problem. This will reveal the difficulties of each method and will help to decide which approach should be used for a concrete applications.

Keywords Lagrange method · Shape derivative · Non-linear PDE · Material derivative · Céa's method · Minimax formulation

AMS 49Q10 · 49Q12

1 Introduction

The objective of this manuscript is to give readers an overview on methods that allow to derive the shape differentiability of PDE (partial differential equation) constrained shape functions. There are several methods available to prove the shape differentiability of shape functions depending on the solution of a PDE. In the recent past two new methods have been proposed: the rearrangement method [14] and an approach using a novel adjoint equation [19]. Other more established methods comprise the material/shape derivative method [18] (also called 'chain rule' approach), the min approach for energy cost functions [7], the minimax approach of [9] and an interesting penalization method [8] to derive sensitivities for a class of variational inequalities. The approach of Céa [5] is frequently used to derive the formulas for the shape derivative, but itself gives no proof for the shape differentiability. Indeed, there are cases where Céa's method fails; cf. [17, 19]. For linear partial differential equations and (semi)-convex cost functions all mentioned methods (except Céa's Lagrange method in some cases) work and the necessary assumptions are readily

K. Sturm (✉)
Fakultät für Mathematik, Universität Duisburg-Essen,
Thea-Leymann-Straße 9, 45127 Essen, Germany
e-mail: kevin.sturm@uni-due.de

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_12

verified. But for non-linear PDEs the situation is quite different as we will see in the presented example. After reading this article the reader may decide which method is suited for his or her problem in hand .

In particular the presented methods are:

- The material derivative method analyzes the sensitivity of the solution of the PDE with respect to the domain. This procedure is similar to the direct approach used in PDE constraint optimal control [21]. Here, the solution of the PDE depends on a control function, which belongs to a usually convex set. The main objective when deriving the necessary optimality conditions is the investigation of the control-solution operator. In shape optimization we have to investigate the "domain-state" operator. The investigation of shape function is more involved, since spaces of shapes admit no vector space structure.
- The rearrangement method exploits a first order expansion of the PDE and the cost function with a remainder which tends to zero with order two. This expansion is combined with the Hölder continuity of the domain-state operator. The main challenge of this method constitute the proof of the Hölder continuity, but more importantly the first order expansion.
- In the minimax approach the cost function is expressed as a minimax of the Lagrangian associated to the optimization problem. By definition a Lagrangian is a function that is the sum of a utility function and a state equation. The problem of the differentiability of the cost function is shifted to the differentiability of a minimax function. The Theorem of Correa-Seeger [6] can be applied to prove the differentiability if (among other requirements) the Lagrangian admits saddle points. A special case of this approach is when the cost function is itself a minimum of an energy. In this case the minimax of the Lagrangian is replaced by the min of the energy and we have to investigate the differentiability of the min function to prove the shape differentiability.
- The averaged adjoint approach can also be seen as a proof for the differentiability of a minimax function. Unlike the Theorem of Correa-Seeger it requires no saddle point assumption. Therefore it constitutes an extension of the Theorem of Correa-Seeger for the special class of Lagrangian functions.

The manuscript is organized as follows:

Section 2, the basic notation is introduced and basic tools from shape optimization, including the Zolésio-Hadamard structure theorem, are recalled. We introduce the basic model problem and make some basic assumptions.

Section 3, the existence of the strong material derivatives associated to this equation is shown under suitable assumption. This proves then the shape differentiability of the cost function.

Section 4, the minimax formulation is reviewed for the particular example. Then the Theorem of Correa-Seeger is applied to prove the differentiability of the minimax function with respect to a parameter, that is, the shape differentiability of the cost function.

Section 5, the rearrangement method is employed to derive the shape differentiability of the semi-linear model problem.

Section 6, the shape differentiability of a special cost function, that is the energy associated to the PDE, is proved. In this case the minimax differentiability reduces to the differentiability of a min function.

Section 7, a recently proposed approach of the averaged adjoint equation is presented and applied to the semi-linear problem.

2 Notations and Problem Description

2.1 Notation

Let E and F be Banach spaces and $U \subset E$ an open subset. We denote by $C(U; F)$ the space of all continuous functions $f : U \to F$. We call a function $f : U \to F$ differentiable in $x \in U$ if it is Fréchet differentiable at x and denote the derivative by $\partial f(x)$. The function is called differentiable if it is differentiable at every point $x \in U$. For $k \geq 1$, the space of all k-times continuously differentiable functions $f : U \to F$ is denoted by $C^k(U; F)$. The directional derivative of f at x in direction v is denoted by $df(x; v)$. When $F = \mathbf{R}$ and $E = \mathbf{R}^d$, we adopt the notation $C^k(\overline{U}; \mathbf{R}^d)$ of [23] for all those functions $f \in C^k(U; F)$ that admit extendable partial derivative $\partial^\alpha f$ to $\overline{U}$ for all indices $\alpha = (\alpha_1, \ldots, \alpha_d)$ with $|\alpha| \leq k$. Also, we identify the derivative $\partial f(x) : \mathbf{R}^d \to \mathbf{R}$ via the Riesz representation theorem by the gradient $\nabla f(x)$, which is for each point $x \in \mathbf{R}^d$ a vector in $\mathbf{R}^d$. For $p \geq 1$, the space of all measurable functions $f : \Omega \to \mathbf{R}$ for which $\|f\|_{L_p(\Omega)} := \left(\int_\Omega |f|^p \, dx\right)^{1/p} < \infty$ is denoted by $L_p(\Omega)$. The space of functions of bounded variations on D is denoted by $BV(D)$. For the one-sided limit (t approaches zero from the right) we write $\lim_{t \searrow 0}$. The right derivative in zero of a function $f : U \subset \mathbf{R} \to \mathbf{R}$ is denoted $f(0^+) := \lim_{t \searrow 0}(f(t) - f(0))/t$.

2.2 The Problem Description

Let $d \in \mathbf{N}^+$. Throughout this manuscript, we consider the following semi-linear **state equation**

$$-\Delta u + \varrho(u) = f \ \text{ in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega. \tag{2.1}$$

on a domain $\Omega \subset \mathbf{R}^d$. The function $u : \Omega \to \mathbf{R}$ is called **state** and $f : D \to \mathbf{R}$ is a function specified below. Without loss of generality, we may assume $\varrho(0) = 0$ otherwise consider $\tilde{\varrho}(x) := \varrho(x) - \varrho(0)$ with right hand side $\tilde{f}(x) := f(x) - \varrho(0)$. To simplify the exposition, we choose as objective function

$$J(\Omega) := \int_\Omega |u - u_r|^2 \, dx, \tag{2.2}$$

where $u_r : \overline{D} \to \mathbf{R}$ is given and $|\ |$ denotes the absolute value. The task is now to derivative the shape derivative of the cost function (2.2) by employing different techniques.

Throughout this manuscript we suppose that the following assumption is satisfied.

Assumption (*Data*)

(i) Let $\Omega \subset D \subset \mathbf{R}^d$ be two simply connected domains with Lipschitz boundaries $\partial\Omega$ and ∂D, respectively.
(ii) The functions u_r, $f : \overline{D} \to \mathbf{R}$ are continuously differentiable.
(iii) The vector field θ belongs to $C_c^2(D, \mathbf{R}^d)$.

For any $k \geq 1$, we define the space

$$\mathcal{C}_c^k(D, \mathbf{R}^d) := \{\theta \in C^k(\mathbf{R}^d; \mathbf{R}^d) : \ \mathrm{supp}(\theta) \subset D\}$$

and set $\mathcal{C}_c^\infty(D, \mathbf{R}^d) = \bigcap_{n\in\mathbf{N}} \mathcal{C}_c^n(D, \mathbf{R}^d)$. The **flow** of a vector field $\theta \in \mathcal{C}_c^k(D, \mathbf{R}^d)$ is defined for each $x_0 \in D$ by $\Phi_t^\theta(x_0) := x(t)$, where $x : [0, \tau] \to \mathbf{R}^d$ solves

$$\dot{x}(t) = \theta(x(t)) \quad \text{in } (0, \tau), \quad x(0) = x_0.$$

In the sequel, we write Φ_t instead of Φ_t^θ.

2.3 Compositions of Functions with Flows

In the following let $\theta \in C_c^1(D, \mathbf{R}^d)$ be a given vector field and $\Phi_t = \Phi_t^\theta$ its associated flow. First, note that by the chain rule $\partial\Phi^{-1}(t, \Phi(t, x)) = (\partial\Phi(t, x))^{-1}$ or briefly $(\partial(\Phi_t^{-1})) \circ \Phi_t = (\partial\Phi_t)^{-1} =: \partial\Phi_t^{-1}$, which implies[1]

$$(\nabla f) \circ \Phi_t = \partial\Phi_t^{-T} \nabla(f \circ \Phi_t).$$

Subsequently the following abbreviations are used

$$\xi(t) := \det(\partial\Phi_t), \qquad A(t) := \xi(t)\partial\Phi_t^{-1}\partial\Phi_t^{-T}, \quad B(t) := \partial\Phi_t^{-T}, \tag{2.3}$$

where $\det : \mathbf{R}^{d,d} \to \mathbf{R}$ denotes the determinant. Step-by-step, we will derive properties of the quantities ξ, B and A.

[1] For any scalar function $f \in H^1(\mathbf{R}^d)$, we have for all $v \in \mathbf{R}^d$ and all $x \in D$

$$\partial(f(\Phi_t(x))v = \partial f(\Phi_t(x))\partial\Phi_t(x)v = \nabla f(\Phi_t(x)) \cdot \partial\Phi_t(x)v = (\partial\Phi_t(x))^T \nabla(f(\Phi_t(x))) \cdot v.$$

Proposition 2.1 *Let a continuous mapping $A \in C([0,\tau]; C(\overline{D}; \mathbf{R}^{d,d}))$ and a function $\xi \in C([0,\tau]; C(\overline{D}))$ be given and assume $A(0) = I$ and $\xi(0) = 1$. Then there are constants $\gamma_1, \gamma_2, \delta_1, \delta_2 > 0$ and $\tau > 0$ such that*

$$\forall \zeta \in \mathbf{R}^d,\ \forall t \in [0,\tau]: \quad \gamma_1|\zeta|^2 \le \zeta \cdot A(t)\zeta \le \gamma_2|\zeta|^2, \qquad (a)$$
$$\delta_1 \le \xi(t) \le \delta_2. \qquad (b)$$

Proof (a) We can estimate

$$\begin{aligned}|\eta|^2 &= (I - A(t))\eta \cdot \eta + A(t)\eta \cdot \eta \\ &\le \|I - A(t)\|_{C(D;\mathbf{R}^{d,d})}\eta \cdot \eta + A(t)\eta \cdot \eta.\end{aligned}$$

By continuity of $t \mapsto A(t)$ there exists for all $\varepsilon > 0$, a $\delta > 0$ such that for all $t \in [0,\delta]$ we have $\|I - A(t)\|_{C(D;\mathbf{R}^{d,d})} \le \varepsilon$. From this the claim follows.
(b) This is clear. □

Proposition 2.2 *Let $B : [0,\tau] \to \mathbf{R}^{d,d}$ be a bounded mapping such that $\|B^{-1}(t)\|_{C(\overline{D};\mathbf{R}^{d,d})} \le C$ for all $t \in [0,\tau]$ for some constant $C > 0$. Then for any $p \ge 1$ there is a constant $C > 0$ such that*

$$\forall t \in [0,\tau],\ \forall f \in W_p^1(D): \ \|\nabla f\|_{L_p(D;\mathbf{R}^d)} \le C\|B(t)\nabla f\|_{L_p(D;\mathbf{R}^d)}$$

Proof Estimating

$$\|\nabla f\|_{L_p(D;\mathbf{R}^d)} = \|(B(t))^{-1}B(t)\nabla f\|_{L_p(D;\mathbf{R}^d)} \le C\,\|B(t)\nabla f\|_{L_p(D;\mathbf{R}^d)}$$

gives the first inequality. □

Lemma 2.3 *Let $\theta \in C^1([0,\tau]; C_c^1(D,\mathbf{R}^d))$ be vector field and Φ its flow. The functions $t \mapsto A(t) := \xi(t)\partial\Phi_t^{-1}\partial\Phi_t^{-T}$, $t \mapsto \xi(t) := \det(\partial\Phi_t)$ and $t \mapsto B(t) := \partial\Phi_t^{-T}$ are differentiable on $[0,\tau]$ and satisfy the following ordinary differential equations*

$$\begin{aligned}B'(t) &= -B(t)(\partial\theta^t)^T B(t)\\ \xi'(t) &= tr(\partial\theta^t B^T(t))\xi(t)\\ A'(t) &= tr(\partial\theta^t B^T(t))A(t) - B^T(t)\partial\theta^t A(t) - (B^T(t)\partial\theta^t A(t))^T,\end{aligned}$$

where $\theta^t(x) := \theta(t, \Phi_t(x))$ and $' := \frac{d}{dt}$.

Proof (i) Let E, F be two Banach spaces. In [2, p. 222. Satz 7.2] it is proved that

$$\mathrm{inv} : \mathcal{L}\mathrm{is}(E;F) \to \mathcal{L}(F;E), \quad A \mapsto A^{-1}$$

is infinitely continuously differentiable with derivative $\partial\mathrm{inv}(A)(B) = -A^{-1}BA^{-1}$. Now by the fundamental theorem of calculus, we have

$$\Phi_t(x) = x + \int_0^t \theta(s, \Phi_s(x))\, ds \;\Rightarrow\; \partial\Phi_t(x) = I + \int_0^t \partial\theta(s, \Phi_s(x)))\, ds,$$

where $I \in \mathbf{R}^{d,d}$ denotes the identity matrix. Therefore $t \mapsto \partial\Phi_t(x)$ is differentiable for each $x \in \overline{D}$ with derivative

$$\frac{d}{dt}(\partial\Phi_t(x)) = \partial\theta^t(x) = \partial\theta(t, \Phi_t(x))\partial\Phi_t(x).$$

Thus if we let $E = F = \mathbf{R}^{d,d}$ and take into account the previous equation, we get by the chain rule

$$\frac{d}{dt}(\mathrm{inv}(\partial\Phi_t(x))) = -(\partial\Phi_t(x))^{-1}\partial\theta^t(x)(\partial\Phi_t(x))^{-1}.$$

(ii) A proof may be found in [22, p. 215, Proposition 10.6].
(iii) Follows from the product rule together with (i) and (ii). □

Remark 2.4 Note that equation (i) can also be proved by differentiating the identity $\partial\Phi_t\partial\Phi_t^{-1} = I$, where I is the identity matrix in $\mathbf{R}^d$. That the inverse $t \mapsto \partial\Phi_t^{-1}$ is differentiable can also be seen by the well known formula $\partial\Phi_t^{-1} = (\det(\partial\Phi_t))^{-1}(\mathrm{cofac}(\partial\Phi_t))^T$, where cofac denotes the cofactor matrix.

Lemma 2.5 *Let $D \subset \mathbf{R}^d$ be an open, bounded set and $p \geq 1$ a real number. Denote by Φ_t the flow of $\theta \in C_c^1(D, \mathbf{R}^d)$.*

(i) For any $f \in L_p(D)$, we have

$$\lim_{t \searrow 0} \|f \circ \Phi_t - f\|_{L_p(D)} = 0 \quad \textit{and} \quad \lim_{t \searrow 0} \|f \circ \Phi_t^{-1} - f\|_{L_p(D)} = 0.$$

(ii) For any $f \in W_p^1(D)$, we have

$$\lim_{t \searrow 0} \|f \circ \Phi_t - f\|_{W_p^1(D)} = 0. \tag{2.4}$$

(iii) For $p \geq 1$ a real number, $k \in \{1, 2\}$ and any $f \in W_p^k(D)$, we have

$$\lim_{t \searrow 0} \left\| \frac{f \circ \Phi_t - f}{t} - \nabla f \cdot \theta \right\|_{W_p^{k-1}(D)} = 0.$$

(iv) Fix $p \geq 1$ and let $t \to u_t : [0, \tau] \to W_p^1(D)$ be a continuous function in 0. Set $u := u_0$. Then $t \mapsto u_t \circ \Phi_t : [0, \tau] \to W_p^1(D)$ is continuous in 0 and

$$\lim_{t \searrow 0} \|u_t \circ \Phi_t - u\|_{W_p^1(D)} = 0.$$

Proof (i) A proof can be found in [10, p. 529].
(ii) In order to prove (2.4) it is sufficient to show

$$\lim_{t\searrow 0} \|\nabla(f\circ\Phi_t - f)\|_{L_p(D)} = \lim_{t\searrow 0} \|\partial\Phi_t^T((\nabla f)\circ\Phi_t - \nabla f)\|_{L_p(D)} = 0.$$

By the triangle inequality, we have

$$\|\partial\Phi_t^T((\nabla f)\circ\Phi_t - \nabla f)\|_{L_p(D)} \le \|(\nabla f)\circ\Phi_t - \nabla f)\|_{L_p(D)} + \|(\partial\Phi_t^T - I)\nabla f)\|_{L_p(D)}.$$

For the first term on the right hand side we can use (i) and the second term tends to zero since $\partial\Phi_t^T \to I$ in $C(\overline{D}; \mathbf{R}^{d,d})$.
(iii) A proof can be found in [14, p. 6, Lemma 3.6].
(iv) By the triangle inequality, we get

$$\|u_t\circ\Phi_t - u\|_{W_p^1(D)} \le \|u_t\circ\Phi_t - u\circ\Phi_t\|_{W_p^1(D)} + \|u\circ\Phi_t - u\|_{W_p^1(D)}.$$

The last term on the right hand side converges to zero as $t \to 0$ due to (ii). For the second inequality note that

$$\begin{aligned}\|u_t\circ\Phi_t - u\circ\Phi_t\|_{W_p^1(D)} &= \left(\int_D \xi^{-1}(t)|u_t - u|^p + \xi^{-1}(t)|B(t)\nabla(u_t - u)|^p\right)^{1/p}\\ &\le C\left(\int_D |u_t - u|^p + |\nabla(u_t - u)|^p\right)^{1/p}\end{aligned}$$

and the right hand side converges to zero. □

Definition 2.6 (*Eulerian semi-derivative*) Let $\Omega \subset D$ and $k \ge 1$ be given. Suppose we are given a shape function $J : \Xi(\Omega) \to \mathbf{R}$ on the set $\Xi(\Omega) := \cup_{t\in[0,\tau]}\{\Phi_t(\Omega)\,|\,\theta \in \mathcal{C}_c^k(D, \mathbf{R}^d)\}$. Then the **Eulerian semi-derivative** of J at Ω in the direction θ is defined as the limit (if it exists)

$$dJ(\Omega)[\theta] := \lim_{t\searrow 0} \frac{J(\Omega_t) - J(\Omega)}{t}.$$

Moreover, if the Eulerian semi-derivative $dJ(\Omega)[\theta]$ exists for all $\theta \in \mathcal{C}_c^\infty(D, \mathbf{R}^d)$ and the map $\theta \mapsto dJ(\Omega)[\theta] : \mathcal{C}_c^\infty(D, \mathbf{R}^d) \to \mathbf{R}$, is linear and continuous, then J is called **shape differentiable** at Ω.

Finally, we state the following theorem from [10, pp. 483–484], which will be used to compute the boundary expression of the shape derivative.

Theorem 2.7 *Let $\theta \in \mathcal{C}_c^k(D, \mathbf{R}^d)$, where $k \ge 1$. Fix $\tau > 0$ and let $\varphi \in C(0,\tau; W_{loc}^{1,1}(\mathbf{R}^d)) \cap C^1(0,\tau; L_{loc}^1(\mathbf{R}^d))$ and an bounded domain Ω with Lipschitz boundary Γ be given. The right sided derivative of the function $f(t) := \int_{\Omega_t}\varphi(t)\,dx$ at $t = 0$ is given by*

$$f'(0^+) = \int_\Omega \varphi'(0)\, dx + \int_\Gamma \varphi(0)\, \theta_n\, ds.$$

In the following, we prove the shape differentiability of J defined in (2.2) by the following methods: the material and shape derivative method, the min-max formulation of Correa-Seeger and the rearrangement method. We present a modification of Céa's Lagrange method which allows a rigorous derivation of the shape derivative in the case of existence of material derivatives.

3 Material and Shape Derivative Method

3.1 Material Derivative Method

In order to compute the Eulerian semi-derivative of J given by (2.2) via material derivative method (chain rule approach), we make the following assumption:

Assumption $(\mathcal{A})$ The function $\varrho : \mathbf{R} \to \mathbf{R}$ is continuously differentiable, bounded and monotonically increasing.

We call $u \in H_0^1(\Omega)$ a **weak solution** of (2.1) if

$$\int_\Omega \nabla u \cdot \nabla \psi\, dx + \int_\Omega \varrho(u)\, \psi\, dx = \int_\Omega f \psi\, dx \quad \text{for all } \psi \in H_0^1(\Omega). \tag{3.1}$$

The weak solution of the previous equation characterizes the unique minimum of the energy $E(\Omega, \cdot) : H_0^1(\Omega) \to \mathbf{R}$ defined by

$$E(\Omega, \varphi) := \frac{1}{2} \int_\Omega |\nabla \varphi|^2 + \hat{\varrho}(\varphi)\, dx - \int_\Omega f \varphi\, dx,$$

where $\hat{\varrho}(s) := \int_0^s 2\, \varrho(s')\, ds'$. In the following, we denote by

$$d_\varphi E(\Omega, \varphi; \psi) := \lim_{t \searrow 0} \frac{E(\Omega, \varphi + t\, \psi) - E(\Omega, \varphi)}{t}$$

$$d_\varphi^2 E(\Omega, \varphi; \psi, \tilde{\psi}) := \lim_{t \searrow 0} \frac{d_\varphi E(\Omega, \varphi + t\, \tilde{\psi}; \psi) - d_\varphi E(\Omega, \varphi; \psi)}{t}$$

the first and second order directional derivative of E at φ in the direction ψ and $(\psi, \tilde{\psi})$, respectively. Then we may write (3.1) as $d_\varphi E(\Omega, u; \psi) = 0$ for all $\psi \in H_0^1(\Omega)$.

Lemma 3.1 *Assume that ϱ is continuously differentiable. Then the mapping*

$$s \mapsto \int_\Omega \varrho(\varphi + s\tilde{\varphi})\psi \, dx$$

is continuously differentiable on $\mathbf{R}$ *for all* $\varphi, \tilde{\varphi} \in L_\infty(\Omega)$ *and* $\psi \in H_0^1(\Omega)$.

Proof Let $\varphi, \tilde{\varphi} \in H_0^1(\Omega) \cap L_\infty(\Omega)$ and $\psi \in H_0^1(\Omega)$. Put $z^s(x) := \varrho(\varphi(x) + s\tilde{\varphi}(x))\psi(x)$. We have for almost all $x \in \Omega$

$$\frac{z^{s+h}(x) - z^s(x)}{h} \to = \varrho'(\varphi(x) + s\tilde{\varphi}(x))\tilde{\varphi}(x)\psi(x) \quad \text{as } h \to 0,$$
$$\left|\frac{d}{ds} z^s(x)\right| \le C|\psi(x)||\tilde{\varphi}(x)|.$$

Then it holds

$$\begin{aligned}\left|\frac{z^{s+h}(x) - z^s(x)}{h}\right| &= \left|\frac{1}{h}\int_s^{s+h} \frac{d}{ds'} z^{s'}(x) ds'\right| \\ &\le C|\psi(x)||\tilde{\varphi}(x)|\frac{1}{h}\int_{s'}^{s'+h} ds' \\ &= C|\psi(x)||\tilde{\varphi}(x)|.\end{aligned}$$

Therefore applying Lebesgue's dominated convergence theorem we conclude

$$\frac{d}{ds}\int_\Omega z^s(x)\, dx = \int_\Omega \varrho'(\varphi(x) + s\tilde{\varphi}(x))\tilde{\varphi}(x)\psi(x)\, dx.$$

As a consequence of the previous lemma, we get the differentiability of $s \mapsto d_\varphi E(\Omega, \varphi + s\tilde{\varphi}, \psi)$. Moreover, we conclude by the monotonicity of ϱ

$$d_\varphi^2 E(\Omega, \varphi; \psi, \psi) = \int_\Omega |\nabla\psi|^2 + \varrho'(\varphi)\psi^2 \, dx \ge C\|\psi\|^2_{H_0^1(\Omega)}$$

for all $\varphi \in H_0^1(\Omega) \cap L_\infty(\Omega)$ and $\psi \in H_0^1(\Omega)$. We now want to calculate the shape derivative of (2.2). For this purpose, we consider the perturbed cost function $J(\Omega_t) = \int_{\Omega_t} |u_t - u_r|^2 \, dx$, where u_t denotes the weak solution of (3.1) on the domain $\Omega_t := \Phi_t(\Omega)$, that is, $u_t \in H_0^1(\Omega_t)$ solves

$$\int_{\Omega_t} \nabla u_t \cdot \nabla\hat{\psi}\, dx + \int_{\Omega_t} \varrho(u_t)\hat{\psi}\, dx = \int_{\Omega_t} f\hat{\psi}\, dx \quad \text{for all } \hat{\psi} \in H_0^1(\Omega_t). \tag{3.2}$$

It would be possible to compute the derivative of $u_t : \Omega_t \to \mathbf{R}$ pointwise by

$$du(x) := \lim_{t \searrow 0} \frac{u_t(x) - u(x)}{t} \quad \text{for all } x \in \left(\bigcap_{t \in [0,\tau]} \Omega_t \right) \cap \Omega.$$

In the literature this derivative is referred to as *local shape derivative of* u in direction θ; cf. [12]. Nevertheless, we go another way and use the change of variables $\Phi_t(x) = y$ to rewrite $J(\Omega_t)$ as

$$J(\Omega_t) = \int_\Omega \xi(t) |u^t - u_r \circ \Phi_t|^2 \, dx, \tag{3.3}$$

where $u^t := \Psi_t(u_t) : \Omega \to \mathbf{R}$ is a function on the fixed domain Ω. We introduce the mapping $\Psi_t(\varphi) := \varphi \circ \Phi_t$ with inverse $\Psi^t(\hat{\varphi}) := \Psi_t^{-1}(\hat{\varphi}) = \hat{\varphi} \circ \Psi_t^{-1}$. To study the differentiability of (3.3), we can study the function $t \mapsto u^t$. Notice that $u_0 = u^0 = u$ is nothing but the weak solution of (3.1).

The limit $\dot{u} := \lim_{t \searrow 0} (u^t - u)/t$ is called **strong material derivative** if we consider this limit in the norm convergence in $H_0^1(\Omega)$ and **weak material derivative** if we consider the weak convergence in $H_0^1(\Omega)$.

The crucial observation of [23, Theorem 2.2.2, p. 52] is that Ψ_t constitutes an isomorphism from $H^1(\Omega_t)$ into $H^1(\Omega)$. Hence using a change of variables in (3.2) shows that u^t satisfies

$$\int_\Omega A(t) \nabla u^t \cdot \nabla \psi \, dx + \int_\Omega \xi(t) \, \varrho(u^t) \psi \, dx = \int_\Omega \xi(t) f^t \psi \, dx \quad \text{for all } \psi \in H_0^1(\Omega), \tag{3.4}$$

where we used the notation from (2.3). The previous equation characterizes the unique minimum of the convex energy $\tilde{E} : [0, \tau] \times H_0^1(\Omega) \to \mathbf{R}$[2]

$$\tilde{E}(t, \varphi) := \frac{1}{2} \int_\Omega \xi(t) |B(t) \nabla \varphi|^2 + \xi(t) \hat{\varrho}(\varphi) \, dx - \int_\Omega \xi(t) f^t \varphi \, dx. \tag{3.5}$$

By standard regularity theory (see e.g. [15]) it follows that $u^t \in C(\overline{\Omega})$ for all $t \in [0, \tau]$. Moreover, the proof of [4, Theorem 3.1] shows that there is a constant $C > 0$ such that

$$\|u^t\|_{C(\overline{\Omega})} + \|u^t\|_{H^1(\Omega)} \le C \quad \text{for all } t \in [0, \tau].$$

As before using Lebesque's dominated convergence theorem it is easy to verify that for fixed $t \in [0, \tau]$ the second order directional derivative $d_\varphi^2 \tilde{E}(t, \varphi; \psi, \eta)$ exists for all $\varphi \in L_\infty(\Omega) \cap H_0^1(\Omega)$ and $\psi, \eta \in H_0^1(\Omega)$. Taking into account Proposition 2.1, we see that

$$C \|\psi\|^2_{H^1(\Omega; \mathbf{R}^d)} \le d_\varphi^2 \tilde{E}(t, \varphi; \psi, \psi). \tag{3.6}$$

[2] Here we mean convex with respect to φ for each $t \in [0, \tau]$.

or all *varphi* $\in L_\infty(\Omega) \cap H_0^1(\Omega), \psi \in H_0^1(\Omega)$ and for all $t \in [0,\tau]$, Note that $d_\varphi \tilde{E}(t,\varphi;\psi)$ is also differentiable with respect to t and Lemma 2.3 shows:

$$\begin{aligned}\partial_t d_\varphi \tilde{E}(t,\varphi;\psi) &= \int_\Omega A'(t)\nabla\varphi\cdot\nabla\psi + \xi'(t)\varrho(\varphi)\,\psi\,dx \\ &\quad - \int_\Omega (\xi'(t)f^t + \xi(t)B(t)\nabla f^t)\varphi\,dx \\ &\le C(1+\|\varphi\|_{H^1(\Omega)})\|\psi\|_{H^1(\Omega)},\end{aligned} \tag{3.7}$$

for all $t \in [0,\tau]$, where $C > 0$ is a constant. By the coercivity property (3.6) of the second order derivative of $\tilde{E}$

$$C\|\nabla(u^t-u)\|^2_{L_2(\Omega;\mathbf{R}^d)} \le \int_0^1 d_\varphi^2 \tilde{E}(t, su^t+(1-s)u; u^t-u, u^t-u) \tag{3.8}$$

$$= d_\varphi \tilde{E}(t,u^t;u^t-u) - d_\varphi \tilde{E}(t,u;u^t-u) \tag{3.9}$$

$$= -(d_\varphi \tilde{E}(t,u;u^t-u) - d_\varphi \tilde{E}(0,u;u^t-u)) \tag{3.10}$$

$$= -t\,\partial_t d_\varphi \tilde{E}(\eta_t\, t,u;u^t-u) \tag{3.11}$$

$$\le Ct\,\|\nabla(u^t-u)\|_{L_2(\Omega;\mathbf{R}^d)}. \tag{3.12}$$

In step (3.8)–(3.9), we applied the mean value theorem in integral form, in step (3.9)–(3.10), we used that $d_\varphi \tilde{E}(t,u^t;u^t-u) = d_\varphi \tilde{E}(0,u;u^t-u) = 0$, and in step from (3.10)–(3.11), we applied the mean value theorem which yields $\eta_t \in (0,1)$. In the last step (3.12), we employed the estimate (3.7). Finally, by the Poincaré inequality, we conclude that there is $c > 0$ such that $\|u^t-u\|_{H^1(\Omega)} \le ct$ for all $t \in [0,\tau]$. From this estimate we deduce that for any real sequence $(t_n)_{n\in\mathbf{N}}$ with $t_n \searrow 0$ as $n \to \infty$, the quotient $w^n := (u^{t_n}-u)/t_n$ converges weakly in $H_0^1(\Omega)$ to some element $\dot{u}$ and by compactness there is a subsequence $(t_{n_k})_{k\in\mathbf{N}}$ such that $(w^{n_k})_{k\in\mathbf{N}}$ converges strongly in $L_q(\Omega)$ to some v, where $0 < q < \frac{2d}{d-2}$; (cf. [11, p. 270, Theorem 6]).[3] Extracting a further subsequence we may assume that $w^{t_k}(x) \to \dot{u}(x)$ as $k \to \infty$ for almost every $x \in \Omega$. Notice that the limit $\dot{u}$ depends on the sequence $(t_{n_k})_{k\in\mathbf{N}}$. However, we will see that this limit is the same for any sequence $(t_n)_{n\in\mathbf{N}}$ converging to zero.

Subtracting (3.4) at $t > 0$ and $t = 0$ yields

$$\begin{aligned}&\int_\Omega A(t)\nabla(u^t-u)\cdot\nabla\psi\,dx + \int_\Omega \xi(t)(\varrho(u^t)-\varrho(u))\,\psi\,dx \\ &= \int_\Omega (\xi(t)-1)\,\varrho(u)\,\psi\,dx - \int_\Omega (A(t)-I)\nabla u\cdot\nabla\psi\,dx \\ &\quad + \int_\Omega (\xi(t)-1)f^t\psi\,dx + \int_\Omega (f^t-f)\psi\,dx.\end{aligned} \tag{3.13}$$

[3] When $d = 2$ this means $H^1(\Omega)$ is compactly embedded into $L_p(\Omega)$ for arbitrary $p > 1$. When $d = 3$ we get that $H^1(\Omega)$ compactly embeds into $L_{6-\varepsilon}(\Omega)$ for any small $\varepsilon > 0$.

We choose $t = t_{n_k}$ in the previous equation and want to pass to the limit $k \to \infty$. The only difficult term in (3.13) is

$$\int_\Omega \xi(t) \frac{\varrho(u^t) - \varrho(u)}{t} \psi \, dx = \int_\Omega \xi(t) \left[\int_0^1 \varrho'(u_s^t) \, ds \right] \left(\frac{u^t - u}{t} \right) \psi \, dx.$$

From the strong convergence of $(u^{t_{n_k}} - u)/t_{n_k}$ to $\dot{u}$ in $L_2(\Omega)$ and the pointwise convergence $\xi(t_{n_k}) \to 1$ and $\varrho'(u_s^{t_{n_k}}) \to \varrho'(u)$, we infer that

$$\int_\Omega \xi(t_{n_k}) \frac{\varrho(u^{t_{n_k}}) - \varrho(u)}{t_{n_k}} \psi \, dx \longrightarrow \int_\Omega \varrho'(u) \, \dot{u} \, \psi \, dx \quad \text{as } k \to \infty.$$

Therefore, choosing $t = t_{n_k}$ in (3.13) and dividing by t_{n_k}, we may pass to the limit:

$$\begin{aligned} &\int_\Omega \nabla \dot{u} \cdot \nabla \psi + \varrho'(u) \, \dot{u} \, \psi \, dx + \int_\Omega A'(0) \nabla u \cdot \nabla \psi \, dx \\ &\quad + \int_\Omega \operatorname{div} \theta \varrho(u) \psi \, dx = \int_\Omega \operatorname{div}(\theta) f \psi \, dx + \int_\Omega \nabla f \cdot \theta \, \psi \, dx. \end{aligned} \tag{3.14}$$

for all $\psi \in H_0^1(\Omega)$. The function $\dot{u}$ is the unique solution of (3.14). Hence for every sequence $(t_n)_{n \in \mathbf{N}}$ converging to zero there exists a subsequence $(t_{n_k})_{k \in \mathbf{N}}$ such that $w^{t_k} \to \dot{u}$ as $k \to \infty$. Moreover,

$$\int_\Omega \xi(t) \frac{\varrho(u^t) - \varrho(u)}{t} \psi \, dx \longrightarrow \int_\Omega \varrho'(u) \, \dot{u} \, \psi \, dx \quad \text{as } t \searrow 0$$

and

$$\int_\Omega A(t) \nabla \frac{u^t - u}{t} \cdot \nabla \psi \, dx \longrightarrow \int_\Omega \nabla \dot{u} \cdot \nabla \psi \, dx \quad \text{as } t \searrow 0.$$

We now show that the strong material derivative exists. For this subtract (3.14) from (3.13) to obtain

$$\begin{aligned} &\int_\Omega A(t) \nabla \left(\frac{u^t - u}{t} - \dot{u} \right) \cdot \nabla \psi \, dx + \int_\Omega \xi(t) \left[\int_0^1 \varrho'(u_s^t) \, ds \right] \left(\frac{u^t - u}{t} - \dot{u} \right) \psi \, dx \\ &= \int_\Omega (A(t) - I) \nabla \dot{u} \cdot \nabla \psi \, dx + \int_\Omega (\xi(t) - 1) \left[\int_0^1 \varrho'(u_s^t) \, ds \right] \dot{u} \, \psi \, dx \\ &\quad + \int_\Omega \left[\int_0^1 \varrho'(u_s^t) - \varrho'(u) \, ds \right] \dot{u} \, \psi \, dx - \int_\Omega \left(\frac{A(t) - I}{t} - A'(0) \right) \nabla u \cdot \nabla \psi \, dx \\ &\quad + \int_\Omega \left(\frac{\xi(t) - 1}{t} - \operatorname{div}(\theta) \right) \varrho(u) \, \psi \, dx + \int_\Omega \left(\frac{\xi(t) - 1}{t} - \operatorname{div}(\theta) \right) f^t \psi \, dx \\ &\quad + \int_\Omega \left(\frac{f^t - f}{t} - \nabla f \cdot \theta \right) \psi \, dx. \end{aligned}$$

Now we insert $\psi = w^t - \dot{u}$ as test function into the previous equation. Using Proposition 2.1 and the fact that $\xi(t) > 0$, $\varrho' \geq 0$ we get

$$\begin{aligned}\gamma_1 \|\nabla(w^t - \dot{u})\|^2_{L_2(\Omega)} \leq & \int_\Omega (A(t) - I)\nabla\dot{u} \cdot \nabla(w^t - \dot{u})\, dx \\ & + \int_\Omega (\xi(t) - 1) \int_0^1 \varrho'(u_s^t)\, ds\, \dot{u}\, (w^t - \dot{u})\, dx \\ & + \int_\Omega \int_0^1 (\varrho'(u_s^t) - \varrho'(u)\, ds)\, \dot{u}\, (w^t - \dot{u})\, dx \\ & - \int_\Omega \left(\frac{A(t) - I}{t} - A'(0) \right) \nabla u \cdot \nabla(w^t - \dot{u})\, dx \\ & + \int_\Omega \left(\frac{\xi(t) - 1}{t} - \operatorname{div}(\theta) \right) (\varrho(u)\, (w^t - \dot{u}) + f^t\, (w^t - \dot{u}))\, dx \\ & + \int_\Omega \left(\frac{f^t - f}{t} - \nabla f \cdot \theta \right) (w^t - \dot{u})\, dx.\end{aligned}$$

Using the convergences $A(t) \to I, (A(t) - I)/t - A'(0) \to 0, (f^t - f)/t - \nabla f \cdot \theta \to 0$, $\xi(t) \to 1$ and $(\xi(t) - 1)/t - \operatorname{div}(\theta)$ in $C(\overline{\Omega})$, and the uniform boundedness of $\|w^t - \dot{u}\|_{H^1(\Omega)}$ and $\|\dot{u}\|_{H^1(\Omega)}$ yields

$$\|w^t - \dot{u}\|_{H^1(\Omega)} \to 0 \quad \text{as } t \searrow 0.$$

We are now in the position to calculate the volume expression of the shape derivative. First, we differentiate (3.3) with respect to t

$$dJ(\Omega)[\theta] = \int_\Omega \operatorname{div}(\theta)|u - u_r|^2\, dx - \int_\Omega 2(u - u_r)\nabla u_r \cdot \theta\, dx + \int_\Omega 2(u - u_r)\dot{u}\, dx.$$

Note that for the previous calculation it was enough to have $\|u^t - u\|_{H^1(\Omega)} \leq ct$ for all $t \in [0, \tau]$. This is sufficient to differentiate the L_2 cost function. Nevertheless, for a cost function that involves gradients of u such as

$$\tilde{J}(\Omega) := \int_\Omega \|\nabla u - \nabla u_r\|^2\, dx,$$

this is not true anymore. Now in order to eliminate the material derivative in the last equation, the so-called adjoint equation is introduced

$$\text{Find } p \in H_0^1(\Omega): \quad d_\varphi E(\Omega, u; p, \psi) = -2 \int_\Omega (u - u_r)\, \psi\, dx \quad \text{for all } \psi \in H_0^1(\Omega). \tag{3.15}$$

The function p is called *adjoint state*. Finally, testing the adjoint equation with $\dot{u}$ and the material derivative Eq. (3.14) with p, we arrive at the volume expression

$$
\begin{aligned}
dJ(\Omega)[\theta] &\overset{(3.15)}{=} \int_\Omega \operatorname{div}(\theta)|u-u_r|^2\,dx \\
&\quad - \int_\Omega 2(u-u_r)\nabla u_r\cdot\theta\,dx - d_\varphi E(\Omega,u;p,\dot u) \\
&\overset{(3.14)}{=} \int_\Omega \operatorname{div}(\theta)|u-u_r|^2\,dx - \int_\Omega 2(u-u_r)\nabla u_r\cdot\theta\,dx \\
&\quad + \int_\Omega A'(0)\nabla u\cdot\nabla p + \operatorname{div}(\theta)\varrho(u)p\,dx - \int_\Omega \operatorname{div}(\theta f)p\,dx. \quad (3.16)
\end{aligned}
$$

Note that the volume expression already makes sense when $u,\ p\in H_0^1(\Omega)$. Assuming higher regularity of the state and adjoint (e.g. $u,\ p\in H^2(\Omega)\cap H_0^1(\Omega)$) would allow us to rewrite the previous volume expression into a boundary expression, that is, an integral over the boundary $\partial\Omega$.

3.2 Shape Derivative Method

Assuming that the solutions u, p and the boundary $\partial\Omega$ are smooth, say C^2, we may transform the volume expression (3.16) into an integral over $\partial\Omega$. This can be accomplished by integration by parts or in the following way. Instead of transporting the cost function back to Ω, one may directly differentiate $J(\Omega_t)=\int_{\Omega_t}|\Psi^t(u^t)-u_r|^2\,dx$ by invoking Theorem 2.7, to obtain

$$
dJ(\Omega)[\theta]=\int_{\partial\Omega}|u-u_r|^2\theta_n ds+\int_\Omega 2\,(u-u_r)(\dot u-\partial_\theta u)\,dx. \quad (3.17)
$$

The function $u':=\dot u-\partial_\theta u$ is called **shape derivative** of u at Ω in direction θ associated with the parametrization Ψ_t. It is linear with respect to θ. Note that since $\Psi^0=id$, we have $\Psi^t\circ\Psi^{-t}=\Psi^0=id_{H_0^1(\Omega)}$ and $\Psi^{-t}\circ\Psi^t=\Psi^0=id_{H_0^1(\Omega_t)}$. Note that setting $u^t:=\Psi_t(u_t)$, we can write

$$
u'=\frac{d}{dt}\Psi^t(u^t)|_{t=0}=\frac{d}{dt}(u^t\circ\Phi_t^{-1})|_{t=0}.
$$

Therefore the shape derivative decomposes into two parts, namely

$$
u'=\underbrace{\partial_t\Psi^t(u^t)_{|t=0}}_{\in L_2(\Omega)}+\underbrace{\Psi^0(\dot u)}_{\in H_0^1(\Omega)},
$$

where $\partial_t\Psi^t(u^t)|_{t=0}:=\lim_{t\searrow 0}(\Psi^t(u^t)-\Psi^0(u^t))/t=-\partial_\theta u$. Assuming that the solution u belongs to $u\in H_0^1(\Omega)\cap H^2(\Omega)$, we have

$$u' = \underbrace{\partial_t \Psi^t(u^t)_{|t=0}}_{\in H^1(\Omega)} + \underbrace{\Psi^0(\dot{u})}_{\in H_0^1(\Omega)\cap H^2(\Omega)}$$

The perturbed state equation (3.2) can be rewritten as

$$\int_{\Omega_t} \nabla(\Psi^t(u^t)) \cdot \nabla(\Psi^t(\varphi)) \; + \; \varrho(\Psi^t(u^t))\,(\Psi^t(\varphi))\,dx = \int_{\Omega_t} f\,\Psi^t(\varphi)\,dx$$

for all $\varphi \in H_0^1(\Omega)$. Suppose that $u, p \in H^2(\Omega) \cap H_0^1(\Omega)$. Hence by formally differentiating the last equation using the transport Theorem 2.7:

$$\begin{aligned} &\int_\Omega \nabla u' \cdot \nabla\varphi + \varrho'(u)u'\,\varphi\,dx - \int_\Omega \nabla u \cdot \partial_\theta\varphi + \varrho(u)\,\partial_\theta\varphi\,dx \\ &+ \int_{\partial\Omega} (\nabla u \cdot \nabla\varphi + \varrho(u)\,p)\;\theta_n ds = \int_{\partial\Omega} f\varphi\;\theta_n ds - \int_\Omega f\partial_\theta\varphi\,dx \end{aligned} \tag{3.18}$$

for all $\varphi \in H^2(\Omega) \cap H_0^1(\Omega)$, where $\theta_n := \theta \cdot n$ and $\partial_\theta := \theta \cdot \nabla$. Note that the adjoint state p vanishes on Γ. This equation can also be derived from (3.14) by partial integration.

Remark 3.2 Note that u' does not belong to $H_0^1(\Omega)$, but only to $H^1(\Omega)$. As the shape derivative does not belong to the solution space of the state equation, it may lead to false or incomplete formulas for the boundary expression. This seems to be first observed in [17].

Note that $u = 0$ on Γ implies that $\nabla_\Gamma u = 0$ and hence $\nabla u|_\Gamma = (\partial_n u)n$. Then integrating by parts in (3.18) and using that u is a strong solution yields

$$\begin{aligned} \int_\Omega \nabla\dot{u} \cdot \nabla\varphi \; + \; \varrho'(u)\dot{u}\,\varphi\,dx = &\int_{\partial\Omega} (\partial_n u\;\partial_n\varphi - 2\,\partial_n u\;\partial_n\varphi)\;\theta_n\;ds \\ &+ \int_\Omega \partial_\theta u\,(-\Delta\varphi + \varrho'(u)\,\varphi)\,dx. \end{aligned} \tag{3.19}$$

Now, one can eliminate $\dot{u}$ in $dJ(\Omega)[\theta]$ given by (3.17) using the previous equation and the adjoint state equation

$$\begin{aligned} dJ(\Omega)[\theta] &\overset{(3.15)}{=} \int_{\partial\Omega} |u - u_r|^2\;\theta_n\,ds + \int_\Omega \nabla\dot{u} \cdot \nabla p + \varrho'(u)\,\dot{u}\,p\,dx \\ &\quad + \int_\Omega \partial_\theta u\,2(u - u_r)\,dx \\ &\overset{(3.19)}{=} \int_{\partial\Omega} |u - u_r|^2\;\theta_n\,ds - \int_{\partial\Omega} 2\partial_n u\;\partial_n p\,\theta_n ds \\ &\quad + \int_\Omega (-\Delta p + \varrho'(u)\;p + 2(u - u_r))\partial_\theta u\,dx. \end{aligned}$$

Finally, assuming that p solves the adjoint equation in the strong sense, we get

$$dJ(\Omega)[\theta] = \int_{\partial\Omega}(|u - u_r|^2 - \partial_n u\, \partial_n p)\, \theta_n\, ds. \tag{3.20}$$

What we observe in the calculations above is that there is no material derivative $\dot{u}$ or shape derivative u' in the final expression (3.16) or (3.20). This suggests that there might be a way to obtain this formula without the computation of $\dot{u}$. In the next section, we get to know one possible way to avoid the material derivatives.

4 The Min-Max Formulation of Correa and Seeger

In this section, we want to discuss the minimax formulation of shape optimization problems and a theorem of Correa and Seeger [6] that gives a powerful tool to differentiate a minimax function with respect to a parameter. The cost function for many optimal control problems can be rewritten as the min-max of a Lagrangian function $\mathcal{L}$, that is, an utility function plus the equality constraints, i.e.,

$$J(u) = \inf_{\varphi\in A} \sup_{\psi\in B} \mathcal{L}(u, \varphi, \psi).$$

Therefore, the directional differentiation of the cost function is equivalent to the differentiation of the inf-sup with respect to u. This method has clear restrictions, but still it is applicable to many commonly used cost functions and to a certain class of non-linear partial differential equations. This method is in particular applicable to linear partial differential equations and convex cost functions.

4.1 Saddle Points and Their Characterization

For the convenience of the reader we recall here the definition of saddle points and their characterization.

Definition 4.1 Let A, B be sets and $G : A \times B \to \mathbf{R}$ a map. Then a pair $(u, p) \in A \times B$ is said to be a **saddle point** on $A \times B$ if

$$G(u, \psi) \le G(u, p) \le G(\varphi, p) \quad \text{for all } \varphi \in A, \quad \text{for all } \psi \in B.$$

We have the following equivalent condition for (u, p) being a saddle point.

Lemma 4.2 *A pair* $(u, p) \in A \times B$ *is a saddle point of* $G(,)$ *if and only if*[4]

$$\min_{\hat{\varphi}\in A} \sup_{\hat{\psi}\in B} G(\hat{\varphi}, \hat{\psi}) = \max_{\hat{\psi}\in B} \inf_{\hat{\varphi}\in A} G(\hat{\varphi}, \hat{\psi}),$$

and it is equal to $G(u, p)$*, where* u *being the attained minimum and* p *the attained maximum, respectively.*

Proof A proof can be found in [20, p. 166–167].

4.2 Min-Max Formulation for the Semi-linear Equation

Let $\varphi, \psi \in H_0^1(\Omega)$ be two functions. Instead of differentiating the cost function and the state equation separately, we incorporate both in the Lagrangian

$$\mathcal{L}(\Omega, \varphi, \psi) := \int_\Omega |\varphi - u_r|^2 \, dx + \int_\Omega \nabla\varphi \cdot \nabla\psi \, dx + \int_\Omega \varrho(\varphi)\psi \, dx - \int_\Omega f\psi \, dx.$$

The point of departure for the **min-max formulation** is the observation that

$$J(\Omega) = \min_{\varphi\in H_0^1(\Omega)} \sup_{\psi\in H_0^1(\Omega)} \mathcal{L}(\Omega, \varphi, \psi),$$

since for any $\varphi \in H_0^1(\Omega)$

$$\sup_{\psi\in H_0^1(\Omega)} \mathcal{L}(\Omega, \varphi, \psi) = \begin{cases} J(\Omega) & \text{when } \varphi = u \text{ solves (3.1)} \\ +\infty & \text{else .} \end{cases}$$

In order to apply the **theorem of Correa-Seeger** to the Lagrangian $\mathcal{L}$, we have to show that it admits saddle points. Reasonable conditions to ensure the existence of saddle points for our specific example is to assume that $\mathcal{L}$ is convex and differentiable with respect to φ.

Assumption $(\mathcal{C})$ The function ϱ is linear, that is, $\varrho(x) = ax$, where $a \in \mathbf{R}$.

Since for every open set $\Omega \subset \mathbf{R}^d$ the Lagrangian $\mathcal{L}$ is convex and differentiable with respect to φ, and concave and differentiable with respect to ψ, we know from [20, Proposition 1.6, p. 169–170] that the saddle points can be characterized by

$$\begin{aligned} u \in H_0^1(\Omega): &\quad \partial_\psi \mathcal{L}(\Omega, u, p)(\hat{\psi}) = 0 \quad \text{for all } \hat{\psi} \in H_0^1(\Omega) \\ p \in H_0^1(\Omega): &\quad \partial_\varphi \mathcal{L}(\Omega, u, p)(\hat{\varphi}) = 0 \quad \text{for all } \hat{\varphi} \in H_0^1(\Omega). \end{aligned}$$

[4] Here the min and max indicate that the infimum and supremum is attained, respectively.

The last equations are exactly the state equation (3.1) and the adjoint Eq. (3.15). To compute the shape derivative of J, we consider for $t > 0$

$$\begin{aligned} J(\Omega_t) &= \min_{\hat{\varphi} \in H_0^1(\Omega_t)} \sup_{\hat{\psi} \in H_0^1(\Omega_t)} \mathcal{L}(\Omega_t, \hat{\varphi}, \hat{\psi}) \\ &= \min_{\varphi \in H_0^1(\Omega)} \sup_{\psi \in H_0^1(\Omega)} \mathcal{L}(\Omega_t, \Psi^t(\varphi), \Psi^t(\psi)), \end{aligned} \tag{4.1}$$

where the saddle points of $\mathcal{L}(\Omega_t, \cdot, \cdot)$ are again given by the solutions of (3.1) and (3.15), but the domain Ω has to be replaced by Ω_t. By definition of a saddle point

$$\mathcal{L}(\Omega_t, u_t, \hat{\psi}) \le \mathcal{L}(\Omega_t, u_t, p_t) \le \mathcal{L}(\Omega_t, \hat{\varphi}, p_t) \quad \text{for all } \hat{\psi}, \hat{\varphi} \in H_0^1(\Omega_t). \tag{4.2}$$

Since $\Psi_t : H_0^1(\Omega_t) \to H_0^1(\Omega)$ is a bijection it is easily seen that the saddle points of $G(t, \varphi, \psi) := \mathcal{L}(\Omega_t, \Psi^t(\varphi), \Psi^t(\psi))$ are given by $u^t = \Psi_t(u_t)$ and $p^t = \Psi_t(p_t)$. It can also be verified that the function u^t solves (3.4) and applying the change of variables $\Phi_t(x) = y$ to (3.15) shows that p^t solves

$$\int_\Omega A(t) \nabla \psi \cdot \nabla p^t + \xi(t)\, \varrho'(u^t)\, p^t\, \psi \, dx = -2 \int_\Omega \xi(t)(u^t - u_r^t)\psi \, dx \tag{4.3}$$

for all $\psi \in H_0^1(\Omega)$. Moreover, the functions u^t, p^t satisfy

$$G(t, u^t, \psi) \le G(t, u^t, p^t) \le G(t, \varphi, p^t) \quad \text{for all } \psi, \varphi \in H_0^1(\Omega),$$

where G takes, after applying the change of variables $\Phi_t(x) = y$, the explicit form

$$G(t, \varphi, \psi) = \int_\Omega \xi(t) |\varphi - u_r^t|^2 \, dx + \int_\Omega A(t) \nabla \varphi \cdot \nabla \psi + \xi(t) \varrho(\varphi)\, \psi \, dx - \int_\Omega \xi(t) f^t \psi \, dx. \tag{4.4}$$

From Lemma 4.2 and the definition of a saddle point (u^t, p^t) of $G(t, \,)$, we conclude

$$g(t) := \min_{\varphi \in H_0^1(\Omega)} \sup_{\psi \in H_0^1(\Omega)} G(t, \varphi, \psi) = G(t, u^t, p^t). \tag{4.5}$$

Moreover, we have the relation

$$g(t) = G(t, u^t, \psi) \quad \text{for all } \psi \in H_0^1(\Omega), \tag{4.6}$$

since u^t solves (3.4). In view of (4.1), we can obtain the shape derivative $dJ(\Omega)[\theta]$ by calculating the derivative of $g(t)$ at $t = 0$. In order to use (4.5), we have to find conditions which show that we are allowed to differentiate the min-max of the function G with respect to t at $t = 0$. On the other hand the relation (4.6) shows that $dJ(\Omega)[\theta] = \frac{d}{dt} G(t, u^t, \psi)_{|_{t=0}}$ for all $\psi \in H_0^1(\Omega)$, that means the differentiability of the min-max of G is equivalent to the differentiability of $G(t, u^t, \psi)$ and it is

independent of ψ. Sufficient conditions for the differentiability are provided by the Theorem of Correa-Seeger (Theorem 4.5). Note the relation (4.5) is also true for a general function G when u^t, p^t are saddle points, but the relation (4.6) only for the special structure (4.4) of G. It is clear that if the functions u^t and G are sufficiently differentiable the derivative $\frac{d}{dt}(g(t))_{t=0}$ exists. The purpose of the reformulation of the cost function as an inf-sup is to avoid the material derivatives $\dot{u}$. Note that when the state equation has no unique solution the cost function is not well-defined, but the the function g is. Without a computation of the material derivative $\dot{u}$ or $\dot{p}$, we can show (cf. also the Theorem 4.5) that $dJ(\Omega)[\theta] = \partial_t G(0, u, p)$. Clearly the functions $t \mapsto u^t$, $t \mapsto p^t$ and G have to satisfy some additional conditions. Let us sketch the proof of this fundamental result when G is given by (4.4). To be more precise we want to establish the following.

Proposition 4.3 *Let $\psi \in H_0^1(\Omega)$. Then the function $[0, \tau] \to \mathbf{R} : t \mapsto G(t, u^t, \psi)$ is differentiable from the right side in 0. Moreover, we have the following*

$$\frac{d}{dt} G(t, u^t, \psi)|_{t=0} = \partial_t G(0, u, p) \tag{4.7}$$

for arbitrary $\psi \in H_0^1(\Omega)$. Here, $p \in H_0^1(\Omega)$ solves the adjoint Eq. (3.15).

Proof By definition of a saddle point (u^t, p^t)

$$G(t, u^t, p^t) \leq G(t, u, p^t), \quad G(0, u, p) \leq G(0, u^t, p)$$

and therefore setting $\Delta(t) := G(t, u^t, p^t) - G(0, u, p)$ gives

$$G(t, u^t, p) - G(0, u^t, p) \leq \Delta(t) \leq G(t, u, p^t) - G(0, u, p^t).$$

Using the mean value theorem, we find for each $t \in [0, \tau]$ numbers $\zeta_t, \eta_t \in (0, 1)$ such that

$$t\partial_t G(t\zeta_t, u^t, p) \leq \Delta(t) \leq t\partial_t G(t\eta_t, u, p^t), \tag{4.8}$$

where the derivative of G with respect to t is given by

$$\begin{aligned} \partial_t G(t, \varphi, \psi) = &\int_\Omega \xi'(t)|\varphi - u_r^t|^2 - 2\xi(t)(\varphi - u_r^t)B(t)\nabla u_r^t \cdot \theta^t \, dx \\ &+ \int_\Omega A'(t)\nabla\varphi \cdot \nabla\psi + \xi'(t)\varrho(\varphi)\,\psi - \xi'(t)\, f^t\psi - B(t)\nabla f^t \cdot \theta^t \psi \, dx \end{aligned} \tag{4.9}$$

and the derivatives ξ' and A' are given by Lemma 2.3. It can be verified from this formula that $(t, \varphi) \mapsto \partial_t G(t, \varphi, p)$ is strongly continuous and $(t, \psi) \mapsto \partial_t G(t, u, \psi)$ is even weakly continuous. Moreover, from (3.4) and (4.3) it can be inferred that $t \mapsto u^t$ and $t \mapsto p^t$ are bounded in $H_0^1(\Omega)$ and therefore for any sequence $(t_n)_{n\in\mathbf{N}}$ we get $u^{t_n} \rightharpoonup w$, $p^{t_n} \rightharpoonup v$ for two elements $w, v \in H_0^1(\Omega)$. Passing to the limit in

(3.4) and (4.3) and taking into account Lemma 2.5, we see that w solves the state equation and v the adjoint equation. By uniqueness of the state and adjoint equation we get $w = u$ and $v = p$. Selecting a further subsequence $(t_{n_k})_{k\in\mathbf{N}}$ yields that $u^{t_{n_k}}$ converges strongly in $L_2(\Omega)$. Thus we conclude from (4.8)

$$\liminf_{t\searrow 0} \Delta(t)/t \geq \partial_t G(0, u, p), \quad \limsup_{t\searrow 0} \Delta(t)/t \leq \partial_t G(0, u, p),$$

which leads to $\limsup_{t\searrow 0} \Delta(t)/t = \liminf_{t\searrow 0} \Delta(t)/t$. This finishes the proof of (4.7) and hence we have shown the shape differentiability of J.

Evaluating the derivative $\partial_t G(t, u, p)|_{t=0}$ leads to the formula (3.16). Note that when $\partial\Omega$ is C^2 then we may extend u, $p \in H^2(\Omega) \cap H_0^1(\Omega)$ to global H^2 functions $\tilde{u}$, $\tilde{p} \in H^2(\mathbf{R}^d)$. Then the boundary expression is obtained by applying the transport theorem (Theorem 2.7) to $\frac{d}{dt}\mathcal{L}(\Omega_t, \Psi^t(\tilde{u}), \Psi^t(\tilde{p}))|_{t=0}$:

$$\begin{aligned} dJ(\Omega)[\theta] = &\int_\Gamma (|u - u_r|^2 + \nabla u \cdot \nabla p + \varrho(u)\, p)\theta_n \, ds + \int_\Omega \nabla \mathring{u} \cdot \nabla p + \varrho'(u)\, \mathring{u}\, p \, dx \\ &+ \int_\Omega (u - u_r)\mathring{u} \, dx + \int_\Omega \nabla u \cdot \nabla \mathring{p} + \varrho(u)\, \mathring{p} \, dx - \int_\Omega f\, \mathring{p} \, dx, \end{aligned}$$

where $\mathring{u} = \partial_t(\Psi^t(\tilde{u}))|_{t=0} = -\nabla u \cdot \theta$, $\mathring{p} = \partial_t(\Psi^t(\tilde{p}))|_{t=0} = -\nabla p \cdot \theta$. To rewrite the equation into an integral over Γ, we integrate by parts and obtain

$$\begin{aligned} dJ(\Omega)[\theta] = &\int_\Gamma (|u - u_r|^2 + \nabla u \cdot \nabla p + \varrho(u)\, p)\, \theta_n \, ds \\ &+ \int_{\partial\Omega} \mathring{u}\, \partial_n p \, ds + \int_{\partial\Omega} \partial_n u\, \mathring{p} \, ds \\ &- \int_\Omega \mathring{u} \left(-\Delta p + \varrho'(u)\, p + 2(u - u_r)\right) dx \\ &- \int_\Omega \mathring{p}\, (-\Delta u + \varrho(u) - f)\, dx. \end{aligned}$$

Finally, using the strong solvability of u and p, and taking into account $\nabla u = (\partial_n u)n$ on $\partial\Omega$, we arrive at (3.20).

Remark 4.4 We point out that the inequalities (4.2) are the key to avoid the material derivatives. Nevertheless, without the assumption of convexity of G with respect to φ it is difficult to prove this inequality.

4.3 The Theorem of Correa-Seeger

Finally, we quote the Theorem of Correa-Seeger, which applies in situations when the state equation admits no unique solution and the Lagrangian admits saddle points.

The proof is similar to the proof of Proposition 4.3. Let a real number $\tau > 0$ and vector spaces E and F be given. We consider a mapping

$$G : [0, \tau] \times E \times F \to \mathbf{R}.$$

For each $t \in [0, \tau]$ we define

$$g(t) := \inf_{u \in E} \sup_{p \in F} G(t, u, p), \qquad h(t) := \sup_{p \in F} \inf_{u \in E} G(t, u, p)$$

and the associated sets

$$E(t) = \left\{ \hat{\varphi} \in E : \sup_{p \in F} G(t, \hat{\varphi}, p) = g(t) \right\}$$
$$F(t) = \left\{ \hat{\psi} \in F : \inf_{u \in E} G(t, u, \hat{\psi}) = h(t) \right\}.$$

For fixed t they are the points in E respectively F where inf respectively the sup are attained in $g(t)$ respectively $h(t)$. We know that if $g(t) = h(t)$ then the set of saddle points is given by

$$S(t) := E(t) \times F(t).$$

Theorem 4.5 (R. Correa and A. Seeger, [9]) *Let the function G and the vector spaces E, F be as before. Suppose the following conditions:*

(HH1) *For all $t \in [0, \tau]$ assume $S(t) \neq \emptyset$.*

(HH2) *The partial derivative $\partial_t G(t, u, p)$ exists for all $(t, u, p) \in [0, \tau] \times E \times F$.*

(HH3) *For any sequence $(t_n)_{n \in \mathbf{N}}$ with $t_n \searrow 0$ there exists a subsequence $(t_{n_k})_{k \in \mathbf{N}}$ and an element $u_0 \in E(0)$, $u_{t_{n_k}} \in E(t_{n_k})$ such that for all $p \in F(0)$*

$$\lim_{\substack{k \to \infty \\ t \searrow 0}} \partial_t G(t, u_{n_k}, p) = \partial_t G(0, u_0, p).$$

(HH4) *For any sequence $(t_n)_{n \in \mathbf{N}}$ with $t_n \searrow 0$ there exists a subsequence $(t_{n_k})_{k \in \mathbf{N}}$ and an element $p_0 \in F(0)$, $p_{t_{n_k}} \in F(t_{n_k})$ such that for all $u \in E(0)$*

$$\lim_{\substack{k \to \infty \\ t \searrow 0}} \partial_t G(t, u, p_{t_{n_k}}) = \partial_t G(0, u, p_0).$$

Then there exists $(u_0, p_0) \in E(0) \times F(0)$ such that

$$\frac{d}{dt} g(t)|_{t=0} = \partial_t G(0, u_0, p_0).$$

4.4 Céa's Classical Lagrange Method and a Modification

Let the function G be defined by (4.4). Assume that G is sufficiently differentiable with respect to t, φ and ψ. Additionally, assume that the strong material derivative $\dot{u}$ exists in $H_0^1(\Omega)$. Then we may calculate as follows

$$dJ(\Omega)[\theta] = \frac{d}{dt}(G(t, u^t, p))|_{t=0} = \underbrace{\partial_t G(t, u, p)|_{t=0}}_{\text{shape derivative}} + \underbrace{\partial_\varphi G(0, u, p)(\dot{u})}_{\text{adjoint equation}},$$

and due to $\dot{u} \in H_0^1(\Omega)$ it implies

$$dJ(\Omega)[\theta] = \partial_t G(t, u, p)|_{t=0}.$$

Therefore, we can follow the lines of the calculation of the previous section to obtain the boundary and volume expression of the shape derivative.

In the original work [5], it was calculated as follows

$$dJ(\Omega)[\theta] = \partial_\Omega \mathcal{L}(\Omega, u, p) + \partial_\varphi \mathcal{L}(\Omega, u, p)(u') + \partial_\psi \mathcal{L}(\Omega, u, p)(p'), \qquad (4.10)$$

where $\partial_\Omega \mathcal{L}(\Omega, u, p) := \lim_{t \searrow 0}(\mathcal{L}(\Omega_t, u, p) - \mathcal{L}(\Omega, u, p))/t$. Then it was assumed that u' and p' belong to $H_0^1(\Omega)$, which has as consequence that $\partial_\varphi \mathcal{L}(\Omega, u, p)(u') = \partial_\psi \mathcal{L}(\Omega, u, p)(p') = 0$. Thus (4.10) leads to the wrong formula

$$dJ(\Omega)[\theta] = \int_\Gamma (|u - u_r|^2 + \partial_n u\, \partial_n p)\, \theta_n \, ds.$$

This can be fixed by noting that $u' = \dot{u} - \partial_\theta u$ and $p' = \dot{p} - \partial_\theta p$ with $\dot{u}, \dot{p} \in H_0^1(\Omega)$:

$$dJ(\Omega)[\theta] = \partial_\Omega \mathcal{L}(\Omega, u, p) - \partial_\varphi \mathcal{L}(\Omega, u, p)(\partial_\theta u) - \partial_\psi \mathcal{L}(\Omega, u, p)(\partial_\theta p),$$

which gives the correct formula. Finally, note that for Maxwell's equations a different parametrization than $v \mapsto v \circ \Phi_t$ of the function space is necessary since the differential operator is modified differently. This leads then to a different definition of the shape derivative and also the formulas will be different. This is well-known from the finite element analysis of Maxwell's equations; cf. [1, 3, 13, 16].

5 Rearrangement of the Cost Function

The rearrangement method introduced in [14] avoids the material derivative and is applicable to a wide class of elliptic problems. We describe the method at hand of our semi-linear example and write subsequently the perturbed cost function (3.3) as

$$J(\Omega_t) = \int_\Omega j(t, u^t)\,dx, \quad j(t, v) := \xi(t)|v - u_r^t|^2. \tag{5.1}$$

In order to derive the shape differentiability, we make the following assumptions:

Assumption $(\mathcal{R})$ Assume that $\varrho \in C^2(\mathbf{R}) \cap L_\infty(\mathbf{R})$, $\varrho'' \in L_\infty(\mathbf{R})$ and $\varrho'(x) \geq 0$ for all $x \in \mathbf{R}$.

Instead of requiring the Lipschitz continuity of $t \mapsto u^t$, we claim that the following holds: there exist constants $c, \tau, \varepsilon > 0$ such that $\|u^t - u\|_{H_0^1(\Omega)} \leq ct^{1/2+\varepsilon}$ for all $t \in [0, \tau]$.

Theorem 5.1 *Let Assumption $(\mathcal{R})$ be satisfied and let $\theta \in C_c^2(D, \mathbf{R}^d)$. Then $J(\Omega_t)$ given by (5.1) is differentiable with derivative:*

$$dJ(\Omega)[\theta] = \partial_t G(0, u, p),$$

where u, p are solutions of the state and adjoint state equation.

Proof The main idea is to rewrite the difference $J(\Omega_t) - J(\Omega)$ and use a first order expansions of the PDE and the cost function with respect to the unknown together with Hölder continuity of $t \mapsto u^t$. To be more precise, write

$$\begin{aligned}\frac{J(\Omega_t) - J(\Omega)}{t} = &\underbrace{\frac{1}{t}\int_\Omega (j(t, u^t) - j(t, u) - j'(t, u)(u^t - u))\,dx}_{B_1(t)} \\ &+ \underbrace{\frac{1}{t}\int_\Omega (j(t, u) - j(0, u))\,dx}_{B_2(t)} \\ &+ \underbrace{\frac{1}{t}\int_\Omega (j'(t, u) - j'(0, u))(u^t - u)\,dx}_{B_3(t)} \\ &+ \underbrace{\frac{1}{t}\int_\Omega j'(0, u)(u^t - u)\,dx}_{B_4(t)},\end{aligned} \tag{5.2}$$

where $j' := \partial_u j$ and $u_s^t := su^t + (1-s)u$. Using the mean value theorem in integral form entails for some constant $C > 0$

$$\begin{aligned}\int_\Omega (j(t, u^t) - j(t, u) - j'(t, u)(u^t - u))\,dx &= \int_0^1 (1-s)\, j''(t, u_s^t)(u^t - u)^2\,dx \\ &\leq C\|u^t - u\|_{L_2(\Omega)}^2 \quad \text{for all } t \in [0, \tau].\end{aligned}$$

Using the $\lim_{t\searrow 0} \|u^t - u\|_{H_0^1(\Omega)}/\sqrt{t} = 0$, we see that B_1 tends to zero as $t \searrow 0$. Let $\tilde{E}(t, \varphi)$ be defined by (3.5). Then the fourth term in (5.2) can be written by using the adjoint Eq. (3.15) as follows

$$\begin{aligned}\int_\Omega j'(0,u)(u^t - u)\,dx &= d_\varphi \tilde{E}(0, u^t; p) - d_\varphi \tilde{E}(0, u; p) - d_\varphi^2 \tilde{E}(0, u; u^t - u, p)\\ &\quad + d_\varphi \bar{E}(t, u^t; p) - d_\varphi \bar{E}(t, u; p)\\ &\quad - (d_\varphi \tilde{E}(0, u^t; p) - d_\varphi \tilde{E}(0, u; p))\\ &\quad + d_\varphi \tilde{E}(t, u; p) - d_\varphi \tilde{E}(0, u; p). \qquad (5.3)\end{aligned}$$

By standard elliptic regularity theory, we may assume that $p \in H_0^1(\Omega) \cap L_\infty(\Omega)$. Therefore by virtue of Taylor's formula in Banach spaces (cf. [2, p. 193, Theorem 5.8]) the first line in (5.3) on the right hand side can be written as

$$\begin{aligned}&d_\varphi \tilde{E}(0, u^t; p) - d_\varphi \tilde{E}(0, u; p) - d_\varphi^2 \tilde{E}(0, u; u^t - u, p)\\ &= \int_0^1 (1-s) d^3 \tilde{E}(0, u_s^t; u^t - u, u^t - u, p)\,ds,\end{aligned}$$

where the remainder can be estimated as follows

$$\begin{aligned}\int_0^1 (1-s) d_\varphi^3 \tilde{E}(0, u_s^t; u^t - u, u^t - u, p)\,ds &= \int_0^1 (1-s)\varrho''(u_s^t)(u^t - u)^2\, p\,ds\\ &\le \frac{1}{2}\|p\|_{L_\infty(\Omega)} \|\varrho''\|_{L_\infty(\mathbf{R})} \|u^t - u\|_{L_2(\Omega)}.\end{aligned}$$

Using $d_\varphi \tilde{E}(t, u^t; p) - d_\varphi \tilde{E}(0, u; p) = 0$, and the differentiability of $t \mapsto \tilde{E}(t, u)$ yields

$$\begin{aligned}\lim_{t\searrow 0} \frac{d_\varphi \tilde{E}(t, u^t; p) - d_\varphi \tilde{E}(t, u; p)}{t} &= \lim_{t\searrow 0} \frac{1}{t}(d_\varphi \tilde{E}(0, u^t; p) - d_\varphi \tilde{E}(0, u; p)),\\ \lim_{t\searrow 0} \frac{d_\varphi \tilde{E}(t, u; p) - d_\varphi \tilde{E}(0, u; p)}{t} &= \int_\Omega A'(0)\nabla u \cdot \nabla p - \operatorname{div}(\theta) fp - \nabla f \cdot \theta\, p\,dx.\end{aligned}$$

Thus from (5.3), we infer

$$\begin{aligned}&\lim_{t\searrow 0} \frac{1}{t} \int_\Omega j'(0,u)(u^t - u)\,dx\\ &= \int_\Omega A'(0)\nabla u \cdot \nabla p + \operatorname{div}(\theta)\varrho(u)p - \operatorname{div}(\theta) fp - \nabla f \cdot \theta\, p\,dx.\end{aligned}$$

Therefore we may pass to the limit in (5.2) and obtain

$$\lim_{t \searrow 0} \frac{J(\Omega_t) - J(\Omega)}{t} = \int_\Omega \partial_t j(0,u)\, dx + \partial_t d_\varphi \tilde{E}(0,u;p).$$

This finishes the proof and shows that $dJ(\Omega)[\theta] = \partial_t G(0,u,p)$. □

6 Differentiability of Energy Functionals

If it happens that the cost function J is the energy of the PDE (2.1), that is,

$$J(\Omega) := \min_{\varphi \in H_0^1(\Omega)} E(\Omega, \varphi),$$

then it is easy to show the shape differentiability of J by using a result from [10, p. 524, Theorem 2.1], see also [7, pp. 139]. First note that $J(\Omega_t) = \min_{\varphi \in H_0^1(\Omega)} \tilde{E}(t,\varphi)$. By definition of the minimum u^t of $\tilde{E}(t,\cdot)$ and u of $\tilde{E}(0,\cdot)$, respectively, we have

$$\tilde{E}(0,u^t) - \tilde{E}(0,u) \geq 0, \qquad \tilde{E}(t,u) - \tilde{E}(0,u) \leq 0$$

and thus

$$\begin{aligned} J(\Omega_t) - J(\Omega) &= \tilde{E}(t,u^t) - \tilde{E}(0,u^t) + \tilde{E}(0,u^t) - \tilde{E}(0,u) \\ &\geq \tilde{E}(t,u^t) - \tilde{E}(0,u^t) \\ J(\Omega_t) - J(\Omega) &= \tilde{E}(t,u^t) - \tilde{E}(t,u) + \tilde{E}(t,u) - \tilde{E}(0,u) \\ &\leq \tilde{E}(t,u) - \tilde{E}(0,u). \end{aligned}$$

Using the mean value theorem, we conclude the existence of numbers $\eta_t, \zeta_t \in (0,1)$ such that

$$t\, \partial_t \tilde{E}(\eta_t\, t, u^t) \leq J(\Omega_t) - J(\Omega) \leq t\, \partial_t \tilde{E}(\zeta_t\, t, u).$$

Thus if

$$\tilde{E}(0,u) \leq \liminf_{t \searrow 0} \partial_t \tilde{E}(\eta_t\, t, u^t), \qquad \tilde{E}(0,u) \geq \limsup_{t \searrow 0} \partial_t \tilde{E}(\zeta_t\, t, u), \tag{6.1}$$

then we may conclude that J is shape differentiable by the squeezing lemma. We obtain

$$\lim_{t \searrow 0} \frac{J(\Omega_t) - J(\Omega)}{t} = \partial_t \tilde{E}(0,u).$$

This result can be seen as a special case of Theorem 4.5. Note that in our example

$$\partial_t \tilde{E}(t,\varphi) = \int_\Omega A'(t)\nabla\varphi\cdot\nabla\varphi + \xi'(t)\varrho(\varphi)\,dx - \int_\Omega \xi'(t) f^t \varphi\,dx + \int_\Omega \xi(t)\, B(t)\nabla f^t\cdot\varphi\,dx.$$

From this identity, the convergence of $u^t \to u$ in $H_0^1(\Omega)$ and the smoothness of $A(t)$, $\xi(t)$ and $B(t)$, we infer that (6.1) are verified.

7 The Averaged Adjoint Approach

Let the Banach spaces E, F and a number $\tau > 0$ be given. Consider a function

$$G : [0,\tau]\times E\times F \to \mathbf{R}, \quad (t,\varphi,\psi)\mapsto G(t,\varphi,\psi)$$

such that $\psi \mapsto G(t,\varphi,\psi)$ is affine for all $(t,\varphi)\in[0,\tau]\times E$. Introduce the *solution set of the state equation*

$$E(t) := \{u\in E |\ d_\psi G(t,u,0;\hat{\psi}) = 0 \ \text{ for all } \hat{\psi}\in F\}.$$

Introduce the following hypothesis.

Assumption Suppose that $E(t) = \{u^t\}$ is single-valued for all $[0,\tau]$.

(i) For all $t\in[0,\tau]$ and $\tilde{p}\in F$ the mapping

$$[0,1]\to\mathbf{R} :\ s\mapsto G(t, su^t + (1-s)u^0, \tilde{p})$$

is absolutely continuous. This implies that for almost all $s\in[0,1]$ the derivative $d_\varphi G(t, su^t + (1-s)u^0, \tilde{p}; u^t - u^0)$ exists and in particular

$$G(t,u^t,\tilde{p}) - G(t,u^0,\tilde{p}) = \int_0^1 d_\varphi G(t, su^t + (1-s)u^0, \tilde{p}; u^t - u^0)\,ds.$$

(ii) For all $t\in[0,\tau]$, $\varphi\in E$ and $\tilde{p}\in F$

$$s\mapsto d_\varphi G(t, su^t + (1-s)u^0, \tilde{p};\varphi)$$

is well-defined and belongs to $L_1(0,1)$.

Introduce for $t\in[0,\tau]$, $u^t\in E(t)$ and $u^0\in E(0)$ the following set

$$Y(t,u^t,u^0) := \left\{ q\in F |\ \forall\hat{\varphi}\in E :\ \int_0^1 d_\varphi G(t, su^t + (1-s)u^0, q;\hat{\varphi})\,ds = 0\right\},$$

which is called solution set of the *averaged adjoint equation* with respect to t, u^t and u^0. For $t = 0$ the set $Y(0, u^0) := Y(0, u^0, u^0)$ coincides with the solution set of the usual adjoint state equation

$$Y(0, u^0) = \left\{ q \in F \mid d_\varphi G(0, u^0, q; \hat{\varphi}) = 0 \quad \text{for all } \hat{\varphi} \in E \right\}.$$

We call any $p \in Y(0, u^0)$ an adjoint state.

Theorem 7.1 *Let linear vector spaces E and F, a real number $\tau > 0$. Suppose that the function*

$$G : [0, \tau] \times E \times F \to \mathbf{R}, \quad (t, \varphi, \psi) \mapsto G(t, \varphi, \psi),$$

is affine in the last argument. Let Assumption (H0) and the following conditions be satisfied.

(H1) For all $t \in [0, \tau]$ and all $(u, p) \in E(0) \times F$ the derivative $\partial_t G(t, u, p)$ exists.
(H2) For all $t \in [0, \tau]$ the set $Y(t, u^t, u^0)$ is nonempty and $Y(0, u^0, u^0)$ is single-valued.
(H3) Let $p^0 \in Y(0, u^0)$. For any sequence $(t_n)_{n \in \mathbf{N}}$ of non-negative real numbers converging to zero, there exist a subsequence $(t_{n_k})_{k \in \mathbf{N}}$ and $p^{t_{n_k}} \in Y(t_{n_k}, u^{t_{n_k}}, u^0)$ such that

$$\lim_{\substack{k \to \infty \\ s \searrow 0}} \partial_t G(s, u^0, p^{t_{n_k}}) = \partial_t G(0, u^0, p^0).$$

Then for any $\psi \in F$:

$$\frac{d}{dt}(G(t, u^t, \psi))|_{t=0} = \partial_t G(0, u^0, p^0).$$

Proof The result was proved in [19].

7.1 Application to the Semi-linear Problem

In this section, we apply Theorem 7.1 to the example (2.1) and (2.2). For convenience, we recall the cost function

$$J(\Omega) = \int_\Omega |u - u_r|^2 \, dx, \tag{7.1}$$

and the weak formulation of (2.1)

$$\int_\Omega \nabla u \cdot \nabla \psi \, dx + \int_\Omega \varrho(u) \, \psi \, dx = \int_\Omega f \psi \, dx \quad \text{for all } \psi \in H_0^1(\Omega). \tag{7.2}$$

Suppose in the following the assumption on the data f, u_r and Ω introduced in the beginning of Sect. 3 is satisfied. Recall that the equation (7.2) on the domain $\Phi_t(\Omega)$ transported back to Ω by $y = \Phi_t(x)$ reads

$$\int_\Omega A(t)\nabla u^t \cdot \nabla \psi \, dx + \int_\Omega \xi(t)\varrho(u^t)\psi \, dx = \int_\Omega \xi(t) f^t \psi \, dx, \quad \text{for all } \psi \in H_0^1(\Omega). \tag{7.3}$$

This equation characterizes the unique minimum of the convex energy (3.5). Recall the definition of the Lagrangian associated to the problem

$$G(t, \varphi, \psi) = \int_\Omega \xi(t)|\varphi - u_r^t|^2 \, dx + \int_\Omega A(t)\nabla\varphi \cdot \nabla\psi + \xi(t)\varrho(\varphi)\,\psi \, dx - \int_\Omega \xi(t) f^t \psi \, dx. \tag{7.4}$$

Theorem 7.2 *Let Assumption* $(\mathcal{A})$ *be satisfied. Then* J *defined in* (7.1) *is shape differentiable and its derivative is given by*

$$dJ(\Omega)[\theta] = \partial_t G(0, u^0, p^0),$$

where $p^0 \in Y(0, u^0)$.

Proof Let us verify the conditions (H0)–(H3) for the function G given by (7.4).

(H0) This has already been proven in Sect. 3.

(H1) This is an easy consequence of $\theta \in C_c^2(D, \mathbf{R}^d)$ and Lemma 2.5. The derivative is given by (4.9).

(H2) Note that for all $t \in [0, \tau]$, we have $\in E(t) = \{u^t\}$, where u^t solves (7.3). We have $p^t \in Y(t, u^t, u^0)$ if and only if

$$\int_\Omega A(t)\nabla\psi \cdot \nabla p^t + \xi(t)k(u, u^t)\,\psi \, dx = -\int_\Omega \xi(t)(u^t + u - 2u_r^t)\psi \, dx, \tag{7.5}$$

for all $\psi \in H_0^1(\Omega)$, where $k(u, u^t) := \int_0^1 \varrho'(u_s^t)\, ds$ and $u_s^t := su^t + (1-s)u$. Due to the Lemma of Lax-Milgram the previous equation has a unique solution $p^t \in H_0^1(\Omega)$. Note that the strong formulation of the averaged adjoint on the moved domain, namely $p_t := p^t \circ \Phi_t^{-1}$ on Ω_t satisfies

$$\begin{aligned} -\Delta p_t + k(u \circ \Phi_t^{-1}, u_t)\, p_t &= -(u_t - u \circ \Phi_t^{-1} - 2u_r) \quad \text{in } \Omega_t \\ p_t &= 0 \quad \text{on } \partial\Omega_t, \end{aligned}$$

where $k(u \circ \Phi_t^{-1}, u^t \circ \Phi_t^{-1}) := \int_0^1 \varrho'(u_s^t \circ \Phi_t^{-1})\, ds = \int_0^1 \varrho'(su_t + (1-s)u \circ \Phi_t^{-1})\, ds$.

(H3) We already know that Assumption $(\mathcal{A})$ implies that $t \mapsto u^t$ is continuous from $[0, \tau]$ into $H_0^1(\Omega)$. But this is actually not necessary as we will show. Suppose

that we do not know that $t \mapsto u^t$ is continuous. Then by inserting $\psi = u^t$ in the state equation (7.3), we obtain after an application of Hölder's inequality $\|u^t\|_{H^1(\Omega)} \leq C$ for some constant $C > 0$. For any sequence of non-negative real numbers $(t_n)_{n\in\mathbf{N}}$ converging to zero there exists a subsequence $(t_{n_k})_{n\in\mathbf{N}}$ such that $u^{t_{n_k}} \rightharpoonup z$ as $k \to \infty$. Setting $t = t_{n_k}$ in the state equation and passing to the limit $k \to \infty$ shows $z = u$. Moreover, inserting $\psi = p^t$ into (7.5) as test function and using Hölder's inequality yields for some constant $C > 0$

$$\|p^t\|_{H_0^1(\Omega)} \leq C\|u^t + u - 2u_r^t\|_{L_2(\Omega)} \quad \text{for all } t \in [0, \tau].$$

Therefore again for any sequence $(t_n)_{n\in\mathbf{N}}$ there exists a subsequence $(t_{n_k})_{n\in\mathbf{N}}$ such that $y^{t_{n_k}} \rightharpoonup q \quad$ as $k \to \infty$ for some $q \in H_0^1(\Omega)$. Selecting $t = t_{n_k}$ in (7.5), we want to pass to the limit $k \to \infty$ by using Lebesque's dominated convergence theorem. It suffices to show that $w^k(x) := \int_0^1 \varrho'(u_s^{t_{n_k}}(x))\, ds$ is bounded in $L_\infty(\mathbf{R}^d)$ independently of k and that this sequence convergences pointwise almost everywhere in Ω to $\varrho'(u)$. The boundedness of w^k follows from the continuity of u^t on $\overline{\Omega}$ and the continuity of ϱ. The pointwise convergence $w^k(x) \to \varrho(u(x))$ as $k \to \infty$ (possibly a subsequence) follows from the fact that ϱ is continuous and $u^{t_{n_k}}$ converges pointwise to u as $k \to \infty$. Therefore there is a sequence $t_n \searrow 0$ such that we may pass to the limit $n \to \infty$ in (7.5), after inserting $t = t_n$. By uniqueness, we conclude $q = p \in Y(0, u^0)$. Finally note that $(t, \psi) \mapsto \partial_t G(t, u, \psi)$ is weakly continuous.

All conditions (H0)–(H3) are satisfied and we finish the proof. □

References

1. V. Akelik, G. Biros, O. Ghattas, D. Keyes, K. Ko, L.-Q. Lee, E.G. Ng, Adjoint methods for electromagnetic shape optimization of the low-loss cavity for the international linear collider. J. Phys.: Conf. Ser. **16**(1), 435 (2005)
2. H. Amann, J. Escher, *Analysis II*. Grundstudium Mathematik. [Basic Study of Mathematics] (Birkhauser Verlag, Basel, 1999)
3. J. Cagnol, M. Eller, Shape optimization for the maxwell equations under weaker regularity of the data. Comptes Rendus Mathematique **348**(2122), 1225–1230 (2010)
4. E. Casas, Boundary control of semilinear elliptic equations with pointwise state constraints. SIAM J. Control Optim. **31**(4), 993–1006 (1993)
5. J. Cèa, Conception optimale ou identification de formes, calcul rapide de la derivee dircetionelle de la fonction cout. Math. Mod. Numer. Anal. **20**, 371–402 (1986)
6. R. Correa, A. Seeger, Directional derivative of a minimax function. Nonlinear Anal. **9**(1), 13–22 (1985)
7. M.C. Delfour, *Introduction to Optimization and Semidifferential Calculus*. MOS-SIAM Series on Optimization (Society for Industrial and Applied Mathematics, Philadelphia, 2012)
8. M.C. Delfour, J.-P. Zolèsio, Shape sensitivity analysis via a penalization method. Ann. Mat. Pura Appl. **4**(151), 179–212 (1988)
9. M.C. Delfour, J.-P. Zolèsio, Shape sensitivity analysis via min max differentiability. SIAM J. Control Optim. **26**(4), 834–862 (1988)

10. M.C. Delfour, J.-P. Zolèsio, Shapes and geometries. *Metrics, Analysis, Differential Calculus, and Optimization*. Advances in Design and Control, vol. 22, 2nd edn. (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2011)
11. L. Evans, *Partial Differential Equations* (American Mathematical Society, Providence, 2002)
12. H. Harbrecht, Analytical and numerical methods in shape optimization. Math. Methods Appl. Sci. **31**(18), 2095–2114 (2008)
13. F. Hettlich, The domain derivative of time-harmonic electromagnetic waves at interfaces. Math. Methods Appl. Sci. **35**(14), 1681–1689 (2012)
14. K. Ito, K. Kunisch, G.H. Peichl, Variational approach to shape derivatives. ESAIM Control Optim. Calc. Var. **14**(3), 517–539 (2008)
15. D. Kinderlehrer, G. Stampacchia, An introduction to variational inequalities and their applications. *Classics in Applied Mathematics*, vol. 31 (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2000). Reprint of the 1980 original
16. P. Monk, Finite element methods for Maxwell's equations. *Numerical Mathematics and Scientifc Computation* (Oxford University Press, New York, 2003)
17. O. Pantz, Sensibilitè de l'èquation de la chaleur aux sauts de conductivitè. C. R. Math. Acad. Sci. Paris **341**(5), 333–337 (2005)
18. J. Sokołowski, J.-P. Zolèsio, Introduction to shape optimization. *Shape Sensitivity Analysis*. Springer Series in Computational Mathematics, vol. 16 (Springer, Berlin, 1992)
19. K. Sturm, Lagrange method in shape optimization for non-linear partial differential equations: a material derivative free approach. WIAS- Preprint No. 1817 (Submitted) (2013)
20. R. Temam, I. Ekeland, Analyse convexe et problemes variationnels. Springer, New York (1974)
21. F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen* (Vieweg, Wiesbaden, 2005)
22. L. Younes, *Shapes and Diffeomorphisms*. Applied Mathematical Sciences, vol. 171 (Springer, Berlin, 2010)
23. W.P. Ziemer, Weakly differentiable functions. *Sobolev Spaces and Functions of Bounded Variation*. Graduate Texts in Mathematics, vol. 120 (Springer, New York, 1989)

Optimal Regularity Results Related to a Partition Problem Involving the Half-Laplacian

Alessandro Zilio

Abstract For a class of optimal partition problems involving the half-Laplacian operator and a subcritical cost functionals, we derive the optimal regularity of the density-functions which characterize the partitions, for the entire set of minimizers. We present a numerical scheme based on the arguments of the proof and we collect some numerical results related to the problem.

Keywords Square root of the Laplacian · Spatial segregation · Strongly competing systems · Optimal regularity of limiting profiles · Singular perturbations

Mathematics Subject Classification (2010) Primary: 49Q10 · Secondary: 35B40 · 35R11 · 45C05 · 81Q05 · 82B10

1 Introduction

In recent time, the study of nonlocal operators has become a dominant subject in the regularity theory of minimization problems and elliptic equations. Originally inspired by modelling reasons, the study of non-local diffusion operators has revealed important in order both to test and to extend already understood theories concerning the behaviour of solutions to local problems.

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ ERC Grant Agreement n. 321186: "Reaction-Diffusion Equations, Propagation and Modelling" held by Henri Berestycki, and under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 339958: "Complex Patterns for Strongly Interacting Dynamical Systems" held by Susanna Terracini.

A. Zilio (✉)
Centre d'analyse et de mathématique sociales (CAMS), École des hautes études en sciences sociales (EHESS), 190-198 avenue de France, 75244 Paris Cedex 13, France
e-mail: azilio@ehess.fr; alessandro.zilio@polimi.it

A. Pratelli and G. Leugering (eds.), *New Trends in Shape Optimization*,
International Series of Numerical Mathematics 166,
DOI 10.1007/978-3-319-17563-8_13

Of the many non-local operators now object of study in the literature, this paper is concerned with possibly the easiest yet most fundamental one: the half-Laplacian. Given a smooth function $u \in \mathcal{C}_0^\infty(\mathbb{R}^N)$, the half-Laplacian operator $(-\Delta)^{1/2}$ is defined as the singular integral

$$(-\Delta)^{1/2}u := C_N \mathrm{pv} \int_{\mathbb{R}^N} \frac{u(x) - u(y)}{|x - y|^{N+1}} \mathrm{d}y$$

where the constant C_N is a normalization constant and pv stands for the principal value. For non-smooth functions, whenever possible, the operator is defined in the distributional sense (see [9], or the more recent [8], for comprehensive theory of the operator). As it is now well known, the above operator is related both to the infinitesimal generator of a Levy α-stable diffusion process and, via the Fourier transform $\mathcal{F}$, to the multiplication operator whose symbol is given by $|\xi|$ (see [8, Proposition 3.3]), that is

$$\forall u \in \mathcal{S}(\mathbb{R}^k) \quad (-\Delta)^{1/2}u = \mathcal{F}^{-1}(|\xi|\hat{u})$$

where $\mathcal{S}(\mathbb{R}^k)$ is the space of Schwartz functions, $\hat{u} = \mathcal{F}(u)$ and $\mathcal{F}^{-1}$ is the inverse transform. Moreover, from a variational point of view, the half-Laplacian can be related to the differential of the fractional Sobolev seminorm of $H^{1/2}(\mathbb{R}^N)$, that is

$$|u|^2_{H^{1/2}(\mathbb{R}^N)} := \langle (-\Delta)^{1/2}u, u\rangle = \frac{C_N}{2} \int_{\mathbb{R}^N \times \mathbb{R}^N} \frac{|u(x) - u(y)|^2}{|x - y|^{N+1}} \mathrm{d}x\mathrm{d}y.$$

The paper is devoted to the study of the regularity of optimal partition problems involving the fractional operator $(-\Delta)^{1/2}$. With this we mean that, given a set $\Omega \subset \mathbb{R}^N$ and a cost functional J associated to a suitable set of partitions of Ω, we wish to find the regularity shared by *all* the partitions that minimize J. More precisely, let us consider the functional space

$$H^{1/2}(\mathbb{R}^N) := \left\{ u : \|u\|^2_{H^{1/2}(\mathbb{R}^N)} := |u|^2_{H^{1/2}(\mathbb{R}^N)} + |u|^2_{L^2(\mathbb{R}^N)} < +\infty \right\}$$

and let $\Omega \subset \mathbb{R}^N$ be bounded and smooth set (i.e., with at least $\mathcal{C}^1$ boundary). Given some suitable functions $F_i : \Omega \times \mathbb{R} \to \mathbb{R}$, we introduce the functional

$$J(u_1, \dots, u_k) := \begin{cases} \displaystyle\sum_{i=1}^{k} \left(\frac{1}{2}|u_i|^2_{H^{1/2}(\mathbb{R}^N)} + \int_\Omega F_i(x, u_i)\mathrm{d}x \right) \\ \qquad\qquad \text{if } u_i \cdot u_j = 0 \text{ a.e. for every } j \neq i \\ +\infty \qquad\qquad \text{otherwise} \end{cases} \tag{1.1a}$$

and set the optimal partition problem on

$$\mathbb{S}^k_{L^2} := \left\{(u_1, \dots, u_k) : u_i \in H^{1/2}_\Omega(\mathbb{R}^N), \|u_i\|_{L^2(\mathbb{R}^N)} = 1\right\} \tag{1.1b}$$

where we used the notation $H^{1/2}_\Omega(\mathbb{R}^N) := \{w \in H^{1/2}(\mathbb{R}^N) : w|_{\mathbb{R}^N\setminus\Omega} = 0\}$. The main results we shall prove in the paper are the followings.

Theorem 1.1 *Let $\Omega \subset \mathbb{R}^N$ be bounded and smooth set. For each $i = 1, \dots, k$, let $F_i \colon \Omega \times \mathbb{R} \to \mathbb{R}$ be a Carathéodory function (that is, $(x, s) \mapsto F_i(x, s)$ is measurable in x and continuous in s) such that*

$$|F_i(x, s)| \le C_i(1 + |s|^p) \qquad \forall x \in \Omega, s \in \mathbb{R}$$

for a suitable constant $C_i \ge 0$, where $p < p^\sharp = \frac{2N}{N-1}$. Then there exists at least a minimizer of J in $\mathbb{S}^k_{L^2}$. Moreover, if $F_i(x, \cdot) \in \mathcal{C}^1(\mathbb{R})$ for a.e. $x \in \Omega$, then any minimizer $\mathbf{u} := (u_1, \dots, u_k) \in \mathcal{C}^{0,\alpha}(\mathbb{R}^N; \mathbb{R}^k)$ for any $\alpha \in (0, 1/2)$.

Theorem 1.2 *Under the assumptions of Theorem 1.1, let us assume also that*

(A) each function F_i is independent of x, $F_i \in \mathcal{C}^{2,\varepsilon}(\mathbb{R})$ for some $\varepsilon > 0$ and $F_i'(0) = 0$.

Then any minimizer $\mathbf{u}$ of J over the set $\mathbb{S}^k_{L^2}$ belongs to $\mathcal{C}^{0,1/2}(\mathbb{R}^N; \mathbb{R}^k)$ and satisfies the following Euler–Lagrange equation

$$u_i\left((-\Delta)^{1/2}u_i - F_i'(u_i)\right) = 0 \qquad \text{a.e. } in\ \Omega.$$

Remark 1.3 One could also consider partition problems of unbounded domains, for example with $\Omega = \mathbb{R}^N$, if the functions F_i can be used to ensure compactness: this can be achieved, for instance, if $F_i(x, s) = V(x)s^2$, with V positive and $V(x) \to +\infty$ for $|x| \to \infty$. In such a case the correct functional setting in given by the space

$$H^{1/2}_V(\mathbb{R}^N) := \left\{w \in H^{1/2}(\mathbb{R}^N) : \|w\|^2_{H^{1/2}_V(\mathbb{R}^N)} := |w|^2_{H^{1/2}(\mathbb{R}^N)} + \int_{\mathbb{R}^N} V w^2 \mathrm{d}x < \infty\right\}.$$

Associating to the bounded set Ω its indicator function

$$\chi_\Omega(x) := \begin{cases} 1 & \text{if } x \in \Omega \\ +\infty & \text{if } x \notin \Omega, \end{cases}$$

we see that $H^{1/2}_\Omega(\mathbb{R}^N) \equiv H^{1/2}_{\chi_\Omega}(\mathbb{R}^N)$. We shall not address this extension in the following, though the theory here developed may be used also to cover this case with little modifications.

At the moment no result asserting the regularity of the partition sets $\{(\omega_1, \dots, \omega_k)\}$ is known. In any case we observe that from Theorem 1.1 we can deduce that any

subset $\omega_i = \{u_i > 0\} \cup \{u_i < 0\}$ is an open set, which is already a non trivial result. Theorems 1.1 and 1.2 are analogous to well established results found in the case of standard diffusion operators, see for instance [3, 10]: in particular, they constitute the first step in the proof of the regularity of the free-boundary $\cup_i \partial\omega_i$, as done in [3, 11].

As a possible application, we can consider the case $F_i \equiv 0$. In such a situation, the optimal partition problem (1.1) is precisely given by the problem of finding k disjoint subsets $\omega_1, \dots, \omega_k$ of Ω such that the functional

$$(\omega_1, \dots, \omega_k) \mapsto \sum_{i=1}^{k} \lambda_1(\omega_i)$$

is minimal. Here $\lambda_1(\omega_i)$ stands for the first eigenvalue of the half-Laplacian in ω_i, defined as

$$\lambda_1(\omega_i) := \inf \left\{ |u|^2_{H^{1/2}(\mathbb{R}^N)} : \|u\|_{L^2(\mathbb{R}^N)} = 1,\ u = 0 \text{ a.e. on } \mathbb{R}^N \setminus \omega_i \right\}.$$

Remark 1.4 From a point of view of the applications, mainly linked to pattern formation in relativistic quantum systems, one could also consider a slightly different formulation of the optimal partition problem, as follows. Let us fix $k \in \mathbb{N}$ non-negative constants $m_1, \dots, m_k$ and let us introduce the operator $(-\Delta + m_i^2)^{1/2}$, which acts on smooth functions as

$$\forall u \in \mathcal{S}(\mathbb{R}^k) \quad (-\Delta + m_i^2)^{1/2} u = \mathcal{F}^{-1}((\xi^2 + m_i^2)^{1/2} \hat{u}).$$

Accordingly, one could introduce as a cost functional

$$R(u_1, \dots, u_k) := \begin{cases} \displaystyle\sum_{i=1}^{k} \left(\frac{1}{2} \langle (-\Delta + m_i^2)^{1/2} u_i, u_i \rangle + \int_\Omega F_i(x, u_i) \mathrm{d}x \right) \\ \qquad\qquad \text{if } u_i \cdot u_j = 0 \text{ a.e. for every } j \neq i \\ +\infty \qquad\qquad \text{otherwise} \end{cases}$$

again defined over the set $\mathbb{S}^k_{L^2}$. The same regularity results available for the functional J can be recast and extended without effort to the case of the functional R.

The last section is devoted to some numerical results. The simulations are obtained using an approximation scheme which is based on the proof of the Theorems 1.1 and 1.2: some comparisons with the results obtained in the case of the standard Laplacian are also presented.

To conclude, we would like to mention that results similar to those discussed above are available also in the case of any fractional power of the Laplacian $(-\Delta)^s$ with $s \in (0, 1)$: some of the needed preliminaries can be already found in [13].

2 Proof of the Results

As a first step, we shall prove that, under the assumptions of Theorem 1.1, the optimal partition problem admits at least a solution. Later we shall concentrate on the regularity of the whole set of solutions.

Lemma 2.1 *Let $\Omega \subset \mathbb{R}^N$ be bounded and smooth set. For each $i = 1, \dots, k$, let $F_i \colon \Omega \times \mathbb{R} \to \mathbb{R}$ be a Carathéodory function (that is, $(x, s) \mapsto F_i(x, s)$ is measurable in x and continuous in s) such that*

$$|F_i(x, s)| \leq C_i(1 + |s|^p) \qquad \forall x \in \Omega, s \in \mathbb{R}$$

for a suitable constant $C_i \geq 0$, where $p < p^\sharp = \frac{2N}{N-1}$. Then there exists a minimizer of J in $\mathbb{S}^k_{L^2}$.

Proof The lemma follows by the direct method of the calculus of variations. Indeed, we evince directly from the assumptions on the functions F_i that the functional J is weakly lower semicontinuous in $H^{1/2}(\mathbb{R}^N; \mathbb{R}^k)$ and moreover, since

$$\lim_{\|\mathbf{u}\|_{H^{1/2}} \to \infty} J(\mathbf{u}) = +\infty,$$

the functional J is also coercive in the weak topology of $H^{1/2}(\mathbb{R}^N; \mathbb{R}^k)$ (see [7, Example 1.14]). □

The regularity of the solutions to the previous minimization problem is in general hard to study directly. In order to simplify the analysis, in what follows we shall introduce two families of functionals which are related to the previous one in a precise way. The first family precisely implements the disjointness constraint in a relaxed way, through a penalization term: in particular we shall show that any sequence of minima to the family of functional converges to a minimum of the original functional. Our goal is to show that the topology of this convergence is sufficiently strong in order to ensure the regularity of the limiting densities. Unfortunately, since no result is known about the uniqueness of the optimal partition, the first proposed approximating procedure may fail to conclude the regularity of the whole set of optimal partitions. To avoid this issue, we need to introduce another family of functionals.

We start with the easier family of functionals.

Definition 2.2 Under the functional setting of Theorem 1.1, for any $\beta > 0$, let us introduce

$$J_\beta(u_1, \dots, u_k) := \sum_{i=1}^k \left(\frac{1}{2} |u_i|^2_{H^{1/2}(\mathbb{R}^N)} + \int_\Omega F_i(x, u_i) \mathrm{d}x \right) + \beta \sum_{j<i} \int_\Omega u_i^2 u_j^2 \mathrm{d}x.$$

Lemma 2.3 *Under the assumptions of Theorem 1.1, for every $\beta > 0$ there exists a minimizer $\mathbf{u}_\beta \in \mathbb{S}^k_{L^2}$ of J_β. Moreover, there exists a constant $C > 0$ (independent of β) such that $\|\mathbf{u}_\beta\|_{H^{1/2}(\mathbb{R}^N;\mathbb{R}^k)} \leq C$.*

Proof The proof is analogous to the one given in the limiting case $\beta = +\infty$. The main difference is represented by the presence of the interaction term, which is not sub-critical if $N \geq 3$. In this situation, it is sufficient to recall that, thanks positivity of β, the last term is lower semicontinuous, as a consequence of the Fatou's Lemma. □

Lemma 2.4 *It holds $\Gamma - \lim_{\beta \to +\infty} J_\beta = J$ (w.r.t. the weak $H^{1/2}(\mathbb{R}^N; \mathbb{R}^k)$-topology). Moreover, any sequence of minimizers $\{\mathbf{u}_\beta\}$ to J_β converges weakly in $H^{1/2}(\mathbb{R}^N; \mathbb{R}^k)$, up to a subsequence, to a minimizer of J.*

Proof The family of functionals J_β is increasing in β and converges pointwise to the functional J. As a consequence $\Gamma - \lim J_\beta = J$. The family J_β is also equi-coercive and this implies, up to subsequences, the convergence of the minimizers. See [7, Proposition 5.4, Corollary 7.20] for further details. □

As mentioned before, even though the family $\{\mathbf{u}_\beta\}$ converges, up to subsequences, to a minimizer of J, at the moment we can not say that any minimizer of J can be approximated in this way. In order to obtain a stronger conclusion, we need another step, involving the introduction of another functional, which will be the main object of the analysis in the following. For this purpose, let

$$e(s) := \sqrt{1 + s^2}$$

(we observe preliminarily that $|e'(s)| < 1$ for any $s \in \mathbb{R}$) and let $\bar{\mathbf{u}} \in \mathbb{S}^k_{L^2}$ be any minimizer of J.

Definition 2.5 Under the functional setting of Theorem 1.1, for any $\beta > 0$, we let

$$J^*_\beta(u_1, \dots, u_k) := \sum_{i=1}^k \left(\frac{1}{2} |u_i|^2_{H^{1/2}(\mathbb{R}^N)} + \int_\Omega [F_i(x, u_i) + e(u_i - \bar{u}_i)] \, \mathrm{d}x \right) + \frac{\beta}{2} \sum_{j<i} \int_\Omega u_i^2 u_j^2 \mathrm{d}x.$$

It is immediate to see that the proof of existence of minimizers developed for the functional J_β covers also the functional J^*_β. Moreover, since the functional J^*_β can be decomposed as

$$J^*_\beta(\mathbf{u}) = J_\beta(\mathbf{u}) + \sum_{i=1}^k \int_\Omega e(u_i - \bar{u}_i) \mathrm{d}x$$

it easily follows that any the sequence of minima $\{\mathbf{u}_\beta\}$ convergence weakly in $H^{1/2}(\mathbb{R}^N; \mathbb{R}^k)$ and strongly in $L^2(\mathbb{R}^N; \mathbb{R}^k)$ to the minimum $\bar{\mathbf{u}}$ of J.

Lemma 2.6 *There exists $C > 0$ independent of β such that*

$$\|\mathbf{u}_\beta\|^2_{H^{1/2}(\mathbb{R}^N;\mathbb{R}^k)} + \beta \int_{\mathbb{R}^N} \sum_{j\neq i} u^2_{i,\beta} u^2_{j,\beta} \mathrm{d}x \leq C.$$

Moreover if $F_i(x,\cdot) \in \mathcal{C}^1(\mathbb{R})$ for a.e. $x \in \Omega$, each function $u_{i,\beta}$ is a smooth solution to the Euler-Lagrange equation

$$(-\Delta)^{1/2} u_{i,\beta} + f_i(x, u_{i,\beta}) + e'(u_{i,\beta} - \bar{u}_i) u_{i,\beta} = \gamma_{i,\beta} u_{i,\beta} - \beta u_{i,\beta} \sum_{j\neq i} u^2_{j,\beta}$$

in Ω together with the boundary condition $u_i \equiv 0$ in $\mathbb{R}^N \setminus \Omega$. (here $f_i(x,s) := \partial_s F_i(x,s)$). The Lagrange multipliers $\gamma_{i,\beta}$ are bounded uniformly with respect to β.

Proof The first conclusion follows from the estimate

$$J^*_\beta(\mathbf{u}_\beta) \leq J^*_\beta(\bar{\mathbf{u}}) = J(\bar{\mathbf{u}}) \leq C$$

and the coercivity of J^*_β. Once the constraints are expressed through the Lagrange multipliers, the Euler-Lagrange equations can be derived classically, considering smooth variation of the minimizers $\mathbf{u}$. To conclude, testing the equation in $u_{i,\beta}$ by $u_{i,\beta}$ itself, the identity

$$\gamma_{i,\beta} = |u_{i,\beta}|^2_{H^{1/2}(\mathbb{R}^N)} + \int_\Omega u_{i,\beta} \left(f_{i,\beta}(x, u_{i,\beta}) + e'(u_{i,\beta} - \bar{u}_i) u_{i,\beta} + \beta u_{i,\beta} \sum_{j\neq i} u^2_{j,\beta} \right) \mathrm{d}x$$

yields the uniform bound for the multipliers. □

Corollary 2.7 *There exists a constant $C > 0$, independent of β, such that $\|\mathbf{u}_\beta\|_{L^\infty(\mathbb{R}^N;\mathbb{R}^k)} \leq C$.*

Proof This a consequence of the Brezis-Kato inequality, suitably generalized to the fractional setting (see [2, Sect. 5], and in particular [2, Theorem 5.2]). We give a sketch of the proof of such result in the appendix. □

We are in a position to apply the result contained in [12], which implies a first uniform regularity estimate for the densities $\mathbf{u}_\beta$.

Theorem 2.8 *For any $\alpha < 1/2$, there exists a constant $C > 0$ which is independent of β, such that*

$$\|\mathbf{u}_\beta\|_{\mathcal{C}^{0,\alpha}(\mathbb{R}^N;\mathbb{R}^k)} \leq C \qquad \forall \beta > 0.$$

In particular, the sequence $\mathbf{u}_\beta$ is compact in the $H^{1/2}_\Omega(\mathbb{R}^N;\mathbb{R}^k)$ topology and the uniform topology, and the limit $\bar{\mathbf{u}}$ of the family belongs to $\mathcal{C}^{0,\alpha}(\mathbb{R}^N;\mathbb{R}^k)$ for any $\alpha < 1/2$.

Proof This is a direct consequence of [12, Theorem 1.3]. The only difference here is that the forcing term in the Euler–Lagrange equation (see Lemma 2.6) here depends also on the variable x. But the same proof of [12, Theorem 1.3] works also in this case, under the verified hypothesis there exists a constant $C > 0$ such that

$$\sup_{\beta>0} \left\| f_i(x, u_{i,\beta}) + e'(u_{i,\beta} - \bar{u}_i)u_{i,\beta} - \gamma_{i,\beta}u_{i,\beta} \right\|_{L^\infty(\Omega)} < C. \qquad \square$$

We can conclude with the optimal regularity result mentioned in the introduction.

Theorem 2.9 *(*Theorem 1.2*) Under the previous assumptions, let us also suppose that*

(A) each function F_i is independent of x, $F_i \in \mathcal{C}^{2,\varepsilon}(\mathbb{R})$ for some $\varepsilon > 0$ and $F_i'(0) = 0$.

Then any minimizer $\mathbf{u}$ of J over the set $\mathbb{S}^k_{L^2}$ belongs to $\mathcal{C}^{0,1/2}(\mathbb{R}^N; \mathbb{R}^k)$ and satisfies the following Euler-Lagrange equation

$$u_i \left((-\Delta)^{1/2} u_i - F_i'(u_i) \right) = 0 \qquad \text{a.e. } in\ \Omega.$$

Proof As of now, we have shown that the minimizer $\mathbf{u} \in C^{0,\alpha}(\mathbb{R}^N; \mathbb{R}^k)$ for any $\alpha < 1/2$ and that the approximating sequence $\mathbf{u}_\beta$ converges to $\mathbf{u}$ strongly in $H^{1/2}(\mathbb{R}^N; \mathbb{R}^k)$ and uniformly in $\mathbb{R}^N$. Passing to the limit in the Euler-Lagrange equation and using the uniform estimate in Lemma 2.6, we infer that $\mathbf{u}$ satisfies

$$\begin{cases} u_i u_j = 0 & \text{in } \Omega, \text{ for any } i \neq j \\ u_i \left((-\Delta)^{1/2} u_i - F_i'(u_i) \right) = 0 & \text{a.e. in } \Omega \\ u_i = 0 & \text{in } \mathbb{R}^N \setminus \Omega. \end{cases}$$

We are then in a position to apply the result in [12, Theorem 1.2] (see also [12, Proposition 9.2.]). $\square$

3 Numerical Simulations

We now present some numerical validations of the theoretical results obtained so far. In the following we shall present a numerical algorithm which is based on the approximation scheme developed in the previous section, which has then no pretensions of being the most suitable from a computational point of view. All the simulations were carried out with a finite element approximation scheme, using the free software FreeFem++, available at http://www.freefem.org/ff++/.

Let us consider a specific example, which has also a possible interest in the applied science. Let us consider the optimal partition in k subsets of the unit ball in $\mathbb{R}^2$, that is, the optimal partition induced by the functional

$$J(u_1,\dots,u_k) := \begin{cases} \displaystyle\sum_{i=1}^{k} \frac{1}{2}|u_i|^2_{H^{1/2}(\mathbb{R}^2)} & \text{if } u_i \cdot u_j = 0 \text{ a.e. for every } j \neq i \\ +\infty & \text{otherwise} \end{cases} \tag{3.1}$$

constrained on the set $\mathbb{S}^k_{L^2}$ with $\Omega = B_1(0)$. Reasoning as in Sect. 2, we recall that the minimizers of the previous functional can be approximated by the minimizers of the approximating functional

$$J_\beta(u_1,\dots,u_k) := \sum_{i=1}^{k} \frac{1}{2}|u_i|^2_{H^{1/2}(\mathbb{R}^2)} + \beta \int_{\mathbb{R}^N} \sum_{j<i} u_i^2 u_j^2$$

and, finally, the minimizers can be obtained as solutions to the Euler–Lagrange equation

$$(-\Delta)^{1/2} u_{i,\beta} + \beta u_{i,\beta} \sum_{j\neq i} u^2_{j,\beta} = \gamma_{i,\beta} u_{i,\beta} \tag{3.2}$$

for suitable Lagrange multipliers $\gamma_{i,\beta}$: a meta-algorithm inspired by this approximation is illustrated in Algorithm 1. Let us observe that, in order to find a solution to the nonlinear system of equations, we have used a fixed point method based on the steepest descent algorithm alternated with a projection on the constraint $\mathbb{S}^k_{L^2}$. Being the underlying problem strongly non-convex, no results about the convergence to the minimal solution are known, if not under the assumptions that the initial guess is already close to the optimal configuration. Similar results may be found for example in [1], where a different algorithm is presented to study the optimal partition problem of the standard Laplace-Dirichlet eigenvalues.

Algorithm 1 Approximating scheme

procedure ApproximatingProcedure
  initialize $\gamma_i, u_i, \bar{u}_i$
  $\beta \leftarrow 1$
  $\bar{\beta} \leftarrow$ a large constant
  repeat
    repeat
      Solve $(-\Delta)^{1/2} u_i + \beta u_i \sum_{j\neq i} \bar{u}_j^2 = \gamma_i \bar{u}_i,\ u_i \equiv 0$ in $\mathbb{R}^2 \setminus B_1$
      $\bar{u}_i \leftarrow \dfrac{\alpha u_i + (1-\alpha)\bar{u}_i}{\|\alpha u_i + (1-\alpha)\bar{u}_i\|_{L^2}}$ ▷ Projection on $\mathbb{S}^k_{L^2}$, $\alpha \in (0,1)$
      $\gamma_i \leftarrow |\bar{u}_i|^2_{H^{1/2}(\mathbb{R}^2)} + \beta \displaystyle\int_{\mathbb{R}^2} \bar{u}_i^2 \sum_{j\neq i} \bar{u}_j^2 \mathrm{d}x$
    until convergence in L^2 with a prescribed tolerance
    $\beta \leftarrow 2\beta$
  until $\beta > \bar{\beta}$ and convergence in L^2
end procedure

The only non trivial task in the algorithm is given by the non-local equation in u_i: to solve it, we can make use of the extensional formulation of the half-Laplacian (see [4] and reference therein), which relates the Eq. (3.2) to

$$\begin{cases} -\Delta v_i = 0 & \text{in } \mathbb{R}^3_+ = \mathbb{R}^2 \times \mathbb{R}_+ \\ \partial_\nu v_i + \beta v_i \sum_{j \neq i} \bar{v}_j^2 = \gamma_i \bar{v}_i & \text{on } B_1 \times \{0\} \\ v_i \equiv 0 & \text{in } \mathbb{R}^2 \setminus B_1 \times \{0\} \end{cases} \tag{3.3}$$

where $v_i, \bar{v}_i \in H^1(\mathbb{R}^3_+)$ satisfy $v_i(\cdot, 0) = u_i$ and $\bar{v}_i(\cdot, 0) = \bar{u}_i$. The advantage of this formulations is that it can be readily approximated using finite element schemes, which are implement, for example, in the free software FreeFem++. To complete the approximating procedure, since (3.3) is defined on an unbounded set, we need to consider a bounding box $Q_L \subset \mathbb{R}^3_+$, $Q_L = (-L, L)^2 \times (0, 2L)$ with $L > 0$ large, and reformulated the equation as

$$\begin{cases} -\Delta v_i = 0 & \text{in } Q_L \\ \partial_\nu v_i + \beta v_i \sum_{j \neq i} \bar{v}_j^2 = \gamma_i \bar{v}_i & \text{on } (-L, L)^2 \times \{0\} \\ v_i \equiv 0 & \text{in } (-L, L)^2 \setminus B_1 \times \{0\} \end{cases}$$

for $v_i, \bar{v}_i \in H^1_{0,+}(Q_L) = \{w \in H^1_{0,+}(Q_L), w = 0 \text{ on } Q_L \setminus (-L, L)^2 \times \{0\}\}$. This last approximation is valid since, by the comparison, it is possible to show the solutions of Eq. (3.3) decay away from the origin $x = 0$. As a result, we can formulate the final Algorithm 2.

Algorithm 2 Approximating scheme revised

procedure ApproximatingProcedure
  $L, \bar{\beta} \leftarrow$ large constants
  initialize $v_i, \bar{v}_i \in H^1_{0,+}(Q_L), \gamma_i \in \mathbb{R}$
  $\beta \leftarrow 1$
  repeat
    repeat
      Solve $\begin{cases} -\Delta v_i = 0 & \text{in } Q_L \\ \partial_\nu v_i + \beta v_i \sum_{j \neq i} \bar{v}_j^2 = \gamma_i \bar{v}_i & \text{on } (-L, L)^2 \times \{0\} \\ v_i \equiv 0 & \text{in } (-L, L)^2 \setminus B_1 \times \{0\} \end{cases}$
      $\bar{v}_i \leftarrow \dfrac{\alpha v_i + (1-\alpha)\bar{v}_i}{\|\alpha v_i + (1-\alpha)\bar{v}_i\|_{L^2((-L,L)^2 \times \{0\})}}$
      $\gamma_i \leftarrow |\bar{v}_i|^2_{H^1_{0,+}(Q_L)} + \beta \displaystyle\int_{(-L,L)^2 \times \{0\}} \bar{v}_i^2 \sum_{j \neq i} \bar{v}_j^2 \mathrm{d}x$
    until convergence in L^2 of the traces up to a prescribed tolerance
    $\beta \leftarrow 2\beta$
  until $\beta > \bar{\beta}$ and convergence in L^2 of the traces
end procedure

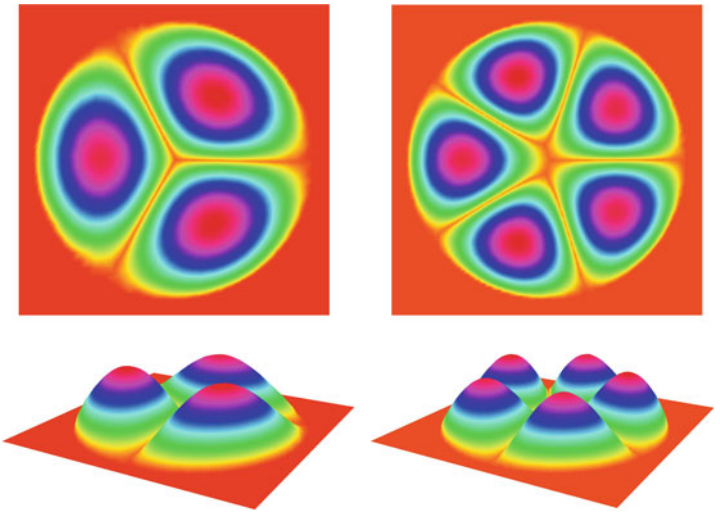


Fig. 1 Optimal partitions related to the problem 3.1. The non complete symmetry of the solutions, which can be expected by comparison with the partition problem involving the standard Laplacian, may be an effect of the presence of the bounding box Q_L: the more L is chosen large, the more such effect should be smoothed out. In any case, even for large values of L this phenomenon seems persistent

Remark 3.1 It should be mentioned that, though the extensional formulation of the fractional Laplacian alleviates us from solving non-local equations, it transforms N-dimensional optimal partition problems in to $N + 1$-dimensional boundary partition problems. For example, a planar problem is solved resorting to a fully three-dimensional one. Since both three-dimensional partition problem and, in general, boundary problems are stiff from a numerical point of view, it may seem surprising that the algorithms presented in this section converge in general, with just simple tunings of the parameters.

Remark 3.2 In order to obtain more accurate solutions, but sacrificing the efficiency, we have also inserted a step involving mesh-refinements.

As a result of the numerical simulation, we collect in Fig. 1 the solutions obtained for the problem (3.1) in the case of $k = 3$ components and $k = 5$ components. In Fig. 2 we show the corresponding solutions in the case of the standard Laplace operator: comparing the two situations, it is possible to see that, even though qualitatively very similar, the solutions *may* be different not only with respect to the regularity of their respective densities, but also in the geometry of the sets constituting the partition. In Fig. 3 we show the numerical results for the optimal partitions related to the problem 3.1, in the case of k = 10 components. It is tempting to extend the *hexagonal conjecture* (see for instance [1, 5]) also to the non-local setting.

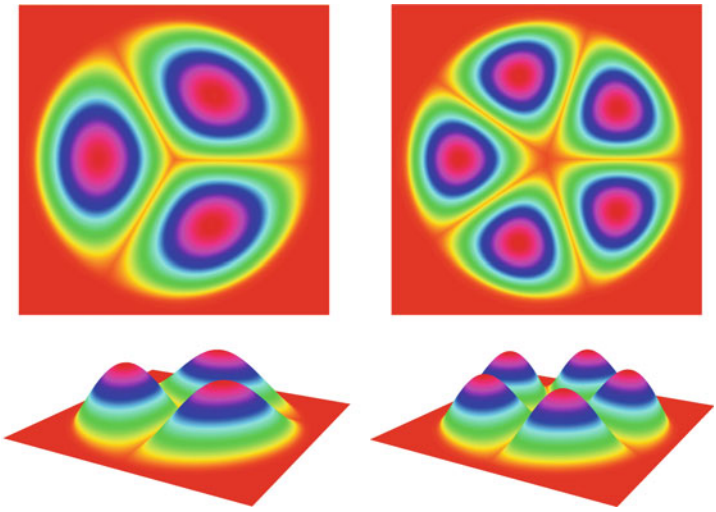


Fig. 2 Optimal partitions related to the problem 3.1, in the case the standard Laplace operator, obtained with the same approximating scheme employed for the half-Laplacian (see also [6] for further examples). The solutions are qualitatively similar, even though in this former case the transition between two different densities is smoother (in particular, solution are Lipschitz-continuous, as shown in [10])

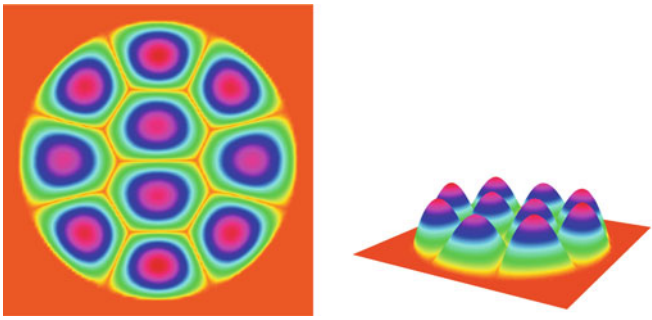


Fig. 3 Optimal partitions related to the problem 3.1, in the case of $k = 10$ components

Acknowledgments The author is indebted with the anonymous referee for suggesting many useful improvements to the original manuscript.

Appendix A: The Brezis-Kato Inequality

In this last section, we will give a proof of Corollary 2.7, using in fact the following version of the Brezis-Kato inequality

Lemma A.1 *Let $\Omega \subset \mathbb{R}^N$ be a smooth and bounded domain and let us consider $\mathbf{u} \in H^{1/2}_{\Omega}(\mathbb{R}^N, \mathbb{R}^k)$ to be solutions to the system*

$$(-\Delta)^{1/2} u_i = a_i(1 + |u_i|) - \beta u_i \sum_{j \neq i} u_j^2. \tag{A.1}$$

where $a_i \in L^N(\mathbb{R}^N)$. Then $u_i \in L^\infty(\mathbb{R}^N)$ for all $i = 1, \dots, k$ and the norm can be bounded uniformly in β with a constant that depends only on the $H^{1/2}$-norm of $\mathbf{u}$ and the L^N-norm of a_i.

Remark A.2 In order to apply the previous result to the setting of Corollary 2.7, it is sufficient to introduce the functions

$$a_{i,\beta} := \frac{(\gamma_{i,\beta} - Ke'(u_{i,\beta} - \bar{u}_i))u_{i,\beta} - f_i(x, u_{i,\beta})}{1 + |u_{i,\beta}|}$$

and to observe that, thanks to the sub-criticality of f_i and the uniform boundedness of Lagrange multipliers, we have $\|a_{i,\beta}\|_{L^N(\mathbb{R}^N)} \leq C$ uniformly in β.

Proof In order to simplify the proof, we resort to the extensional formulation of the half-Laplacian, relating the system (A.1) to

$$\begin{cases} -\Delta v_i = 0 & \text{in } \mathbb{R}^{N+1}_+ \\ \partial_\nu v_i = a_i(1 + |v_i|) - \beta v_i \sum_{j \neq i} v_j^2 & \text{on } \Omega \subset \partial\mathbb{R}^{N+1}_+ \\ v_i = 0 & \text{on } \mathbb{R}^N \setminus \Omega \end{cases}$$

where $v_i \in H^1(\mathbb{R}^{N+1}_+)$ satisfy $v_i(\cdot, 0) = u_i$. Let $g_\varepsilon \in \mathcal{C}^\infty(\mathbb{R})$ be a smooth approximation of the modulus functions, that is, $g_\varepsilon(t) = \sqrt{\varepsilon + t^2}$. The Stampacchia's lemma and the Lebesgue's theorem ensure that

$$g_\varepsilon(v_i) \to |v_i| \text{ in } H^1(\mathbb{R}^{N+1}_+), \quad g'_\varepsilon(v_i)v_i \to |v_i| \text{ in } L^2(\mathbb{R}^N)$$

For any test function $\phi \in H^1(\mathbb{R}^{N+1}_+)$ such that $\phi \geq 0$, we have

$$\begin{aligned} &\int_{\mathbb{R}^{N+1}_+} \nabla g_\varepsilon(v_i) \nabla \phi + \int_{\mathbb{R}^N} \beta g'_\varepsilon(v_i) v_i \sum_{j \neq i} v_j^2 \phi \\ &= \int_{\mathbb{R}^N} g'_\varepsilon(v_i) a_i(1 + |v_i|)\phi - \int_{\mathbb{R}^{N+1}_+} g''_\varepsilon(v_i) |\nabla v_i|^2 \phi \end{aligned}$$

and letting $\varepsilon \to 0^+$ we obtain

$$\int_{\mathbb{R}^{N+1}_+} \nabla |v_i| \nabla \phi + \int_{\mathbb{R}^N} \beta |v_i| \sum_{j \neq i} v_j^2 \phi \leq \int_{\mathbb{R}^N} \operatorname{sgn}(v_i) a_i(1 + |v_i|)\phi.$$

(similar computations are present in [12, Lemma 5.5]). As a result, each $|v_i| \in H^1(\mathbb{R}^{N+1}_+)$ is a subsolution of the equation in $w_i \in H^1(\mathbb{R}^{N+1}_+)$

$$\begin{cases} -\Delta w_i = 0 & \text{in } \mathbb{R}^{N+1}_+ \\ \partial_\nu w_i = |a_i|(1 + w_i) & \text{on } \Omega \subset \partial\mathbb{R}^{N+1}_+ \\ w_i = 0 & \text{on } \mathbb{R}^N \setminus \Omega \end{cases}$$

Thus, if we show a uniform bound for the functions w_i in L^∞, by the comparison principle we could evince that the same bounds holds for the functions $|v_i|$. To conclude it is then sufficient to recall the Brezis-Kato estimate for the half-Laplacian, shown in [2, Theorem 5.2], which implies the sought L^∞ bound. □

References

1. B. Bourdin, D. Bucur, É. Oudet, Optimal partitions for eigenvalues. SIAM J. Sci. Comput. **31**(6), 4100–4114 (2009/10)
2. X. Cabré, J. Tan, Positive solutions of nonlinear problems involving the square root of the Laplacian. Adv. Math. **224**(5), 2052–2093 (2010)
3. L.A. Caffarelli, F.-H. Lin, Singularly perturbed elliptic systems and multi-valued harmonic functions with free boundaries. J. Am. Math. Soc. **21**(3), 847–862 (2008)
4. L.A. Caffarelli, L. Silvestre, An extension problem related to the fractional Laplacian. Commun. Partial Differ. Equ. **32**(7–9), 1245–1260 (2007)
5. L.A. Cafferelli, F.H. Lin, An optimal partition problem for eigenvalues. J. Sci. Comput. **31**(1–2), 5–18 (2007)
6. S.-M. Chang, C.-S. Lin, T.-C. Lin, W.-W. Lin, Segregated nodal domains of two-dimensional multispecies Bose-Einstein condensates. Phys. D **196**(3–4), 341–361 (2004)
7. G. Dal Maso, *An Introduction to* Γ*-Convergence*. Progress in Nonlinear Differential Equations and their Applications, vol. 8. (Birkhäuser Boston, Inc., Boston, 1993)
8. E. Di Nezza, G. Palatucci, E. Valdinoci, Hitchhiker's guide to the fractional Sobolev spaces. Bull. Sci. Math. **136**(5), 521–573 (2012)
9. N.S. Landkof, *Foundations of Modern Potential Theory*. (Springer, New York, 1972) Translated from the Russian by A.P. Doohovskoy, Die Grundlehrender mathematischen Wissenschaften, Band 180
10. B. Noris, H. Tavares, S. Terracini, G. Verzini, Uniform Hölder bounds for nonlinear Schrödinger systems with strong competition. Commun. Pure Appl. Math. **63**(3), 267–302 (2010)
11. H. Tavares, S. Terracini, Regularity of the nodal set of segregated critical configurations under a weak reflection law. Calc. Var. Partial Differ. Equ. **45**(3–4), 273–317 (2012)
12. S. Terracini, G. Verzini, A. Zilio, Uniform Hölder bounds for strongly competing systems involving the square root of the Laplacian. Preprint arXiv:1211.6087
13. S. Terracini, G. Verzini, A. Zilio, Uniform Hölder regularity with small exponent in competition-fractional diffusion systems. Discret. Contin. Dyn. Syst. **34**(6), 2669–2691 (2014)