DISSERTATION

submitted

to the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of

Ruprecht Karl University of Heidelberg

for the degree of

Doctor of Natural Sciences

Put forward by

Karen Estefanía Loayza Romero, M.Sc.

born in Quito, Ecuador

Oral examination:

A Discrete Perspective on PDE-Constrained Shape Optimization Problems

Advisor: Prof. Dr. Roland Herzog

Abstract

It is well known among practitioners that the numerical solution of shape optimization problems constrained by partial differential equations (PDEs) often exhibits several difficulties. In particular, when the PDE is discretized by a finite element method, and the underlying mesh is used to represent the shape of the domain to be optimized directly, one often experiences a degeneracy of the mesh quality as the optimization progresses. The degeneracy manifests itself in some of the mesh cells thinning in the sense that at least one of its heights approaches zero.

Various techniques have been developed to circumvent this major obstacle in computational shape optimization. This thesis offers a new perspective on understanding the particularities of PDE-constrained shape optimization problems when they are treated under the discretize-then-optimize paradigm. We focus on two-dimensional problems, where the PDE is discretized using a finite element method, and the underlying mesh represents the discrete shape. Under these considerations, we make three main contributions. First, we study the set of all possible configurations of node positions a mesh of a given connectivity can attain. Then, using the language of simplicial complexes, we provide theoretical evidence that this set is a smooth manifold, and we term it the *manifold of planar triangular meshes*.

Secondly, we construct two complete Riemannian metrics for the manifold of planar triangular meshes, which avoid all possible self-intersections on a mesh. Moreover, they enjoy certain invariance properties under rigid body motions, and the latter is additionally invariant under uniform mesh refinements. In practice, endowing the manifold of triangular meshes with these metrics allows us to update the meshes following geodesics in any direction and as long as we want without jeopardizing their quality. This property can also be understood as degenerate meshes being infinitely far away from any regular mesh in terms of their geodesic distance.

Finally, alongside the newly proposed notion of the complete manifold of planar triangular meshes, we focus on the theoretical and computational aspects of discretized PDE-constrained shape optimization problems. We provide numerical evidence that such problems generally possess no solutions within the manifold of planar triangular meshes, even when the shape functional is bounded below. To overcome this drawback, we introduce a penalty functional which, briefly speaking, controls the mesh quality. When added to the shape functional, it renders well-posed discrete shape optimization problems, i. e., they possess at least one globally optimal solution. Subsequently, we solve the penalized problem using four different variants of the Riemannian steepest descent method. These variants depend on the metric used to transform cotangent vectors into tangent vectors and the metric used to update the meshes. Our numerical experiments reveal that using the proposed complete metrics to navigate the manifold is practically convenient since the optimization scheme does not need explicit monitoring of the mesh quality and can take arbitrarily large steps. Unfortunately, exploiting the properties of the complete metric is computationally challenging since the numerical integration of the respective geodesics is prohibitively expensive. However, we demonstrate that using the proposed Riemannian metric in gradient methods is still beneficial, even when combined with the inexpensive Euclidean retraction. Furthermore, the numerical evidence suggests gradient methods perform well in absence of the mesh quality penalty term when utilizing the complete metric.

Zusammenfassung

Es ist bekannt, dass die numerische Lösung von Formoptimierungsproblemen, die durch partielle Differentialgleichungen (PDEs) beschrieben sind, oft verschiedene Schwierigkeiten aufweist. Insbesondere, wenn die PDE durch eine Finite-Elemente-Methode diskretisiert wird und das zugrundeliegende Gitter verwendet wird, um direkt auch die Form des zu optimierenden Gebietes darzustellen, kommt es im Verlauf der Optimierung häufig zu einer Entartung der Gitterqualität. Die Entartung äußert sich daarin, dass Gitterzellen “dünn” werden, also dass mindestens eine ihrer Höhe gegen Null geht.

Es wurden verschiedene Methoden entwickelt, um dieses Problem in der numerischen Formoptimierung zu umgehen. Diese Arbeit bietet eine neue Perspektive zum Verständnis der Besonderheiten von PDE-beschränkten Formoptimierungsproblemen, wenn sie gemäß dem Vorgehen "diskretisieren, dann optimieren"behandelt werden. Wir konzentrieren uns auf zwei-dimensionale Probleme, bei denen die PDE mit einer Finite-Elemente-Methode diskretisiert wird und das zugrundeliegende Gitter die diskrete Form repräsentiert. Unter diesen Gesichtspunkten leisten wir drei wesentliche Beiträge: Zunächst untersuchen wir die Menge aller möglichen Konfigurationen von Knotenpositionen, die ein Gitter mit einer gegebenen Konnektivität erreichen kann. Dann liefern wir mit Hilfe des Begriffs der simplizialen Komplexe den theoretischen Beweis, dass es sich bei dieser Menge um eine glatte Mannigfaltigkeit handelt, und bezeichnen diese als *Mannigfaltigkeit der ebenen Dreiecksgitter* (*manifold of planar triangular meshes*).

Weiterhin konstruieren wir zwei vollständige Riemannsche Metriken für diese Mannigfaltigkeit der ebenen Dreiecksgitter, die alle möglichen Selbstüberschneidungen auf einem Gitter vermeiden. Darüber hinaus besitzen diese Metriken bestimmte Invarianz-Eigenschaften bei Starrkörperbewegungen. Die zweite Metrik ist zudem invariant gegenüber gleichmäßiger Gitterverfeinerung. Der Einsatz dieser Metriken in der Mannigfaltigkeit der ebenen Dreiecksgitter ermöglicht es uns, die Gitter zu aktualisieren, indem wir Geodäten in beliebiger Richtung und so lange wie gewünscht folgen, ohne die Qualität der Gitter zu beeinträchtigen. Diese Eigenschaft kann auch so verstanden werden, dass entartete Gitter in Bezug auf ihren geodätischen Abstand unendlich weit von jedem regulären Gitter entfernt sind.

Schließlich konzentrieren wir uns mit dem neu vorgeschlagenen Begriff der vollständigen Mannigfaltigkeit der ebener Dreiecksgitter auf die theoretischen und numerischen Aspekte von diskretisierten PDE-beschränkten Formoptimierungsproblemen. Wir liefern numerische Beweise dafür, dass solche Probleme im Allgemeinen keine Lösungen in der Mannigfaltigkeit der ebenen Dreiecksgitter besitzen, selbst wenn das Formfunktional nach unten beschränkt ist. Um dieses Problem zu überwinden, führen wir eine Straffunktion ein, welche die Gitterqualität kontrolliert. Wenn sie zum Formfunktional hinzuaddiert wird, ergibt sich ein gut gestelltes diskretes Formoptimierungsproblem, d. h., dass mindestens eine global optimale Lösung existiert. Anschließend lösen wir das Problem mit Strafterm mit vier verschiedenen Varianten der Riemannschen Methode des steilsten Abstiegs. Diese Varianten hängen von zwei Metriken ab: der Metrik zur Umwandlung von Kotangentialvektoren in Tangentialvektoren und der Metrik zur Aktualisierung der Gitter. Unsere numerischen Experimente zeigen, dass die Verwendung der vorgeschlagenen vollständigen Metriken zur Navigation auf der Mannigfaltigkeit praktisch ist, da das Optimierungsverfahren keine explizite Überwachung der Gitterqualität benötigt und beliebig große Schritte machen kann. Leider ist

die Ausnutzung der Eigenschaften der vollständigen Metrik numerisch eine Herausforderung, weil die numerische Integration der entsprechenden Geodäten unerschwinglich teuer ist. Wir zeigen jedoch, dass die Verwendung der vorgeschlagenen Riemannschen Metrik in Gradientenverfahren immer noch von Vorteil ist, selbst wenn sie mit der kostengünstigen euklidischen Retraktion kombiniert wird. Darüber hinaus legen die numerischen Resultate nahe, dass Gradientenmethoden auch ohne den Strafterm für die Gitterqualität gut funktionieren, wenn die vollständige Metrik verwendet wird.

Acknowledgements

Working on the development of this thesis is one of the most exciting and challenging tasks I have ever faced, none of which would have been possible without the support of many people to whom I would like to dedicate the following lines.

First of all, I would like to express my heartfelt gratitude to my supervisor and mentor, Prof. Roland Herzog, for believing in my potential and giving me the opportunity to work together. Thank you for all the fruitful discussions, for all the guidance I received during this time, for always being willing to help me, and for encouraging me to pursue my own projects.

I would also like to thank all the people from the working groups "Numerical Mathematics (Partial Differential Equations)" at the Chemnitz University of Technology and "Scientific Computing and Optimization" at Heidelberg University. Even though the pandemic took some time away from us, I enjoyed working with all of you.

I gratefully acknowledge the German Academic Exchange Service (DAAD) which funded my entire studies, and supported the publication of this thesis through the program: "Research Grants - Doctoral Programmes in Germany, 2017/18".

There are no words to describe how grateful I feel for always having the support of my family and friends from Ecuador; despite being 10,000 km away, their warmth, love, and encouragement never failed me. In particular, I want to thank my parents Leonardo and Saida, for having taught me not to be afraid of challenges, and for always having encouraged me to follow my dreams. To my brothers Nicolás and Agustín for always being by my side in every step I take.

Last but not least, I want to thank the person who has been there for me throughout this journey, Konstantin, my partner in life, home office mate, and coffee buddy. Thank you for always being there listening, advising, encouraging, and always believing in me. Without you this journey would not have been so much fun.

Contents

1 Introduction

The origins of shape optimization can be traced back to the 9th century B.C., when queen Dido asked for as much land as could be bound by the skin of a bull. Mathematically speaking, she was looking for the shape with maximal area for a given perimeter, and this problem is now called the *isoperimetric problem*. Another example is the optimal design of a ship which minimizes resistance, under the assumption that the shape of the hull is the rotation about the x-axis of a curve $y(x)$, as proposed by Newton. Thanks to Hadamard, 1908, who developed the notions of differentials of functions with respect to boundary variations, the constraints on the axissymmetry were no longer required, and more general designs were allowed. Nevertheless, the principle remained intact: finding the optimal design or shape according to certain criteria. The applications became more interesting when, additionally, the systems under study were governed by partial differential equations (PDEs). Nowadays, it is common to encounter applications like the design of electric motors Gangl et al., 2015, acoustic horns Schmidt, Wadbro, Berggren, 2016, aerodynamic structures Schmidt, Ilic, et al., 2013; Schillings, Schmidt, Schulz, 2011, elastic structures Allaire, Jouve, Toader, 2004, high-voltage devices Bandara et al., 2015. Furthermore, geometric inverse problems like impedance tomography can also be solved in a shape optimization framework, see e. g., Laurain, Sturm, 2016; Schulz, Siebenborn, Welker, 2015b; Hintermüller, Laurain, Yousept, 2015; Afraites, Dambrine, Kateb, 2008.

Most, if not all, of the aforementioned problems are solved numerically, and often under the optimize-then-discretize framework. In other words, the derivation of the first-order optimality conditions is performed on a continuous level with the help of shape calculus. The discretization of the problem takes place only at a later stage. Usually, a finite element method is employed to discretize the state equation, since it gives natural meaning to discrete shapes via the underlying triangulations. It is well known among practitioners that this approach often exhibits a number of difficulties. For example, a degeneracy of the mesh quality as the optimization progresses is to be expected.

A number of possible solutions to this major obstacle in computational shape optimization have been proposed in the literature. We do not aim to give a comprehensive overview at this stage but we mention that remeshing Wilke, Kok, Groenwold, 2005, mesh regularization and spatial adaptivity Doğan et al., 2007; Morin et al., 2012, and nearly-conformal transformations Iglesias, Sturm, Wechsung, 2018 have been considered as remedies.

In this thesis we shed new light on the phenomenon of mesh degeneracy in computational shape optimization. To this end, we restrict our discussion to two-dimensional PDE-constrained shape optimization problems. Our main goal is to analyze and numerically solve these problems under the discretize-then-optimize paradigm. Consequently, we make three major contributions, which we detail in section 1.1. Moreover, the outline of the thesis can be found in section 1.2.

1.1 Main Contributions

We consider the two-dimensional PDE-constrained problem proposed in Etling et al., 2020, and given by the following expression:

$$\text{Minimize} \quad \int_\Omega y \,\mathrm{d}x \quad \text{s. t.} \quad -\Delta y = r \text{ in } \Omega \quad \text{w.r.t.} \quad \Omega \subset \mathbb{R}^2. \tag{1.1}$$

The state y is subject to Dirichlet boundary conditions $y = 0$ on $\partial\Omega$ and the right-hand side function $r\colon \mathbb{R}^2 \to \mathbb{R}$ is given. Our main concern is to analyze and numerically solve problem (1.1) under a discretize-then-optimize approach. Even though we restrict our results to the discretized version of problem (1.1), we believe they can be naturally extended to more complex problems. We use the finite element method to discretize the state equation in (1.1). In particular, we use the space of piecewise linear, globally continuous functions, defined on a triangulation of Ω, which has N_V nodes and N_T triangles. Along the optimization process, we keep unchanged the connectivity of the triangulation. The node positions, on the other hand, are collected in a matrix $Q = [q_1, \dots, q_{N_V}] \in \mathbb{R}^{2\times N_V}$, and we use them as the optimization variables. Unfortunately, not all node positions $Q \in \mathbb{R}^{2\times N_V}$ give rise to an admissible triangulation, in the sense that, the nonempty intersection of triangular cells is either a vertex or an edge.

Manifold of Planar Triangular Meshes. The question whether a given mesh is admissible or not can be answered intuitively; however, the mathematical description of *all* admissible meshes is nontrivial. To this end, we use the language of simplicial complexes.

The connectivity information of triangular meshes in $\mathbb{R}^2$ is a pure, 2-path connected abstract simplicial 2-complex, which we will denote by Δ, and in definition 4.2.5 we term it a *connectivity complex*. This implies that for a given distribution of nodes over $\mathbb{R}^2$, the union of all the resulting triangles forms a connected subset of $\mathbb{R}^2$. However, abstract simplicial complexes are purely combinatorial objects, which can be thought as recipes for constructing meshes. The latter is only achieved after a correct assignment of the node positions collected in $Q = [q_1, \dots, q_{N_V}] \in \mathbb{R}^{2\times N_V}$, and thus, further notation is required.

Given a connectivity complex Δ, whose vertex set is $\{1, \dots, N_V\}$ and an assignment of the node positions $Q \in \mathbb{R}^{2\times N_V}$, we consider the associated collection of convex hulls, defined in (4.17), by the following expression:

$$\Sigma_\Delta(Q) \coloneqq \{\mathrm{conv}_Q(i_0, \dots, i_k) \,|\, \{i_0, \dots, i_k\} \in \Delta\} \subset \mathcal{P}(\mathbb{R}^2).$$

Thanks to this definition, we propose the following subsets of $\mathbb{R}^{2\times N_V}$. First, in definition 4.3.2 we introduce the **set of admissible meshes with connectivity** Δ as follows:

$$\mathcal{M}_0(\Delta) \coloneqq \left\{ Q \in \mathbb{R}^{2\times N_V} \,\middle|\, \begin{array}{l} \Sigma_\Delta(Q) \text{ is a geometric simplicial 2-complex} \\ \text{whose associated abstract simplicial complex is } \Delta \end{array} \right\}.$$

We prove in proposition 4.3.4, that it is an open submanifold of $\mathbb{R}^{2\times N_V}$. However, it is easy to devise examples of node positions which lie in different connected components of $\mathcal{M}_0(\Delta)$, as the ones depicted in figure 4.16. Therefore, we use the notion of orientation of simplicial complexes, and in definition 4.3.7 we consider the **set of admissible oriented meshes with connectivity** Δ as follows:

$$\mathcal{M}_+(\Delta) \coloneqq \big\{ Q \in \mathcal{M}_0(\Delta) \,\big|\, A_Q(i_0, i_1, i_2) > 0 \text{ for all 2-faces } [i_0, i_1, i_2] \text{ of } \Delta \big\},$$

where A_Q is the so-called *signed area* given in (4.2). In proposition 4.3.8 we prove the set $\mathcal{M}_+(\Delta)$ is an open submanifold of $\mathbb{R}^{2\times N_V}$. Once again, it is possible to construct examples of node positions which lie in different connected components of $\mathcal{M}_+(\Delta)$, as the ones suggested in example 4.3.9. In this case, one way to avoid this behavior is to consider meshes with no holes. However, we take a more practical approach and consider in definition 4.3.10 the **manifold of planar triangular meshes** given by the following expression:

$$\mathcal{M}_+(\Delta; Q_{\mathrm{ref}}) \coloneqq \left\{ Q \in \mathcal{M}_+(\Delta) \,\middle|\, \begin{array}{l} \text{there exists a continuous path} \\ \text{in } \mathcal{M}_+(\Delta) \text{ from } Q_{\mathrm{ref}} \text{ to } Q \end{array} \right\},$$

where $Q_{\text{ref}} \in \mathcal{M}_+(\Delta)$ is reference oriented mesh. Moreover, $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ represents the set of all oriented meshes which can be generated through continuous deformations of Q_{ref}. We prove that this set is indeed a connected, open submanifold of $\mathbb{R}^{2\times N_V}$ in theorem 4.3.11. In remark 4.3.13, we endow $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ with the natural smooth structure, which in turn, allows us to state that $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is a smooth manifold and therefore characterize its tangent space.

Complete Metrics for the Manifold of Planar Triangular Meshes. Since our main goal is to solve optimization problems on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$, having a smooth manifold is not enough. We need to endow it with more structure, namely a Riemannian metric. From all the possible choices of Riemannian metrics, we consider complete metrics, because they render geodesics which can be extended infinitely without leaving the manifold. This allows us to achieve mesh deformations along geodesic curves with any given initial velocity and for arbitrarily long times while maintaining nondegenerate meshes. This is a desired property from the computational point of view since it will naturally overcome the already mentioned problem of mesh degeneracy.

Given a Riemannian metric, it is relatively simple to verify if it is complete or not. However, the construction of complete metrics on smooth manifolds is not as straightforward. The results in this thesis are based on the following theorem.

Theorem (Gordon, 1973, Thm. 1). *Suppose that $\mathcal{M}$ is a connected manifold of class $\mathcal{C}^3$, endowed with a (not necessarily complete) Riemannian metric $\widetilde{g}$ with component functions $\widetilde{g}_{ab}$. If $f\colon \mathcal{M} \to \mathbb{R}$ is any proper function of class $\mathcal{C}^3$, then the Riemannian metric g defined by:*

$$g_{ab} = \widetilde{g}_{ab} + \frac{\partial f}{\partial x^a}\frac{\partial f}{\partial x^b}$$

is geodesically complete.

By virtue of this theorem, we focus on the construction of proper functions for the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. We construct two functions whose main aim is to penalize all the directions pending to self-intersection of a mesh. The first function f_1 is given in (5.6), and defined for the positive parameters $\alpha_1, \alpha_2, \alpha_3$ through the following expression:

$$f_1(Q; Q_{\text{ref}}) := \sum_{k=1}^{N_T}\sum_{\ell=0}^{2} \frac{\alpha_1}{h_Q^\ell\left(i_0^k, i_1^k, i_2^k\right)} + \sum_{[j_0,j_1]\in E_\partial} \sum_{\substack{i_0\in V_\partial \\ i_0\neq j_0, j_1}} \frac{\alpha_2}{D_Q(i_0; [j_0, j_1])} + \frac{\alpha_3}{2}\|Q - Q_{\text{ref}}\|_F^2,$$

where $h_Q^\ell(i_0^k, i_1^k, i_2^k)$ is the ℓ-th height of the triangle defined by the vertices $\{q_{i_0^k}, q_{i_1^k}, q_{i_2^k}\}$ defined in (5.5), D_Q is the 1-norm based distance from a boundary vertex $\{q_{i_0}\}$ to the boundary edge formed by the vertices $\{q_{j_0}, q_{j_1}\}$ given in (4.13), and $\|\cdot\|_F$ is the Frobenius norm. We prove in lemma 5.2.2 that the function f_1 is well-defined and continuous. Moreover, as a preparation for the proof of properness of f_1, we present in lemma 5.2.3 bounds on the geometric measurements of a triangle: edge lengths, heights, interior angles, among others, in terms of the values of f_1. Furthermore, inproposition 5.2.5 we show bounds on the distance from a vertex to an edge also in terms of the values of f_1. Then, the proof of f_1 being proper is given in theorem 5.2.6.

Assuming the existence of a $\mathcal{C}^3$-regularization for f_1, we construct the first complete metric for $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ in theorem 5.2.9. However, the function f_1 has one drawback: it is not invariant under uniform mesh refinements, which would be desirable from the PDE-constrained shape optimization perspective. Therefore, we propose a second function f_2

given in (5.22) as follows:

$$f_2(Q;Q_{\text{ref}}) \coloneqq \sum_{k=1}^{N_T} \frac{1}{N_T} \frac{\beta_1}{\psi_Q(i_0^k, i_1^k, i_2^k)} + \frac{\beta_2}{\sum_{k=1}^{N_T} A_Q\big(i_0^k, i_1^k, i_2^k\big)} + \sum_{[j_0,j_1]\in E_\partial} \sum_{\substack{i_0\in V_\partial \\ i_0\neq j_0,j_1}} \frac{1}{\#E_\partial \# V_\partial} \frac{\beta_3}{D_Q(i_0;[j_0,j_1])} + \frac{\beta_4}{2N_V}\|Q-Q_{\text{ref}}\|_F^2$$

with

$$\frac{1}{\psi_Q(i_0,i_1,i_2)} \coloneqq \frac{\big(E_Q^0(i_0,i_1,i_2)\big)^2 + \big(E_Q^1(i_0,i_1,i_2)\big)^2 + \big(E_Q^2(i_0,i_1,i_2)\big)^2}{4\sqrt{3}\, A_Q(i_0,i_1,i_2)},$$

where the parameters $\beta_1, \beta_2, \beta_3, \beta_4$ are positive, $E_Q^\ell(i_0,i_1,i_2)$ is the length of the ℓ-th edge, and $A_Q(i_0,i_1,i_2)$ is the area of the triangle defined by the vertices $\{q_{i_0}, q_{i_1}, q_{i_2}\}$. Moreover, the function $1/\psi_Q$ is a well-known triangle quality measure introduced in Bhatia, Lawrence, 1990. The proof of properness of f_2 is given in theorem 5.3.3 and builds up on the fact that in a sublevel set of f_2 it is possible to bound the edge lengths and heights of a triangle as given in proposition 5.3.2,. Thus, the second complete metric is given in proposition 5.3.5, assuming there exists a $\mathcal{C}^3$-regularization for f_2. We provide an example of a family of such regularizations in appendix B.

Discretized, PDE-Constrained Shape Optimization Problems. Alongside with the notion of the manifold of planar triangular meshes and its corresponding complete metrics, we study the discrete version of problem (1.1) given by the following expression:

$$\begin{aligned} \text{Minimize} \quad & \int_{\Omega_Q} y \,\mathrm{d}x \quad \text{w.r.t.} \quad Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}}),\ y \in S_0^1(\Omega_Q) \\ \text{s. t.} \quad & \int_{\Omega_Q} \nabla y \cdot \nabla v \,\mathrm{d}x = \int_{\Omega_Q} r\, v \,\mathrm{d}x \quad \text{for all } v \in S_0^1(\Omega_Q), \end{aligned} \tag{1.2}$$

where Ω_Q denotes the domain covered by the mesh with node positions Q, $S^1(\Omega_Q)$ is the finite element space of piecewise linear, globally continuous functions defined over Ω_Q, and $S_0^1(\Omega_Q)$ is the subspace of functions with zero Dirichlet boundary conditions.

We provide numerical evidence that problems like (1.2), in which the node positions of the mesh serve as the optimization variables, generally possess no solution in the manifold of planar triangular meshes, even when the shape functional is bounded below.

To overcome this drawback, we propose a penalized version of problem (1.2). We use the $\mathcal{C}^3$-regularization of the function f_2 given in (5.22) as the penalization, which naturally controls the mesh quality as the optimization progresses. In corollary 6.2.2, we present the proof of existence of at least one globally optimal solution for a general discretized, PDE-constrained shape optimization problem, which up to the author's knowledge is the first one of this kind.

Numerical Investigations. We use the Riemannian steepest descent method, and devise four different variants depending on the Riemannian metric chosen to compute gradients and the Riemannian metric chosen to update the meshes. The variants are termed Euclidean-Euclidean, Elasticity-Euclidean, Complete-Euclidean and Complete-Complete. The first component of their names refers to the metric used to compute gradients. The second component of their names refers to the metric used to update meshes. In other words, when the

second component is "Complete", it means we update the mesh following geodesics associated with the complete metric. The numerical experiments confirm that the unpenalized problem does not possess a solution in the manifold of planar triangular meshes. However, it can also be observed that by following geodesics associated with the proposed complete metric there is no need to explicitly monitor the mesh quality. In fact, the algorithm can take larger steps than in the Euclidean case. Unfortunately, the numerical integration of the aforementioned geodesic is prohibitively expensive. For that reason, the last experiments aim to compare among the most inexpensive variants of the steepest descent method, i. e., updating the meshes with the Euclidean metric. In this case, the numerical evidence suggests that the proposed Riemannian metric is still beneficial to use since the obtained meshes have better quality.

We remark that parts of this thesis have been taken from the following published and submitted papers.

(a) Roland Herzog, Estefanía Loayza-Romero;
A Discretize-Then-Optimize Approach to PDE-Constrained Shape Optimization;
Submitted to ESAIM: Control, Optimisation and Calculus of Variations;
https://arxiv.org/abs/2109.00076;

(b) Roland Herzog, Estefanía Loayza-Romero;
A Manifold of Planar Triangular Meshes with Complete Riemannian Metric;
Submitted to Mathematics of Computation;
https://arxiv.org/abs/2012.05624

(c) Caroline Geiersbach, Estefanía Loayza-Romero, Kathrin Welker;
Stochastic approximation for optimization in shape spaces;
SIAM Journal on Optimization;
https://doi.org/10.1137/20M1316111

(d) Tommy Etling, Roland Herzog, Estefanía Loayza, Gerd Wachsmuth;
First and second order shape optimization based on restricted mesh deformations;
SIAM Journal on Scientific Computing;
https://doi.org/10.1137/19M1241465

1.2 Outline of the Thesis

The thesis is structured as follows.

Chapter 2 – Fundamentals of PDE-Constrained Optimization gives an introduction into the mathematical background for PDE-constrained optimization, from both, the analytical and computational points of view. In section 2.1 we present the main results about the existence of linear, elliptic partial differential equations, the existence of solutions for optimal control problems, and their optimality conditions. Section 2.2 collects the main concepts of the discretization of PDE-constrained optimization problems, with special emphasis on the differences between the optimize-then-discretize and discretize-then-optimize approaches. Finally, in section 2.3 we introduce the generalities of the finite element method.

Chapter 3 – Optimize-then-Discretize Approach for PDE-Constrained Shape Optimization shows the state-of-the-art of computational shape optimization. In section 3.1 we give an overview of the most common representations of continuous shapes. Since it is well-known that shape optimization problems often do not possess a solution, we recall, in section 3.2, two examples with no optimal solutions. In the first example topological changes of the shapes are allowed, while in the second they are not. Section 3.3 briefly describes

the main notions of shape calculus, which are commonly used to derive the first-order optimality conditions for shape optimization problems. The main concepts of the discretization of shape optimization problems are gathered in section 3.4. Section 3.5 collects some of the techniques developed in the last years to preserve the quality of the meshes along the optimization process.

Chapter 4 – Discrete Shape Manifolds focuses on the construction of the manifold of planar triangular meshes. However, we first provide an overview of previously proposed discrete shape manifolds in section 4.1. Since the description of the manifold of planar triangular meshes is based on the language of simplicial complexes, we introduce the fundamentals on simplicial complexes in section 4.2. We collect important inequalities involving the geometric measurements of a triangle like edge lengths, heights, interior angles, in– and circumradius, among others. Section 4.3 presents a step-by-step construction of the manifold of planar triangular meshes, as the set of all oriented meshes which can be generated through continuous deformations of a reference oriented mesh. We prove the most important properties of this manifold in theorem 4.3.11 and we characterize its tangent space.

Chapter 5 – Complete Metrics for the Manifold of Triangular Meshes describes four different Riemannian metrics for the manifold of planar triangular meshes. The first two, are already known, the Euclidean and linear elasticity metrics are introduced in section 5.1. Then, in section 5.2, we present the first complete metric based on a proper function f_1. Since this metric is not invariant under uniform mesh refinements, we propose a second metric. The metric proposed in section 5.3 inherits all the properties from the first one, and additionally is invariant under uniform mesh refinements. Unfortunately, the geodesic equations associated to these metrics can be solved only numerically. For this purpose, in section 5.4 we present the Hamiltonian formulation of geodesics and describe the Störmer–Verlet scheme. The numerical experiments from section 5.5 investigate how meshes deform under the proposed complete metrics. They provide numerical evidence of the completeness of the metrics. Moreover, it can be observed that the meshes deformed under this metric keep their aspect ratios around satisfactory values.

Chapter 6 – Discretize-then-Optimize Approach for PDE-Constrained Shape Optimization addresses our main concern: the analysis and numerical solution of discretized, PDE-constrained shape optimization problems. In section 6.1, we present numerical evidence that, unfortunately, these problems possess no solution in the manifold of planar triangular meshes, even when the shape functional is bounded below. Therefore, we propose a penalized version of the problem in section 6.2 for which we prove existence of at least one globally optimal solution in the manifold of planar triangular meshes. We also derive the first-order optimality conditions of the penalized problem. We present in section 6.3 four different variants of the Riemannian steepest descent method, which we use to numerically approximate the solutions of the problem. We conduct three numerical experiments described in section 6.4 which compare the performance of the four variants of the Riemannian steepest descent method, and exhibit the differences between the penalized and unpenalized problems.

Chapter 7 – Conclusions and Outlook summarizes the main results of the thesis and points out some future research directions.

2 Fundamentals of PDE-Constrained Optimization

Contents

Optimization problems involving partial differential equations are ubiquitous in our daily lives. Applications as simple as the optimal cooling of a plate or as complex as the generation of weather forecasts can be considered in this field. These problems are interesting from both practical and theoretical points of view, and in this chapter we aim to collect, as briefly as possible, some of the fundamental concepts on this topic.

The chapter is structured as follows. Section 2.1 introduces the background knowledge on linear elliptic PDEs and the theory behind optimization problems subject to them. In section 2.2, we discuss, from a general perspective, the most common approaches for the numerical solutions of PDE-constrained problems. We emphasize the difference between the *optimize-then-discretize* and *discretize-then-optimize* approaches. Finally, in view of our applications to PDE-constrained shape optimization problems, section 2.3 is devoted to the presentation of the basic notions of the finite element method.

This chapter is based on the books Ciarlet, 2002; De los Reyes, 2015; Hinze et al., 2009; Tröltzsch, 2010.

2.1 Analytical Background

In this section, we focus on the notions required for understanding PDE-constrained optimization problems from a continuous perspective. It is divided into three main topics: first, generalities of elliptic PDEs, gathered in subsection 2.1.1. Second, in subsection 2.1.2, we detail the definitions required to establish the existence of solutions for general PDE-constrained problems. Finally, subsection 2.1.3 shows the different approaches which can be used to compute optimality conditions with particular emphasis on the Lagrangian approach.

2.1.1 Theory of Partial Differential Equations

We focus on domains $\Omega \subset \mathbb{R}^2$, whose interior, boundary, and closure are denoted by $\mathrm{int}(\Omega)$, $\partial\Omega$ and $\overline{\Omega}$, respectively. Definition 2.1.1 presents the primary assumption regarding the sufficient regularity of their boundaries.

Definition 2.1.1. *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with boundary $\Gamma = \partial\Omega$. We say Ω has a $\mathcal{C}^{k,1}$-boundary or belongs to the class $\mathcal{C}^{k,1}$, $k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$ if there exist finitely many local coordinate systems $S_1, \dots, S_M$, functions $h_1, \dots, h_M$, and constants $a, b > 0$ having the following properties:*

(*i*) *The functions h_i, $1 \leq i \leq M$, are k-times differentiable on an open superset of the interval $I = \{x \,|\, |x| \leq a\}$, and their partial derivatives of order k are Lipschitz continuous on I.*

(*ii*) *For any* $P \in \Gamma$ *there is some* $i \in \{1, \dots, M\}$*, such that in the coordinate system* S_i *there is some* x *belonging to the open interval* $]-a, a[$ *with* $P = (x, h_i(x))$.

(*iii*) *In the local coordinate system* S_i *we have:*

$$(x_1, x_2) \in \Omega \Leftrightarrow x_1 \in I,\ h_i(x_1) < x_2 < h_i(x_1) + b,$$
$$(x_1, x_2) \notin \Omega \Leftrightarrow x_1 \in I,\ h_i(x_1) - b < x_2 < h_i(x_1).$$

If Ω satisfies definition 2.1.1 with $k = 0$ we say it is a *domain with Lipschitz boundary* or a *regular* domain.

Let us fix some notation. We denote by $\mathcal{C}^k(\Omega)$, with $k \in \mathbb{N}_0$, the linear space of all real-valued functions on Ω that, together with their partial derivatives up to order k, are continuous in Ω. Moreover, we write $\mathcal{C}(\Omega)$ instead of $\mathcal{C}^0(\Omega)$. The linear space $\mathcal{C}^k(\overline{\Omega})$ is the set of all elements of $\mathcal{C}^k(\Omega)$ that together with their partial derivatives up to order k can be continuously extended to $\overline{\Omega}$.

We focus on linear elliptic problems, and especially on the Poisson's equation with homogeneous Dirichlet boundary conditions given by:

$$\begin{cases} -\Delta y &= r, \quad \text{in } \Omega, \\ y &= 0, \quad \text{on } \Gamma, \end{cases} \tag{2.1}$$

where r is a given function on Ω. If $r \in \mathcal{C}(\overline{\Omega})$, then a function $y \in \mathcal{C}^2(\overline{\Omega})$ satisfying the equation in the usual sense is said to be *classical* solution of problem (2.1). On the other hand, its weak formulation is given by:

$$\text{Find } y \in H_0^1(\Omega) \text{ s.t. } \int_\Omega \nabla y \cdot \nabla v \, \mathrm{d}x = \int_\Omega r v \, \mathrm{d}x \quad \text{for all } v \in H_0^1(\Omega), \tag{2.2}$$

where $\nabla y \cdot \nabla v$ denotes the Euclidean inner product between the two-dimensional vectors ∇y and ∇v.

Now we briefly recall the definition of the function spaces which appear in equations (2.1) and (2.2).

Definition 2.1.2. *Let* $\Omega \subset \mathbb{R}^2$ *be a bounded, and open domain with Lipschitz boundary. The space* $L^2(\Omega)$ *is the Lebesgue space of all equivalence classes of Lebesgue measurable functions* y*, which differ only on a set of zero measure, such that* $\int_\Omega |y|^2 \, \mathrm{d}x < +\infty$*. Moreover, we denote by* $H^1(\Omega)$ *the linear space of all functions from* $L^2(\Omega)$ *whose first-order partial derivatives also belong to* $L^2(\Omega)$*. In other words,*

$$H^1(\Omega) = \left\{ y \in L^2(\Omega) \,\middle|\, \frac{\partial y}{\partial x_i} \in L^2(\Omega),\ \textit{for } i = 1, 2 \right\}. \tag{2.3}$$

The space $H^1(\Omega)$ *is a Hilbert space if endowed with the following scalar product*

$$(y, v)_{H^1(\Omega)} = (y, v)_{L^2(\Omega)} + \left(\frac{\partial y}{\partial x_1}, \frac{\partial v}{\partial x_1} \right)_{L^2(\Omega)} + \left(\frac{\partial y}{\partial x_2}, \frac{\partial v}{\partial x_2} \right)_{L^2(\Omega)},$$

where $y, v \in H^1(\Omega)$.

- *The space* $H^1(\Omega)$ *is a special case of the Sobolev space* $W^{1,p}(\Omega)$ *with* $p = 2$.
- *We define* $W_0^{1,p}(\Omega)$ *as the closure of* $\mathcal{C}_0^\infty(\Omega)$ *in* $W^{1,p}(\Omega)$*, then* $H_0^1(\Omega) = W_0^{1,2}(\Omega)$*.* $H_0^1(\Omega)$ *can be understood as the space of functions that vanish at the boundary.*

The existence of weak solutions of equation (2.1) is obtained using Lax–Milgram's Theorem (see e. g., Brezis, 2011, Cor. 5.8, p. 140).

2.1.2 Theory of Optimal Control

This section aims to formulate, as generally as possible, the kind of problems we consider in this thesis, and state conditions under which these problems have a solution.

A general formulation for a PDE-constrained optimization problem reads as follows:

$$\text{Minimize} \quad J(y,u) \quad \text{w.r.t. } u \in U_{\text{ad}} \quad \text{s.t. } E(y,u) = 0, \tag{2.4}$$

where $J\colon Y \times U \to \mathbb{R}$ is the so-called *objective function*. The equation $E(y,u) = 0$ is called *state equation*, and the operator E is defined on $E\colon Y \times U \to Z$. The spaces U, Y, and Z are usually assumed to be Banach spaces and we denote by U', Y' and Z' their topological dual spaces, respectively. The *control* is denoted by u, and the set of *admissible controls* by $U_{\text{ad}} \subset U$.

It is customary to assume that for each $u \in U_{\text{ad}}$, there exists a unique $y \in Y$ such that $E(y,u) = 0$, which allows us to write the relation $U \ni u \mapsto y(u) \in Y$. This mapping, denoted by $G\colon U \to Y$, is referred to as the *control-to-state* operator.

Remark 2.1.3. *If the state equation is associated with a linear elliptic PDE, as the Poisson's equation (2.1), the well-posedness of G is directly obtained from the Lax–Milgram's Theorem.*

The introduction of the control-to-state operator G allows us to write the *reduced functional* j as $j(u) = J(G(u),u)$, which simplifies the problem as follows:

$$\text{Minimize } \; j(u) = J(G(u),u) \quad \text{w.r.t. } \; u \in U_{\text{ad}}. \tag{2.5}$$

The existence of solutions of problem (2.5) depends on the properties of the reduced objective function j, the set of admissible controls U_{ad}, the control-to-state operator G and the function spaces where j and G are defined.

If U_{ad} is empty, of course, problem (2.5) has no solution and by convention the value of the infimum is $+\infty$. On the other hand, if U_{ad} is nonempty, and j fails to be bounded from below, we write $\inf_{u\in U_{\text{ad}}} j(u) = -\infty$. Conversely, if U_{ad} is nonempty and j is bounded from below, then the infimum is finite, yet may fail to be attained by an element of U_{ad}. Thus, we say problem (2.5) has a global solution if the infimum $\inf_{u\in U_{\text{ad}}} j(u)$ can be attained by an element of U_{ad}.

Definition 2.1.4. *A control $\overline{u} \in U_{\text{ad}}$ is called globally* optimal *if and only if*

$$j(\overline{u}) \leq j(u), \tag{2.6}$$

for all $u \in U_{\text{ad}}$.

We present some definitions required to state existence results.

Definition 2.1.5. *A functional $j\colon U \to \mathbb{R}$ is called* weakly lower semi continuous *if for every weakly convergent sequence $u_n \rightharpoonup u$ in U, $n \to +\infty$ it follows that*

$$j(u) \leq \liminf_{n\to+\infty} j(u_n). \tag{2.7}$$

Definition 2.1.6. *A functional $j\colon U \to \mathbb{R}$ is called* radially unbounded *if*

$$\lim_{\|u\|_U\to\infty} j(u) = +\infty, \tag{2.8}$$

where $\|\cdot\|_U$ stands for the norm of the Banach space U.

The most common results of the existence of solutions for problem (2.5) are given in what follows.

Theorem 2.1.7. *Let U be a reflexive Banach space, and U_{ad} a nonempty, closed, convex and bounded subset of U. If j is weakly lower semi continuous, then problem (2.5) has a solution.*

Proof. This proof can be found in Hinze et al., 2009, Thm. 1.45, p. 55. □

If the set U_{ad} is unbounded, the existence of solutions can be obtained through the following result.

Theorem 2.1.8. *Let U be a reflexive Banach space, and U_{ad} a nonempty, closed, and convex subset of U. If j is weakly lower semi continuous and radially unbounded, then problem (2.5) has a solution.*

Proof. The proof of this theorem can be found in De los Reyes, 2015, Thm. 3.1, p. 26. □

Given our applications, which most of the objective functions lack convexity, we cannot, in general, discuss the uniqueness of the solutions. In section 6.2, we will show a more general proof for nonreflexive spaces, where the existence of solutions is obtained from to the properties of j.

2.1.3 Optimality Conditions

We end this review by introducing the techniques used to characterize stationary points of problem (2.5) with particular emphasis on the *Lagrangian approach*.

We recall the notation used in the previous section; j stands for the reduced functional, which is defined on the Banach space U. For the rest of the chapter, we focus on the case $U_{\text{ad}} = U$, i. e., we consider only unconstrained problems. The following theorem gives first-order necessary optimality conditions.

Theorem 2.1.9. *Let $j\colon U \to \mathbb{R}$ be a Fréchet differentiable mapping on an open subset of U and $\overline{u} \in U$ be a minimum of j. Then, $j'(\overline{u})d = 0$ for all $d \in U$, where $j'(\overline{u})$ stands for the first-order derivative of j at $\overline{u}$.*

Proof. The proof of this theorem can be found in De los Reyes, 2015, Cor. 3.1, p. 30. □

We also define, formally, stationary points.

Definition 2.1.10. *Let $j\colon U \to \mathbb{R}$ be Fréchet differentiable around $\overline{u} \in U$. Then, we say $\overline{u}$ is a stationary point of j if $j'(\overline{u}) = 0$.*

It is also customary to assume that the derivative of the operator $E(y, u)$ with respect to y, denoted by $\mathrm{d}_y E(y, u)$ (which is a bounded linear operator from Y to Z) is continuously invertible, so that by the implicit function theorem (see e. g., Ciarlet, 2013, Thm. 7.13-1, p.548) we can ensure $y(u)$ is continuously differentiable. The expression for $y'(u)$ is obtained by differentiating $E(y(u), u) = 0$ in a direction $d \in U$, as follows:

$$\mathrm{d}_y E(y(u), u) y'(u) d + \mathrm{d}_u E(y(u), u) d = 0. \tag{2.9}$$

To compute the derivative of the reduced functional, one could use two equivalent formulations: the sensitivity or the adjoint approach.

The *sensitivity* approach aims to compute the derivative $j'(u)$ in terms of directional derivatives, $j'(u)d$. For a fixed direction $d \in U$, directional derivatives can be obtained by using the chain rule, i. e., it holds:

$$j'(u)d = \langle \mathrm{d}_y J(y(u), u)\,, y'(u)d \rangle_{Y',Y} + \langle \mathrm{d}_u J(y(u), u)\,, d \rangle_{U',U},$$

where $y'(u)$ can be computed by solving the linearized state equation (2.9), and $\langle \cdot\,, \cdot \rangle_{Y',Y}$ stands for the dual pairing between Y' and Y.

If one is interested in computing the entire derivative $j'(u)$, this approach will mean one needs to evaluate the directional derivatives for all elements of a basis of U. We recall this implies solving the linear system (2.9) once for each direction. Of course, this procedure

will increase its computational costs as the dimension of U increases. For more information we refer the reader to Hinze et al., 2009, Sec. 1.6.1, p. 58.

The *adjoint* approach provides a direct and more efficient way of representing the derivative of j, given by the following expression:

$$j'(u)d = \langle y'(u)^{\mathrm{T}}\, \mathrm{d}_y J(y(u),u)\,, d\rangle_{U',U} + \langle \mathrm{d}_u J(y(u),u)\,, d\rangle_{U',U}, \tag{2.10}$$

obtained by using the adjoint operator $y'(u)^{\mathrm{T}} \in \mathcal{L}(Y', U')$ of $y'(u)$. From here, it follows that we do not need the entire operator $y'(u)$, but only the vector $y'(u)^{\mathrm{T}}\, \mathrm{d}_y J(y(u),u) \in U'$, which can be computed by rearranging terms in (2.9), and multiplying both sides of the resulting equation by $\mathrm{d}_y J(y(u),u)$ as follows:

$$y'(u)^{\mathrm{T}}\, \mathrm{d}_y J(y(u),u) = -\,\mathrm{d}_u E(y(u),u)^{\mathrm{T}}\; \mathrm{d}_y E(y(u),u)^{-\mathrm{T}}\, \mathrm{d}_y J(y(u),u). \tag{2.11}$$

Now, by defining $p(u) = -\,\mathrm{d}_y E(y(u),u)^{-\mathrm{T}}\, \mathrm{d}_y J(y(u),u)$, we obtain that

$$j'(u) = \mathrm{d}_u E(y(u),u)^{\mathrm{T}} p(u) + \mathrm{d}_u J(y(u),u). \tag{2.12}$$

Of course, we do not want to invert the operator $\mathrm{d}_y E(y(u),u)$; therefore, we define the *adjoint equation*.

Definition 2.1.11. *Let us consider problem (2.5), then the adjoint state $p(u) \in Z'$ satisfies the following linear system:*

$$\mathrm{d}_y E(y(u),u)^{\mathrm{T}} p(u) = -\,\mathrm{d}_y J(y(u),u). \tag{2.13}$$

The representation of the derivative of the reduced functional j under the adjoint approach can also be derived by the Lagrangian method, which is the approach we will use in this thesis.

If we consider the problem (2.4), its associated Lagrangian is given by $L\colon Y \times U \times Z' \to \mathbb{R}$, such that:

$$(y,u,p) \mapsto L(y,u,p) = J(y,u) + \langle p\,, E(y,u)\rangle_{Z',Z}. \tag{2.14}$$

Then, for the reduced functional it holds:

$$j(u) = L(y(u),u,p) = J(y(u),u) + \langle p\,, E(y(u),u)\rangle_{Z',Z}.$$

By differentiation we obtain:

$$j'(u)d = \langle \mathrm{d}_y L(y(u),u,p)\,, y'(u)d\rangle_{Y',Y} + \langle \mathrm{d}_u L(y(u),u,p)\,, d\rangle_{U',U}. \tag{2.15}$$

Choosing $p = p(u)$ such that $\mathrm{d}_y L(y(u),u,p) = 0$, which coincides with the adjoint equation (2.13), we get:

$$\begin{aligned} j'(u)d &= \langle \mathrm{d}_u L(y(u),u,p(u))\,, d\rangle_{U',U}, \\ &= \langle \mathrm{d}_u J(y(u),u)\,, d\rangle_{U',U} + \langle p(u)\,, \mathrm{d}_u E(y(u),u)d\rangle_{Z',Z}, \\ &= \langle \mathrm{d}_u J(y(u),u)\,, d\rangle_{U',U} + \langle \mathrm{d}_u E(y,u)^{\mathrm{T}} p(u)\,, d\rangle_{U',U}, \end{aligned}$$

from which we obtain the expression

$$j'(u) = \mathrm{d}_u J(y(u),u) + \mathrm{d}_u E(y(u),u)^{\mathrm{T}} p(u). \tag{2.16}$$

Having computed the expression for the derivative of the reduced functional, we are ready to present the first-order necessary optimality conditions of problem (2.4).

Proposition 2.1.12. *Let $(\overline{y}, \overline{u})$ be a local solution of the problem (2.4). Furthermore, let us assume the following holds:*

- *(α) $J\colon Y \times U \to \mathbb{R}$ and $E\colon Y \times U \to Z$ are continuously Fréchet differentiable.*
- *(β) For all $u \in V$ in a neighborhood $V \subset U$, the state equation $E(y,u) = 0$ has a unique solution $y = y(u) \in Y$.*
- *(γ) $\mathrm{d}_y E(y(u),u)$ is continuously invertible.*

Then, there exists a unique $\overline{p} \in Z'$ *such that*

$$\begin{aligned} E(\overline{y}, \overline{u}) &= 0 && \text{(State Equation)}, \\ \mathrm{d}_y E(\overline{y}, \overline{u})^{\mathrm{T}} \overline{p} + \mathrm{d}_y J(\overline{y}, \overline{u}) &= 0 && \text{(Adjoint Equation)}, \\ \mathrm{d}_u E(\overline{y}, \overline{u})^{\mathrm{T}} \overline{p} + \mathrm{d}_u J(\overline{y}, \overline{u}) &= 0 && \text{(Design Equation)}. \end{aligned} \tag{2.17}$$

Proof. See e. g., De los Reyes, 2015, Thm. 3.3, p. 31. □

2.2 Discretization Concepts

For most PDE-constrained optimization problems, the solutions can not be computed explicitly. Therefore, one can only approximate their solutions numerically. In other words, one proposes an approximation of the objective function and the state equation as close as possible to the original ones, which can also be encoded in a computer. These functions can be obtained, for example, by approximating all the function spaces under consideration by finite-dimensional ones, by approximating all the appearing differential operators, among other methods. We refer to this process as *discretization*.

In this section, we will treat discretization as an abstract process. We assume the existence of a procedure which allows us to write J_h opposite to J as the discrete objective function, y_h instead of y as the discrete state variable, u_h instead of u as the discretized control, and $E_h(y_h, u_h) = 0$ as the discrete counterpart of the state equation $E(y, u) = 0$, referring to the elements of problem (2.4). Notice that in this case, the functionals J_h, E_h are defined on finite-dimensional vector spaces, which we denote by U_h, Y_h, Z_h as a counterpart of U, Y, Z.

Two different approaches can be derived depending on when we decide to discretize. They are known as the *optimize-then-discretize* approach and, oppositely, the *discretize-then-optimize* approach. In what follows, we describe their main features and the rationale for choosing one over the other.

Let us start with the *optimize-then-discretize* method. As its name indicates, the discretization is performed in a later stage. To be more precise, for a given PDE-constrained problem as in (2.4), one derives its optimality conditions given in (2.17). Only after the optimality conditions have been obtained, one proceeds to discretize. In our notation this means we need to use an appropriate numerical method to solve the following system of equations:

$$\begin{aligned} E_{h_1}(y_{h_1}, u_{h_1}) &= 0 && \text{(Discrete State Equation)}, \\ \mathrm{d}_y E_{h_2}(y_{h_2}, u_{h_2})^{\mathrm{T}} p_{h_2} + \mathrm{d}_y J_{h_2}(y_{h_2}, u_{h_2}) &= 0 && \text{(Discrete Adjoint Equation)}, \\ \mathrm{d}_u E_{h_3}(y_{h_3}, u_{h_3})^{\mathrm{T}} p_{h_3} + \mathrm{d}_u J_{h_3}(y_{h_3}, u_{h_3}) &= 0 && \text{(Discrete Design Equation)}. \end{aligned} \tag{2.18}$$

We use different subindices h_1, h_2, h_3 since the discretization methods used for each equation do not need to be the same. We will comment more on this matter later in this subsection.

For numerical methods to solve systems of equations, we refer to Quarteroni, Valli, 1994, Part I, Sec. 2, Saad, 2003; Golub, van Loan, 1983.

Now, we focus on the *discretize-then-optimize* method, which suggests discretization is performed first. In other words, the problem under consideration is now given by:

$$\text{Minimize} \quad J_h(y_h, u_h) \quad \text{w.r.t.} \ u_h \in U_{\mathrm{ad},h} \quad \text{s. t.} \ E_h(y_h, u_h) = 0, \tag{2.19}$$

where $U_{\mathrm{ad},h} = U_h$ is the set of all admissible discretized controls. The theory presented in section 2.1 can also be adapted to the study of this finite-dimensional problem. The numerical solution of problem (2.19) can be done by employing iterative methods like the

steepest descent method, quasi-Newton methods like SR1 or BFGS, and second-order methods like Newton. For a complete overview of these methods and the optimality theory of finite-dimensional problems, we refer the reader to Nocedal, Wright, 2006.

We discuss the main differences between the aforementioned approaches and provide reasons to use one or the other.

Each approach has its advantages and disadvantages. We can compare them in terms of accuracy and consistency, as suggested in Van Keulen, Haftka, Kim, 2005, which was studied for structural design problems. However, we think it can be applied for more general PDE-constrained problems.

Accuracy can be measured as the difference, either absolute or relative, of the approximated derivatives and the derivatives of the continuous model. Alternatively, the difference of the approximated derivatives and the derivatives of the numerical model is called *consistency*. Using these terms, one can conclude that the *optimize-then-discretize* approach is accurate (which depends on the size of the discretization) but may not be consistent. In other words, by using the discretized derivatives, we could not assure an algorithm will reach the exact optimality conditions of the continuous model. Conversely, the *discretize-then-optimize* approach is not accurate (because the derivative of the continuous and the discrete problem may differ) but is consistent, i. e., solving a problem under this paradigm will return a solution of the discretized problem (since the derivatives coincide). However, this solution may not be a good solution for the continuous problem.

It is also worth to highlight that in the *discretize-then-optimize* method, the adjoint variable, denoted by p, is entirely determined by the method used to discretize the state variable and the objective function. In the *optimize-then-discretize* method, one could choose different discretization approaches for the state and adjoint variables. However, the following must be taken into account: the design equation from (2.18) relates the control and adjoint variables, and the chosen discretization method needs to consider this relation. We can refer to this property as conservative discretization for the control (see e. g., Hinze et al., 2009, Not. 3.2, p. 164).

If either the state or the control variables have extra constraints, it is preferable to use the *optimize-then-discretize* method. For example, if the problem has control constraints, the regularity of the state will be often lower (restricted by the regularity of the control) than that of the adjoint. Therefore, using this method will allow us to choose a different discretization on the adjoint variable, in order to exploit this feature. The same strategy can be applied when the problem has state constraints, since the adjoint variable may have less regularity than the state. We reflect this possible choices in the different indices h_1, h_2, h_3 used in the system of equations (2.18).

Conversely, one reason to use the *discretize-then-optimize* method is to avoid a complex analysis of the state equation in functional spaces; especially, when a rather simplified one can be performed from a discrete perspective. For example, we refer the reader to De los Reyes, Loayza-Romero, 2019, where this approach was used, as a way to overcome the complex structure of the underlying hyperbolic conservation law, to solve an inverse problem subject to the inviscid Burgers' equation.

Choosing different discretization methods for the state and adjoint variables may result in optimality systems that are no longer square nor symmetric. Conversely, using the same discretization method for the state and adjoint variables will lead to the straight forward equivalence of the methods.

2.3 The Finite Element Method

Up until this point, we have discussed the discretization of problem (2.4) as an abstract process. This section aims to give an specific example of such procedure. We focus on the finite element method, since it gives a natural meaning to discretized domains via *triangulations*, which we will use in our applications to shape optimization.

The main aim of the finite element method is to provide a way of constructing the spaces U_h, Y_h, Z_h, which are the finite-dimensional counterparts of the spaces where problem (2.4) was initially defined. Particularly, we use this method to approximate the solution of equation (2.2). This section is based on the books Grossmann, Roos, Stynes, 2007; Quarteroni, 2009; Quarteroni, Valli, 1994.

We start by introducing the main idea of the Galerkin method. We consider the problem

$$\text{Find } y \in Y \quad \text{s.t. } A(y,v) = \langle r\,, v\rangle_{Y',Y} \quad \text{for all } v \in Y, \tag{2.20}$$

where $A(\cdot,\cdot)$ is a continuous and coercive bilinear form, and $r \in Y'$. Then, by the Lax–Milgram's Theorem, we know this problem has a unique solution. For the weak formulation of (2.1), we have $Y = H_0^1(\Omega)$, the bilinear form $A(y,v) = \int_\Omega \nabla y \cdot \nabla v \,\mathrm{d}x$ and the linear form $\langle r\,, v\rangle_{Y',Y} = \int_\Omega rv \,\mathrm{d}x$.

Let Y_h be a family of spaces that depends on a positive parameter h, such that

$$Y_h \subset Y, \quad \dim Y_h = N_h < +\infty, \quad \text{for all } h > 0. \tag{2.21}$$

The *Galerkin problem* is then given by:

$$\text{Find } \; y \in Y_h \quad \text{s.t. } A(y,v) = \langle r\,, v\rangle_{Y_h',Y_h} \quad \text{for all } v \in Y_h. \tag{2.22}$$

Let us consider now $\{e_a\}_{a=1}^{N_h}$, a basis of Y_h. Then the condition of problem (2.22) can be verified for each element of the basis, instead of all the elements of the space Y_h. Moreover, since $y \in Y_h$, it can also be expressed as linear combinations of the basis functions, as $y = \sum_{a=1}^{N_h} \vec{y}_a e_a$, where $\vec{y} = [\vec{y}_1, \ldots, \vec{y}_{N_h}]^{\mathrm{T}}$, is a vector of real numbers. Thus, the solution of the Galerkin problem is reduced to the solution of the following linear system:

$$\mathbb{A}\vec{y} = \vec{r}$$

where $\mathbb{A}_{ab} = A(e_a, e_b)$ and $\vec{r}_b = \langle r\,, e_b\rangle_{Y_h',Y_h}$. The matrix $\mathbb{A}$ is called *stiffness matrix*.

In a nutshell, the finite element method considers the decomposition of Ω into small cells (elements). It defines finite-dimensional spaces as the set of all functions of a certain regularity when restricted to a cell which, at the same time, have global continuity. We focus in this thesis on linear Lagrange elements. Their description is given in terms of the general finite element definition from Ciarlet, 2002, Sec. 2.3, p. 78, i. e., the element domain, the space of shape functions and the set of degrees of freedom. Briefly, the degrees of freedom can be understood as the values that must be assigned to univoquely define the functions themselves.

Definition 2.3.1. *A finite element is called a linear Lagrange element ($\mathbb{P}_1$) if the following properties are satisfied:*

(a) The cell $K \subset \mathbb{R}^2$ is the convex hull of three distinct vertices of $\mathbb{R}^2$, denoted by q_0, q_1, q_2.

(b) We consider the finite-dimensional space of all polynomials of degree at most one defined on each cell K.

(c) The degrees of freedom $\sigma_0, \sigma_1, \sigma_2$ are the point evaluations in the vertices of K, i. e., $\sigma_\ell(v) = v(q_\ell)$. The set of all degrees of freedom of a cell is denoted by Σ.

(d) The local nodal basis *functions N_1, N_2, N_3 are given by $N_\ell(x) = \lambda_{\ell-1}(x)$, where λ_ℓ is the ℓ-th barycentric coordinate of x w.r.t. K.*

Remark 2.3.2. *For linear Lagrange elements, consider the affine map T_K given by:*

$$T_K(\widehat{x}) = \left[q_1 - q_0,\ q_2 - q_0\right] \widehat{x} + q_0, \tag{2.23}$$

where $K = \operatorname{conv}\{q_0, q_1, q_2\}$ and $\widehat{K} = \operatorname{conv}\{(0,0)^{\mathrm{T}}, (1,0)^{\mathrm{T}}, (0,1)^{\mathrm{T}}\}$, with $\operatorname{conv}\{q_0, q_1, q_2\}$ the convex hull of the vectors $\{q_0, q_1, q_2\}$. Then, $K = T_K(\widehat{K})$. We call $\widehat{K}$ a reference cell, *and all the elements K which can be obtained through the mapping T_K are called* world elements. *Formally, one can guarantee that $K = T_K(\widehat{K})$ are also linear Lagrange elements (cf., Ciarlet, 2002, Thm. 2.3.1, p. 86) by considering the local nodal basis functions as $\widehat{N}_\ell = N_\ell \circ T_K$.*

Now, we have all the ingredients we need to define a *finite element space*. We consider a family of linear Lagrange elements denoted by $\{(K_k, P_k, \Sigma_k)\}_{k=1}^{N_T}$. We denote by N_T the number of elements. Our primary goal is to glue them together to define the space Y_h in terms of the collection of Lagrange elements. However, certain assumptions need to be made about the distribution of the elements on the plane to guarantee that the functions defined on Y_h have the required properties, i. e., we want $Y_h \subset H^1(\Omega)$ to obtain a conforming discretization.

The collection of elements needs to satisfy two properties:

- they have to approximate the domain Ω as well as possible, and
- they have to relate to each other so global continuity can be achieved.

It is customary to assume that Ω is a polyhedral domain, i. e., is an open, bounded, connected subset such that its closure is the union of a finite number of polyhedra. This allows us to avoid technical discussions about "faces" of non-polygonal elements. We will assume from now on that all the domains are polyhedral.

Regarding the relationship between the elements K_k, we start by recalling definition 2.3.1, from where we know that each cell K_k is the convex hull of three distinct vertices, also known as nodes, which belong to $\mathbb{R}^2$. Moreover, an edge of K_k is defined as being the convex hull of two of the three nodes from K_k. These nodes cannot be arbitrarily distributed in the space. Their distribution must satisfy specific properties, which we describe in definition 2.3.3.

Definition 2.3.3. *Let $\Omega \subset \mathbb{R}^2$ be a polyhedral domain. Moreover, let us denote by $\mathcal{T}_h$ a collection of N_T cells $\{K_k\}_{k=1}^{N_T}$. We say $\mathcal{T}_h$ is a triangulation of Ω if the following conditions are satisfied:*

(i) It holds that:

$$\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K.$$

(ii) The interior of K is nonempty for all $K \in \mathcal{T}_h$.

(iii) For all $K_k, K_{\overline{k}} \in \mathcal{T}_h$ such that $k \neq \overline{k}$, the intersection of $\operatorname{int}(K_k)$ and $\operatorname{int}(K_{\overline{k}})$ is empty.

(iv) For all $K_k, K_{\overline{k}} \in \mathcal{T}_h$ such that $k \neq \overline{k}$, the intersection of K_k and $K_{\overline{k}}$ is either the empty set, a vertex or an edge of both cells.

Moreover, if h_K is the diameter of K for each $K \in \mathcal{T}_h$, then $h = \max_{K \in \mathcal{T}_h} h_K$.

Thanks to definition 2.3.3, we gave meaning to the sub index h used to describe the abstract discretization process from section 2.2.

Condition *(iv)* of definition 2.3.3 restricts the triangulations under consideration to so-called *conforming* ones. Even though it is possible to establish the finite element method on nonconforming triangulations, we restrict our study to the conforming type.

From definition 2.3.3, one can infer that a triangulation is completely defined by two different pieces of information: the way its nodes are connected, and their distribution on the space. Figure 2.1 shows three examples of distributions of nodes and connectivity

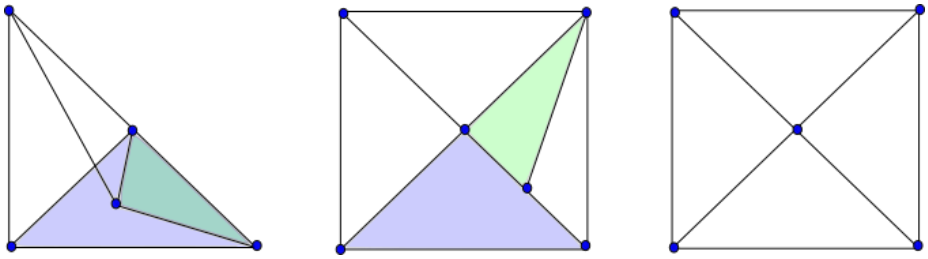


Figure 2.1. Illustration of triangulation with: incorrect distribution of vertices (left), incorrect connectivity information (center), and correct distribution of vertices and correct connectivity information (right), according to definition 2.3.3.

information, which represent triangulations. In the first two examples at least one condition is not satisfied; in the last example, all conditions are satisfied. Figure 2.1 (left) shows a set of nodes for which the connectivity information is correct, however the distribution of the nodes is not. In fact, this can be observed because the intersection of interior of the cells K_k and $K_{\overline{k}}$ (depicted in blue and green, respectively) is nonempty, i. e., they do not satisfy condition *(ii)*. In figure 2.1 (center), we depict a distribution of nodes which is correct, however, their connectivity information is not. Again, this can be observed because the intersection between the cells K_k and $K_{\overline{k}}$ (blue and green, respectively) is something other than the empty set, a vertex, or an edge of both cells. In other words, this triangulation does not satisfy condition *(iv)*. Finally, in figure 2.1 (right), we present an example of a triangulation for which both the connectivity information and the distribution of its nodes are correct.

We proceed to present the space of linear Lagrange finite elements.

Definition 2.3.4. *The space Y_h associated to a triangulation $\mathcal{T}_h$ is the set of globally continuous functions $v_h \colon \overline{\Omega} \to \mathbb{R}$ such that the restriction of v_h denoted by $v_h|_K$ is a polynomial of degree at most one on K, for each $K \in \mathcal{T}_h$. In other words, we consider the following set:*

$$Y_h = \{v \in \mathcal{C}(\overline{\Omega}) \,|\, v_h|_K \in \mathbb{P}_1, \text{for all } K \in \mathcal{T}_h\}. \tag{2.24}$$

Moreover, the degrees of freedom for each element can be collected in the set

$$\Sigma_h = \{v_h(q_i) \,|\, q_i \text{ is a vertex of the triangulation, with } i = 1, \dots, N_V\},$$

where N_V denotes the number of vertices of the triangulation.

Together with this definition we also consider the *global degrees of freedom* given by

$$\sigma_{K,\ell}(v|_K) = \sigma_i(v)$$

for $\ell = 0, 1, 2$ and some i which will only depend on the numbering of the global degrees of freedom, and $i = 1, \dots, N_V$. These relations can be stored in a vector $c^{K,\ell} \in \mathbb{R}^{N_V}$ which satisfies

$$c_i^{K,\ell} = \begin{cases} 1 & \text{if the position of the global degree of freedom } i \text{ coincides} \\ & \text{with the } \ell\text{-th local degree of freedom of } K, \\ 0 & \text{otherwise.} \end{cases} \tag{2.25}$$

Moreover, this vector allows us to define the *global basis functions* as follows.

Definition 2.3.5. *Let $\{\sigma_i\}_{i=1}^{N_V}$ be the global degrees of freedom of Y_h. The functions $\{e_a\}_{a=1}^{N_V} \subset Y_h$ with $\sigma_i(\mathrm{e}_a) = \delta_a^i$, where δ_a^i is the Kronecker delta symbol, are said to be the global nodal basis functions, and they can be expressed in terms of the vector containing the relation*

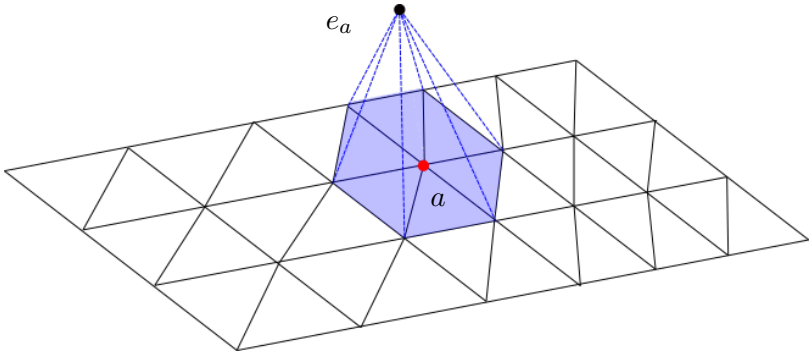


Figure 2.2. Example of a two-dimensional global nodal basis function associated to linear Lagrange elements described in definition 2.3.5, also known as hat function.

between local and global degrees of freedom $c^{K,\ell}$ (2.25) *and the local nodal basis functions* $N_{K,\ell}$ *introduced in definition 2.3.1, as follows:*

$$e_a|_K = \sum_{\ell=0}^{2} c_a^{K,\ell} N_{K,\ell}. \tag{2.26}$$

In the case of linear Lagrange elements they are also known as "hat functions". See figure 2.2 for an illustration.

We end this subsection by describing the assembly of the stiffness and mass matrices for (2.2).

Thanks to the bilinearity of $A(\cdot,\cdot)$ associated to the problem (2.1), the stiffness matrix is given by the expression:

$$\mathbb{A}_{ab} = A(e_a, e_b) = \sum_{k=1}^{N_T} \int_{K_k} \nabla e_a \cdot \nabla e_b \,\mathrm{d}x = \sum_{k=1}^{N_T} \sum_{m,n=0}^{2} (\mathbb{A}_{K_k})_{m,n} c_b^{K_k,m} c_a^{K_k,n},$$

where $\{e_a\}_a^{N_V}$ are global nodal basis functions, and the matrix

$$(\mathbb{A}_K)_{m,n} = \int_K \nabla N_{K,n} \cdot \nabla N_{K,m} \,\mathrm{d}x \tag{2.27}$$

is called the *local stiffness matrix* (element of $\mathbb{R}^{3\times 3}$), associated to the cell K. Analogously, one could assemble the right-hand side of the Poisson's equation and obtain the *local force vector*

$$(\vec{r}_K)_m = \int_K r N_{K,m} \,\mathrm{d}x, \tag{2.28}$$

with $m = 0, 1, 2$. The local contributions $\mathbb{A}_K$ and $\vec{r}_K$ can be computed in virtue of remark 2.3.2, and the substitution rule of integration, which reads:

$$\int_K g(x)\,\mathrm{d}x = \int_{\widehat{K}} g(T_K(\widehat{x}))|\det DT_K(\widehat{x})|\,\mathrm{d}\widehat{x} = |\det T_K| \int_{\widehat{K}} g(T_K(\widehat{x}))\,\mathrm{d}\widehat{x}. \tag{2.29}$$

The mapping T_K is given in (2.23), and $\widehat{K}$ is the reference element. Then,

$$\int_K \nabla N_{K,n}(x) \cdot \nabla N_{K,m}(x)\,\mathrm{d}x = |\det T_K| \int_{\widehat{K}} \left[T_K^{-\mathrm{T}} \widehat{\nabla} \widehat{N}_n(\widehat{x})\right] \cdot \left[T_K^{-\mathrm{T}} \widehat{\nabla} \widehat{N}_m(\widehat{x})\right] \mathrm{d}\widehat{x},$$

where we have used the definition of the local nodal basis functions for the reference element given in remark 2.3.2. The symbol $\widehat{\nabla}$ denotes the gradient of a function defined on the reference element $\widehat{K}$.

Since the local nodal basis functions $\widehat{N}_n$, $n = 0, 1, 2$, are polynomials, their integrals can be computed exactly. For the assembly of the local force vector $\vec{r}_K$ given in (2.28), quadrature formulas can be used. For example, by using one quadrature point, the local contribution of the force vector can be approximated as follows:

$$(\vec{r}_K)_m = |\det T_K| \omega_{\widehat{K}} r(T_K(\xi_{\widehat{K}})) \widehat{N}_m(\xi_{\widehat{K}}), \tag{2.30}$$

where $\omega_{\widehat{K}}$ denotes the weight of the quadrature rule, and $\xi_{\widehat{K}}$ the quadrature point. We use $\omega_{\widehat{K}} = |\widehat{K}|$, and $\xi_{\widehat{K}} = (1/3, 1/3, 1/3)$ (in barycentric coordinates).

3 Optimize-then-Discretize Approach for PDE-Constrained Shape Optimization

Contents

The most common approach for the numerical solution of PDE-constrained shape optimization problems is optimize-then-discretize. As commented in section 2.2, the main idea behind this approach is to compute the optimality conditions in a continuous setting and discretize only at a later stage.

In the context of shape optimization, this means we firstly need to study the possible representations of shapes in a continuous framework. An overview of these representations is considered in section 3.1. Then, in section 3.2, we study the particularities on the existence of solutions for these problems. Section 3.3 presents the generalities of the sensitivity analysis. Section 3.4 aims to discuss about the discretization of PDE-constrained shape optimization problems. In most cases, the finite element method for the discretization of the state equation is preferred. This choice implies discretized shapes are decomposed in simple cells. In this thesis, we focus in the kind of finite element method which consider triangulations. For this reason, in section 3.5 we focus on the presentation of the techniques developed to preserve or improve the quality of the underlying meshes. Sections 3.4 and 3.5 are based on the bibliographic study presented in Etling et al., 2020, Sec. 1.

Briefly, this chapter can be understood as the state-of-the-art of the numerical solution of PDE-constrained shape optimization problems.

PDE-constrained shape optimization problems can be formulated as follows:

$$\text{Minimize} \quad J(\Omega, y) \quad \text{w.r.t.} \ \ \Omega \subset \mathcal{P}(\mathbb{R}^2), \quad \text{s.t.} \ \ E(\Omega, y) = 0, \tag{3.1}$$

where Ω is the optimization variable and an element of the power set of $\mathbb{R}^2$ denoted by $\mathcal{P}(\mathbb{R}^2)$. Moreover, Ξ_{ad} stands for the set of all admissible domains. The functional $J\colon \Xi_{\text{ad}} \times U \to \mathbb{R}$, is the *shape functional*. The equation $E(\Omega, y) = 0$ is the *state equation*. Note the resemblance of problem (3.1) with problem (2.4), which suggests shape optimization problems are nothing else than optimal control problems whose unknowns are the underlying geometries. Despite this connection it needs to be emphasized that, unlike usual optimal control problems, the set of admissible domains Ξ_{ad} does not have any linear nor convex structure. Therefore, common techniques from calculus of variations can not be directly applied, or it is not as straightforward as one could desire.

The techniques used to analyze these problems depend on two aspects:

– the way we represent shapes, and

– the way we perturb shapes.

These choices must be related, but many different methods can be derived depending on them.

3.1 Continuous Shape Representations

Up until now, there is no consensus on an a priori choice for shape representations. However, one thing is clear; whichever representation is chosen, it should be a compromise between being flexible enough to allow mechanical representations of the underlying shapes and allowing their explicit deformation. By mechanical representation, we mean it should allow the use of, for example, the finite element method or finite differences to approximate the state equation.

In what follows, we provide a list of possible shape representations in the continuous case, although not exhaustive, demonstrates their versatility. This overview is an extension of the one presented in Fuchs et al., 2009.

Direction or curvature functions: In the seminal work of Klassen et al., 2004, they assume curves are parameterized by arclength with period 2π, and the shape representation can be given either by a direction or curvature function. The authors describe the direction function as the angle of the tangent vector to the curve. In these cases the shape space is a subspace of the periodic L^2-functions on $[0, 2\pi]$.

Polar coordinates of the tangent vectors: This representation was proposed in Mio, Srivastava, Joshi, 2006. The main idea is to represent the velocity vector of a curve in terms of two time-dependent variables. A quantifier of the rate at which an interval I was stretched or compressed and the angle describing how the interval I was bent to form a curve. Briefly, the curves are represented by the polar coordinates of their tangent vectors.

Smooth Embeddings of the unit circle: The work of Michor, Mumford, 2005 proposes considering shapes as compact, simply connected regions in the plane whose boundary is a simple closed curve. Moreover, they assume certain degrees of smoothness over the boundary of the shape. The proposed shape representation is then given by a $\mathcal{C}^\infty$-embedding of the unit circle in the plane, up to reparametrizations. Usually this space is denoted by $B_e := \mathrm{Emb}(S^1, \mathbb{R}^2)/\mathrm{Diff}(S^1)$, where S^1 is the unit circle, $\mathrm{Emb}(S^1, \mathbb{R}^2)$ denotes the set of all embeddings from S^1 into $\mathbb{R}^2$; and $\mathrm{Diff}(S^1)$ is the set of all difeomorphisms between S^1 and itself.

Characteristic functions of measurable sets: This approach was proposed in Zolésio, 2007; besides considering the shapes as characteristic functions of measurable sets, the author defines the so-called tubes, which in a nutshell, are paths between two shapes. These tubes are associated with time-dependent vector fields, which prescribe the deformation of the measurable set in a weak Eulerian sense.

Phase fields: Representing shapes as phase fields was considered as a relaxation of the material distribution problem in topology optimization. The main idea of this representation is to consider an interpolated material density, which naturally provides geometric information. One could expect that the super-level set of the material density contains the level-set of the unrelaxed material property, see e. g., Burger, Stainko, 2006 for more details. We also refer the reader to the articles Garcke et al., 2018 where a phase-field representation is considered for solving shape optimization problems.

Level Sets: This method represents the evolution of the shapes as level sets of continuous functions. The motion can be formulated as a Hamilton–Jacobi equation for the function defining the level set. The main advantage of this method is that it can

handle topological changes without extra effort. The level set method has initially been introduced in Osher, Sethian, 1988. However, we refer the reader to Burger, Osher, 2005 for a survey with special emphasis on inverse problems and shape optimization. Many studies in shape optimization are conducted using level sets, we refer for example to Laurain, Sturm, 2016; Sturm, 2016; Allaire, Jouve, Toader, 2004.

Deformations of a reference domain: Shapes can also be represented by considering all possible deformations of a reference shape. In this case, the optimization process is posed over the set of such transformations. Usually, these transformations are associated with a vector field, which needs to be smooth enough to guarantee the preservation of the topology or the regularity of the resulting (transformed) shapes. In this way, the shape optimization problem is reformulated as an optimal control problem, where we look for the optimal transformation of the reference domain. This approach is also known as the method of mappings and can be traced back to Murat, Simon, 1977; Simon, 1980. It has recently regained importance, for example we refer the reader to Iglesias, Sturm, Wechsung, 2018; Onyshkevych, Siebenborn, 2021; Haubner, Siebenborn, Ulbrich, 2021; Deckelnick, Herbert, Hinze, 2021; Paganini, Wechsung, Farrell, 2018; Hiptmair, Paganini, 2015

It is worth mentioning that most of these representations allow us to obtain well-posed problems; for example, by considering the level set method, phase-field approach, method of mappings, among others.

It can be proved, that the spaces which collect the shape representations from embeddings, curvature functions, and polar coordinates of tangent vectors constitute Riemannian manifolds. This fact opened a whole new world of possibilities from the differential geometry perspective, which can be exploited by considering shapes as elements of Riemannian manifolds. A detailed contribution, in this sense, is the construction of different Riemannian metrics according to the requirements of each problem. The view of shape spaces as manifolds was first introduced in shape statistics and shape analysis. The main goal in this context was to measure similarities between shapes, disregarding their rigid body motions transformations. We refer the reader to Younes, 2012 for an accessible overview of different Riemannian metrics for shape manifolds.

In Schulz, 2014, shapes are represented as embeddings of the unit circle into the plane as suggested in Michor, Mumford, 2005, which is an infinite-dimensional manifold. To the author's knowledge, this was the first time a shape optimization problem was considered as a problem on a Riemannian manifold. Moreover, the author established a connection between differentials from the differential geometry perspective and shape calculus. In later works, Schulz, Siebenborn, Welker, 2014; Schulz, Siebenborn, 2016; Schulz, Siebenborn, Welker, 2015b; c; 2016 the authors proposed a different Riemannian metric, called the Steklov-Poincaré metric, to numerically compute shape gradients. Since then, many extensions of this approach, even for shape optimization problems under uncertainties, have been considered; see for example Geiersbach, Loayza, Welker, 2019; Geiersbach, Loayza-Romero, Welker, 2021a.

As already mentioned, a Riemannian perspective on shape optimization allows us to consider measures of similarity or distance between shapes and also helps in establishing well-behaved algorithms. However, they come with additional difficulties; as pointed out in Geiersbach, Loayza-Romero, Welker, 2021b. For example, for the manifold described in Michor, Mumford, 2005, the optimization problem will be posed on an infinite-dimensional manifold. As mentioned in Bauer, Bruveris, Michor, 2014, working in infinite-dimensional manifolds, many difficulties arise, and there are still open questions. For one, most of

the Riemannian metrics defined over these spaces are weak, and hence the gradient is not necessarily defined. Furthermore, the existence and uniqueness of solutions of the geodesic equation are not guaranteed and need to be checked for each metric; in some cases, the exponential map is not well-defined. Another problem involves the fact that distances on an infinite-dimensional Riemannian manifold can be degenerate. In Michor, Mumford, 2005, it is shown that the reparametrization invariant L^2-metric on the infinite-dimensional manifold of smooth planar curves induces a geodesic distance equal to zero. Then, in the same work, a curvature weighted L^2 -metric is employed as a remedy, and it is proved that the vanishing phenomenon does not occur for this metric. Summarizing, working with infinite-dimensional manifolds is very challenging and remains an active area of research. One way to overcome these challenges is considering finite-dimensional manifolds, and particularly geodesically complete Riemannian manifolds, as the one we propose in chapters 4 and 5.

3.2 Existence of Optimal Shapes

Now, we study if the problem under consideration has a solution. Unfortunately, for PDE-constrained shape optimization problems, this is not always true. This section is devoted to describe two examples that corroborate this statement. Furthermore, we briefly enumerate the previously proposed solutions, to fix the ill-posedness of these problems.

The first one, taken from Dapogny, 2013, Sec. 2.1.2, p. 51 or Bucur, Buttazzo, 2005, Sec. 4.2, p. 78, aims to optimize the distribution of two materials within a hold-all domain $D \subset \mathbb{R}^2$. One of the materials is assumed to be thermally conductive, and the other is thermally insensitive. The optimal distribution of the material needs to guarantee that the resulting temperature in D is as close as possible to a given profile denoted by $\bar{y}$ when D is heated. Mathematically, this means we consider the following problem:

$$\text{Minimize} \quad J(\Omega, y) = \int_D |y - \bar{y}|^2 \, \mathrm{d}x \quad \text{w.r.t. } \Omega \subset D, \Omega \in \Xi_{\text{ad}}, \quad \text{s. t. equation (2.1)},$$

where $r \equiv 1$ in (2.1), and $\bar{y}$ is chosen to be small enough. The set of admissible shapes Ξ_{ad} is the set of shapes with Lipschitz boundary which are completely contained in D. The proof of lack of existences of solutions for this problem can be sketched as follows:

- It can be verified that a global minimum of the problem is strictly contained in D, since $\bar{y}$ is assumed to be small enough.
- Now, we proceed by contradiction, i. e., let us assume there exists a global minimum $\overline{\Omega} \subset D$ which is going to be completely contained in D.
- Next, consider a new domain $\widetilde{\Omega} = \overline{\Omega} \cup B_{\varepsilon}^{x_0}$, where $B_{\varepsilon_0}^{x_0}$ stands for the open ball with center x_0 and radius ε_0. The center of the ball satisfies $x_0 \in D \setminus \overline{\Omega}$ and the radius $\varepsilon_0 > 0$, is small enough such that the distance from x_0 to $\overline{\Omega}$ is greater than ε_0. See figure 3.1 for an illustration of such a construction.
- The state equation (2.1) can be solved explicitly over $\widetilde{\Omega}$, and therefore the corresponding value of the shape functional $J(\widetilde{\Omega}, y)$.
- Finally, it holds that the value of the shape functional on $\widetilde{\Omega}$ is strictly lower that the value of the assumed to be solution $\overline{\Omega}$, i. e., $J(\widetilde{\Omega}, y(\widetilde{\Omega})) < J(\overline{\Omega}, y(\overline{\Omega}))$.
- One could repeat this process arbitrarily many times and obtain each time an even lower value of the shape functional. In other words, reaching the value of $\bar{y}$ (small enough) requires that the shape is a collection of infinitesimally small inclusions, which clearly has not a Lipschitz boundary.

Since this thesis is focused on pure shape optimization problems, we present a second counterexample, which was studied in De Gournay, Fehrenbach, Plouraboué, 2014, Sec. 5. Its main aim is to optimize the section of a pipe, which maximizes or minimizes the

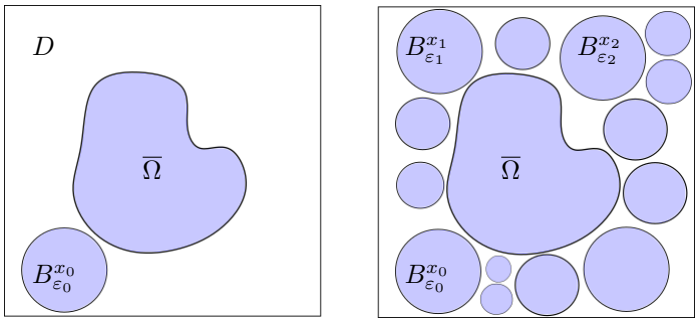


Figure 3.1. Illustration of the construction used for the counterexample of existence of solutions. Left: only one inclusion is considered. Right: multiple inclusions are considered. Clearly a porous structured can be recognized.

characteristic length of heat transport amounts to find the optimal insulating pipe (large characteristic length) or the optimal heat exchanger (small characteristic length). The characteristic lengths are the reciprocal of the Graetz operator's eigenvalues. More precisely, the dominant downstream (resp. upstream) characteristic length is the inverse of the smallest negative (resp. positive) eigenvalue λ_1 (resp. λ_{-1}).

The eigenproblem of the Graetz operator can be described through the following expression:

$$\begin{cases} c\lambda_k^2 T_k + \operatorname{div}(\sigma \nabla T_k) - \lambda_k u T_k &= 0, \quad \text{in } D, \\ T_k &= 0, \quad \text{on } \partial D, \end{cases} \tag{3.2}$$

where T_k denotes the temperature, u is the velocity amplitude, and c, σ are the components of the conductivity matrix. Moreover, the components of the conductivity matrix are defined in terms of the domain Ω, $c = I_\Omega c_1 + (1 - I_\Omega)c_2$ and $\sigma = I_\Omega \sigma_1 + (1 - I_\Omega)\sigma_2$, with $c_i, \sigma_i \in \mathcal{C}^\infty(D)$, I_Ω is the characteristic function of Ω, and the domain satisfies $\Omega \subset D \subset \mathbb{R}^2$. The velocity amplitude is given by $u = \alpha v$, where v solves (2.1) with $r \equiv 1$. For this operator, it can be proved that T_k is a solution of (3.2) if and only if $\phi_k = (T_k, \lambda_k T_k)$ solves $\mathcal{A}\phi_k = \lambda_k \mathcal{B}\phi_k$, with $\mathcal{A}\colon (T, s) \mapsto (-\operatorname{div}(\sigma \nabla T), cs)$ and $\mathcal{B}\colon (T, s) \mapsto (cs - uT, cT)$. To obtain an explicit expression for the eigenvalues of the problem, the authors use the Rayleigh's quotient, i. e.,

$$\lambda_1^{-1} = \max_{\phi \in \mathcal{G}} \frac{(\mathcal{B}\phi, \phi)}{(\mathcal{A}\phi, \phi)}, \quad (\lambda_{-1})^{-1} = \min_{\phi \in \mathcal{G}} \frac{(\mathcal{B}\phi, \phi)}{(\mathcal{A}\phi, \phi)},$$

where $\mathcal{G} \coloneqq H_0^1(\Omega) \times L^2(\Omega)$. Therefore, they aim to maximize or minimize the value of the smallest positive or biggest negative eigenvalue by changing the domain Ω. We recall that the functions c, σ and u from (3.2) depend on the domain Ω.

Moreover, it can be proved that if we do not consider any normalization constraint, the best insulating pipe is empty, and the best conducting pipe is full. Therefore, the authors consider various normalization processes. The counterexample is given for the so-called *prescribed work of the pump* constraint, which implies $\alpha = P|\Omega|^{-1}$ (where P is the work of the pump) in the definition of the velocity amplitude, and $|\Omega|$ stands for the area of the domain Ω. Assuming Ξ_{ad} (the set of admissible shapes) is the set of all domains with $\mathcal{C}^2$-boundary, the sketch of the proof is described in what follows.

– For a given domain Ω with $\mathcal{C}^2$-boundary, it can be proved that there exists a sequence of regular domains Ω_n such that, first, every point in Ω_n is at a distance at most $1/n$ along the vertical direction from a point of the boundary $\partial\Omega_n$. Second, the

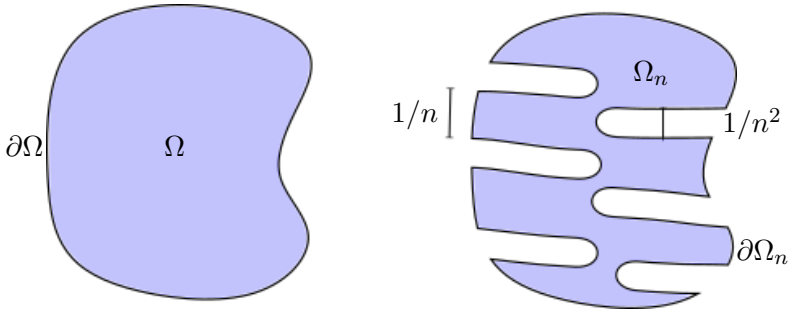


Figure 3.2. The original domain Ω (left), and the domain Ω_n (right), which satisfies that every point in Ω_n is at a distance at most $1/n$ along a vertical line from a point of the boundary. The width of the stripes is $1/n^2$.

characteristic function of Ω_n converges to the characteristic function of Ω strongly in $L^1(D)$. See figure 3.2 for an illustration of such a sequence of domains.

– Consider the eigenvector relative to the first eigenvalue for the steady problem denoted by $\overline{\phi}$ which satisfies

$$(\lambda_1^{\mathrm{st}})^{-1} = \frac{(\mathcal{B}^{\mathrm{st}}\overline{\phi},\overline{\phi})}{(\mathcal{A}(\Omega)\overline{\phi},\overline{\phi})},$$

where $(\mathcal{B}^{\mathrm{st}}\phi,\phi) = \int_D 2c(\Omega)Ts$.

– It can be proved that the following holds: $(\mathcal{B}(\Omega_n)\overline{\phi},\overline{\phi}) \xrightarrow{n\to\infty} (\mathcal{B}^{\mathrm{st}}\overline{\phi},\overline{\phi})$, and $(\mathcal{A}(\Omega_n)\overline{\phi},\overline{\phi}) \xrightarrow{n\to\infty} (\mathcal{A}(\Omega)\overline{\phi},\overline{\phi})$, which implies:

$$(\mathcal{B}(\Omega_n)\overline{\phi},\overline{\phi})/(\mathcal{A}(\Omega_n)\overline{\phi},\overline{\phi}) \xrightarrow{n\to\infty} (\lambda_1^{\mathrm{st}})^{-1}.$$

– It follows from here that Ω cannot be optimal for the minimization of λ_1, because $\limsup \lambda_1(\Omega_n) \leq \lambda_1^{\mathrm{st}}$, and the eigenvalues for the steady problem are strictly smaller than the eigenvalues of the original problem in Ω.

– Now we focus on the case λ_{-1}, and consider $\phi_n = (T_n, \lambda_{-1}(\Omega_n)T_n)$ a sequence of eigenvectors such that $(\mathcal{A}(\Omega_n)\phi_n,\phi_n) = 1$. Relying on the boundedness of $\lambda_{-1}(\Omega)$ in $\mathbb{R}$, the boundedness of T_n on $H_0^1(D)$ and the homogenization theory, one can ensure that exist $\overline{\lambda} \in \mathbb{R}$ and $\overline{T}$ such that $-\operatorname{div}(\overline{\sigma}\nabla\overline{T}) = -c(\Omega)(\overline{\lambda})^2\overline{T}$, from where it follows $\overline{T}$ is an eigenvector of the steady problem, and therefore $(\overline{\lambda})^{-1} \geq (\lambda_{-1}^{\mathrm{st}})^{-1} > (\lambda_{-1})^{-1}(\Omega)$. In other words, for n sufficiently large $\lambda_{-1}(\Omega_n) < \lambda_{-1}(\Omega)$.

– Note that Ω_n for a sufficiently large value of n is no longer of class $\mathcal{C}^2$.

Both examples show that the lack of solutions is an immediate consequence of the lack of compactness of the set of admissible shapes. In fact, in a great variety of shape optimization problems their solutions will contain porous or fractal structures. A natural way to overcome this drawback is to enlarge the set of admissible sets and accept solutions with fractal or porous structures. The mathematical background behind this relaxation is given by the homogenization theory, whose principal aim is to reformulate the topology optimization problem as a problem that finds an optimal distribution of a mixture of material and void. We refer the reader to Tartar, 2000; Allaire, 2012 for more information in this direction. Conversely, one can also obtain existence of solutions for a given functional by restricting the class of admissible shapes via the addition of a uniform geometric constraint on the set of admissible shapes. One example is to define the problem on the class of domains which satisfy a uniform cone condition. Briefly, the uniform cone condition guarantees that the boundary of the admissible shapes is uniformly Lipschitz (see Chenais, 1975 for more

details). One could also add a perimeter regularization to the shape functional as suggested in Ambrosio, Buttazzo, 1993. For example, the minimizing sequence of domains used in the counterexample of the Graetz operator does not satisfy the uniform cone condition either the finite perimeter condition.

The previously mentioned approaches for proving the existence of solutions of optimal shapes aim to:

(*i*) Choose a specific set of admissible shapes.
(*ii*) Endow the set of admissible shapes with a specific topology (which guarantees at the same time continuity of the shape functional and compactness of the set of admissible shapes).
(*iii*) Prove that the set of admissible shapes under the chosen topology is compact.
(*iv*) Prove that the shape functional under the chosen topology is at least lower semi continuous.

Once all these properties hold, the existence of solutions follows immediately from theorem 2.1.7. We will show in chapter 6 a different penalization for the shape function, which is merely inspired on the discrete view of shapes. The addition of this penalization will allow us to prove the existence of a solution for the discretized problem.

3.3 Shape Calculus

After discussing the existence of solutions, we are ready to talk about the optimality conditions for shape optimization problems. In chapter 2, we have shown that the first-order optimality conditions for PDE-constrained problems are stated in terms of the derivatives of the objective function with respect to the control, state, and adjoint variables.

In the context of shape optimization, requiring the derivative of the shape functional with respect to a geometric quantity means we have to provide a way to measure the rate of change of the shape function when the shape is subject to small perturbations. To this end, we first describe how to perturb shapes.

In general, a shape perturbation is given in terms of a family of one-to-one mappings $\{F_t\}_{t\in[0,T]}$ such that $F_t\colon \mathbb{R}^2 \to \mathbb{R}^2$, with $F_0 = \mathrm{id}$ and $T > 0$. It follows then,

$$\Omega_t \coloneqq F_t(\Omega) = \{F_t(x) \,|\, x \in \Omega\},$$

where Ω_t is said to be the perturbed shape at time t, with perturbed boundary

$$\Gamma_t \coloneqq F_t(\Gamma) = \{F_t(x) \,|\, x \in \Gamma = \partial\Omega\}.$$

The way we construct the family of mappings $\{F_t\}_{t\in[0,T]}$ will lead to different methods of representing perturbed shapes. The minimum requirements for these families are:

(α) The transformations $F_t(\cdot)$ and $F_t^{-1}(\cdot)$ belong to $\mathcal{C}^k(\mathbb{R}^2, \mathbb{R}^2)$ for all $t \in [0, T]$.
(β) The mappings $t \mapsto F_t(x)$ and $t \mapsto F_t^{-1}(x)$ belong to $\mathcal{C}^1([0, T])$ for all $x \in \mathbb{R}^2$.

One possibility is to consider only small deformations about a shape Ω, called *perturbations of identity*.

Definition 3.3.1. *Let us consider $\Omega \subset \mathbb{R}^2$, and $T > 0$. The family of mappings $\{F_t\}_{t\in[0,T]}$, associated to the vector field $V \in W^{k,\infty}(\mathbb{R}^2, \mathbb{R}^2)$ or $V \in \mathcal{C}^k(\mathbb{R}^2, \mathbb{R}^2)$ which define the perturbation of the identity method is given by:*

$$F_t(x) = x + t\,V = (\mathrm{id} + t\,V)(x),$$

for all $x \in \Omega$, and where $\mathrm{id}\colon \mathbb{R}^2 \to \mathbb{R}^2$ is the identity map.

Notice that we parameterize perturbed shapes in terms of the vector field V and the time $t \in [0, T]$. Another essential property of this method is that, thanks to the regularity of the vector field V, we can guarantee the topology of the shapes will be preserved for all

time $t \in [0,T]$, see e. g., Dapogny, 2013, Sec. 2.2.1, p. 53. For completeness, we also refer the reader to the *velocity method*, whose description can be found in Sokołowski, Zolésio, 1992, Ch. 2, p. 49.

Now, we can introduce the notion of Eulerian semi-derivative, and shape differentiability.

Definition 3.3.2. *Let $\Omega \subset \mathbb{R}^2$, and Ω_t be the domain obtained by the perturbation of identity with $t \in [0,T]$. Furthermore, y_t is the solution of the state equation on the domain Ω_t. Then, the Eulerian semi-derivative of J at Ω, in the direction V is defined as follows:*

$$\mathrm{d}J(\Omega;V) = \lim_{t\to 0} \frac{J(\Omega_t, y_t) - J(\Omega, y)}{t}. \tag{3.3}$$

Moreover, the shape functional J is said to be shape differentiable at Ω w.r.t. V if the Eulerian semi-derivative defined in (3.3) exists at Ω for all $V \in W^{1,\infty}(\mathbb{R}^2,\mathbb{R}^2)$ and the mapping $V \mapsto \mathrm{d}J(\Omega;V)$ is linear and continuous. In this case, we refer to $\mathrm{d}J(\Omega;V)$ as the shape derivative.

Shape differentiability of PDE-constrained problems, like the one given in (3.1), involves the computation of the derivative of the state and adjoint equations with respect to the domain. Thus, proving shape differentiability is not straightforward, and various techniques have been developed over the last years to do it efficiently. Just to mention some of them: Lagrangian Sturm, 2015, min-max Delfour, Zolésio, 2001, chain rule Sokołowski, Zolésio, 1992, variational Ito, Kunisch, Peichl, 2008.

The shape derivative can be represented in two equivalent formulations. The boundary or strong formulation is given by Hadamard's structure theorem, see Delfour, Zolésio, 2011, Thm. 9.3.6 and Cor. 9.1, p. 479-480. This theorem states that if Γ is smooth enough, then the shape derivative $\mathrm{d}J(\Omega;\cdot)$ admits a representative $g(\Gamma) \in \mathcal{D}^k(\Gamma)$ (a scalar distribution), such that:

$$\mathrm{d}J(\Omega;V) = \langle g(\Gamma)\,, V\cdot n|_\Gamma \rangle_{\mathcal{D}^{-k}(\Gamma)\times\mathcal{D}^k(\Gamma)},$$

where $V\cdot n|_\Gamma$ is the normal component of V restricted to the boundary Γ. In a few words, this implies the shape functional J is insensitive to perturbations of the domain Ω which do not affect its boundary. However, this method has a significant disadvantage; it assumes high regularity of the boundary Γ, which in practice is not always true. For example, when we consider discretized or polygonal shapes, Γ is only piecewise smooth, and thus, Hadamard's structure theorem does not hold anymore.

The second equivalent formulation is called volume or weak formulation (referring to the weak formulation of a PDE). In its beginnings, it was obtained only as an intermediate step along with the computation of boundary formulations (cf., Berggren, 2010). In the last years special attention has been given to the differences between these two formulations. For example, their order of convergence on a finite element setting, the extra work required for the computation of the strong formulation, and of course, the weaker regularity requirements on the domains for the volume formulation. We refer the reader to Hiptmair, Paganini, Sargheini, 2015; Hardesty, Kouri, et al., 2020 for a comparison in this direction.

3.4 Discretization of Shape Optimization Problems

To recap, the optimize-then-discretize approach to solve PDE-constrained shape optimization problems usually proceeds along the following lines. First, one derives an expression for the *shape derivative* of the shape functional w.r.t. vector fields which describe the perturbation of the current domain Ω, as described in section 3.3. Second, the shape derivative, which represents a linear functional on the perturbation vector fields, needs to be converted into a vector field itself, often, but not necessarily, referred to as the *shape gradient*, and here it will be denoted as $\operatorname{grad} J(\Omega)$. The computation of the shape gradient can be achieved, for

example, by evaluating the Riesz representative of the derivative w.r.t. an inner product. The latter is often chosen as the bilinear form associated with the Laplace-Beltrami operator on $\partial\Omega$, or with the linear elasticity (Lamé) system on Ω, see e.g. Schmidt, Schulz, et al., 2011; Schulz, Siebenborn, 2016; Schmidt, Schulz, 2009; 2010. More sophisticated techniques include quasi-Newton or Hessian-based inner products; see Eppler, Harbrecht, 2005; Novruzi, Roche, 2000; Schulz, Siebenborn, Welker, 2015a; Schulz, 2014. It is also worth mentioning that it is not necessary to rely on the Riesz representation of the derivative. One could also determine descent directions in the $W^{1,\infty}$-topology directly, as suggested in Deckelnick, Herbert, Hinze, 2021; Müller et al., 2021.

In any case, the obtained perturbation field is then used to update the domain Ω inside a line search method, where the transformed domain

$$\Omega_s = \{x + s \operatorname{grad} J(\Omega)(x) : x \in \Omega\}, \tag{3.4}$$

associated with the step size s is obtained from the perturbation of identity approach. However, this is not the only option. A widely used alternative approach, after discretization, is to parametrize the possible displacements of the boundary nodes only. The movements of the interior nodes then follow as a second step as the result of some possibly nonlinear map in response to the boundary node displacements. As above, the latter can be obtained utilizing either the volume or the boundary expressions of the shape derivative. We refer the reader to Schmidt, Ilic, et al., 2011; 2013; Lozano, 2017; Bobrowski et al., 2017 for examples of this strategy. In any case, we want to highlight that, while the computation of the shape derivative is either based on the continuous or some discrete formulation of problem (3.1), the computation of the shape gradient or respectively a descent direction and the subsequent updating steps will always be carried out in the discrete setting.

It has been observed in many publications that a straightforward discretization approach has one major drawback: it often leads to a degeneracy of the computational mesh. This degeneracy manifests itself in different ways, mainly through degrading cell aspect ratios or even mesh nodes entering neighboring cells. Doğan et al., 2007 for instance, observe that such mesh distortions impair computations and lead to numerical artifacts. We attribute this behavior to a discretization artifact, by which the positions of *all* nodes of a computational mesh have an impact on the discrete solution of the PDE present in the problem. This presents optimization routines with an opportunity to shift the mesh nodes so that the discrete solution of the PDE exhibits features that allow further descent in the shape functional but at the expense of mesh quality and solution accuracy of the PDE. Notice that this issue does not arise in the continuous setting, where the redistribution of material points in the interior of the domain does not affect the PDE solution and thus on the shape functional.

Unlike usual PDE-constrained optimization problems, the approaches *discretize-then-optimize* and *optimize-then-discretize*, generally, do not commute in shape optimization. This is reflected by obtaining different expressions of the optimality conditions coming from the continuous and discretized problems. Recalling the analysis presented in section 2.2, for general PDE-constrained problems, the equivalence of both approaches holds provided we choose the same discretization method for the state and adjoint variables. If the FEM is the chosen discretization method, this also holds, no matter the order of the polynomials defining the finite element space. However, as highlighted in Berggren, 2010 «... the discretization of the necessary conditions of optimality does not generally lead to the same expression as when deriving the necessary conditions for the discretized problem.» This statement refers specifically to the boundary expression of the shape derivative when the FEM is the chosen discretization method. Based on these two choices, the author provides a way to unify the sensitivity analysis for shape optimization problems from a continuous or

discrete perspective. Moreover, the author specifies when the discretization of the boundary expression of the continuous shape derivative coincides with the boundary expression of the shape derivative from the discretized problem. In particular, this holds when the state equation and the vector field where we evaluate the shape derivative are discretized using piecewise linear finite elements, and it also mentioned that for higher-order finite elements, this is no longer true.

Moreover, we also want to highlight the conclusion provided in Glowinski, He, 1998, Sec. 2.2, p. 156, where the authors mention that, not only are there cases when the optimality conditions do not coincide, but even worse, it may happen that solving a problem under the optimize-then-discretize paradigm with the steepest descent method does not converge. The authors consider the following optimal cooling problem, where the domains are parameterized by a function v as follows:

$$\text{Minimize } J(\Omega, y) = \int_\Omega |y|^2 \, \mathrm{d}x \quad \text{w.r.t. } \Omega \in \Xi_{\text{ad}} \quad \text{s.t. } \begin{cases} -\Delta y &= C \text{ in } \Omega, \\ y &= 0 \text{ on } \Gamma_0, \\ \left|\frac{\partial y}{\partial n}\right| &= 0 \text{ on } \Gamma \setminus \Gamma_0, \end{cases}$$

where $C > 0$, and $\Gamma_0 = \{(x_1, x_2) \,|\, x_2 = v(x_1),\, x_1 \in (0,1)\}$. Moreover, $\Omega \in \Xi_{\text{ad}}$ if and only if it can be expressed as $\{(x_1, x_2) \,|\, x_1 \in (0,1),\, x_2 \in (0, v(x_1))\}$, with

$$v \in \left\{ v \in H_0^1(0,1) \,\middle|\, 0 \le \alpha \le v(x_1) \le \beta, \int_0^1 v(x_1)\, \mathrm{d}x_1 = M, \left|\frac{\mathrm{d}v}{\mathrm{d}x_1}\right| \le c, \text{ on } (0,1) \right\}.$$

By mapping back the domain Ω to the unit square $[0,1]^2$, the problem can be transformed into an identification problem with the design parameter v appearing in the coefficients of the elliptic operator. Using a gradient method for the optimize-then-discretize approach on this problem resulted in the nonconvergence of the algorithm. While using a discretize-then-optimize approach together with a gradient method led to obtaining a solution of the problem with good convergence properties. The authors also comment on the possible reasons for this behavior, which we summarize in what follows:

- After the domain transformation, this model (state equation) is an elliptic equation with variable coefficients.
- It is well-known that the solutions of this kind of models enjoy smoothness properties and sometimes even compactness, which also implies the compactness of the shape functional.
- Unfortunately, the smoothness properties make the related inverse problems hard to solve since large variations on the design variables may have little impact on the solutions and the shape functionals.
- Therefore, the derivatives of these functions can be easily polluted by rounding or truncation errors.

This behavior is usually not observed when the state equation is more complex, because in this case, minor variations in the parameters induce significant variations on the solutions. This also implies that away from the solution, the gradient of the shape functional is large, thus they are less sensitive to rounding and truncation errors.

Both mesh degeneracy and lack of convergence for certain problems under the optimize-then-discretize approach are our main motivations to develop a suitable framework to solve shape optimization problems under the discretize-then-optimize approach.

3.5 Techniques to Preserve Mesh Quality

As already mentioned, one main issue in the numerical solution of PDE-constrained shape optimization problems is how to preserve the quality of the underlying meshes. Over

the past 10 years, a range of various techniques has been proposed to circumvent this significant obstacle. One could classify these techniques into two main groups. The first group aims to correct the errors resulting after updating the mesh. In Etling et al., 2020, the authors refer to them as post-processing techniques. The second one generates descent directions that correct specific behavior known to decrease the quality of the mesh. In what follows, we describe some of these techniques.

Post-processing techniques: Let us assume we have a triangulation Ω_h from Ω, and we have computed the discrete shape gradient $\operatorname{grad} J(\Omega_h)$. Then, we proceed to update the shape Ω_h as suggested in (3.4) for a certain step size s. Now, we check the quality of the resulting mesh $\Omega_{h,s}$ and we realize its quality is not optimal, the following techniques can be used to correct this behavior.

- *Remeshing:* it can be considered the most natural choice, see for instance Wilke, Kok, Groenwold, 2005; Morin et al., 2012; Sturm, 2016; Dokken et al., 2018; Feppon et al., 2018. Remeshing can be carried out either in every iteration or whenever some measure of mesh quality falls below a certain threshold. Drawbacks of remeshing include the high computational cost and the discontinuity introduced into the history of the shape functional.
- *Mesh regularization:* is a redistribution of nodes, with the only goal to keep all angles on element stars of the same size. For example, they impose a geometric restriction that limits the tangential motion of the nodes, see Bänsch, Morin, Nochetto, 2005; Doğan et al., 2007.
- *Space adaptivity:* keeps an accurate representation of the boundary of the domain by refining/coarsening the mesh as required.
- *Goal-oriented mesh refinement:* Giacomini, Pantz, Trabelsi, 2017 addressed the issue of discretization errors in the underlying PDE model and use them to compute an aposteriori error for the shape derivative. This error is then used to develop a certified algorithm to find a descent direction (shape gradient) in the discrete setting, which is also a descent direction for the continuous problem. If for some iteration, the computed descent direction is not a genuine descent direction for the original problem, then a refinement of the entire mesh is performed.
- *Angle width control:* the main idea of this technique is splitting the elements whose angles are wider than a certain threshold, see e. g., Doğan et al., 2007.
- *Geometric line search:* in Morin et al., 2012 the authors consider a line search method that aims to avoid mesh distortion due to tangential movements of the boundary nodes, combined with a geometrically consistent mesh modification (GCMM) proposed in Bonito, Nochetto, Pauletti, 2010.
- *Geometrically consistent mesh modification:* ensures the health of the mesh by computing the position of the new nodes as the solutions of the geometric identity $-\Delta_\Gamma X = H$, where H is an approximation of the curvature of Γ. See Morin et al., 2012; Bonito, Nochetto, Pauletti, 2010 for more information.
- *Mesh smoothing based on Voronoi reparametrizations:* this technique is related to the geometrically consistent mesh modification. The authors define a mesh smoothing technique based on centroidal Voronoi reparametrizations and construct tangent deformation fields and correct the degenerate cell of the mesh, as described in Schmidt, 2014.
- *Time adaptivity:* mainly developed to be applied within the velocity method. It allows large time steps when the normal of the velocity field does not exhibit large variations and force small step sizes otherwise, cf., Doğan et al., 2007.

- *Overlapping meshes:* proposed in Dokken et al., 2019 aims to represent the computational domain by multiple, independent meshes. The authors use a Nitsche-based finite element method to enforce the continuity over the non-matching mesh interfaces in a weak sense.

Improved descent directions: The main difference with the previously mentioned techniques is that in this case, for a given triangulation Ω_h, we compute the shape gradient grad $J(\Omega_h)$ aiming to generate improved descent directions.

- *Nearly conformal transformations:* the aim is to enforce shape gradients to be generated from nearly conformal transformations. It is known that this kind of transformation preserves angles and ensures a good quality of the mesh along the optimization process. This approach was proposed in Iglesias, Sturm, Wechsung, 2018.
- *Restricted mesh deformations:* motivated by a discrete counterpart of the Hadamard's structure theorem, the idea of restricting the space where we look for descent directions was described in Etling et al., 2020.
- *Pre-shape calculus:* under the recently proposed paradigm of pre-shape calculus, the authors of Luft, Schulz, 2021a; b, propose to add certain regularization terms to the shape functional based on the so-called pre-shape parameterization tracking problem.
- *Restricted method of mappings:* impose certain restrictions on the maps which aim to preserve mesh quality. Of course, this technique is used together with the method of mappings. To cite some examples, we refer the reader to Onyshkevych, Siebenborn, 2021 where a nonlinear extension operator is considered, or Haubner, Siebenborn, Ulbrich, 2021 where a continuous extension operator is considered, chosen specifically to meet the regularity requirements of the mappings. Namely, the authors consider extension operators based on the Laplace-Beltrami equation, elliptic equation, or vector-valued elliptic equations.
- *Linear elasticity without interior contributions:* First proposed in Schulz, Siebenborn, Welker, 2015b the authors neglect the contribution of the shape derivative associated with the interior nodes coming from the volume expression of the shape derivative to avoid choosing directions of negative curvature. Despite the aim of the authors was different, the so-called *strip method*, proposed in Hardesty, Antil, et al., 2020 can also be classified within this category.
- *Weighted linear elasticity:* In Schulz, Siebenborn, 2016, the authors propose to set the Lamé parameter μ, associated with the bilinear form from the linear elasticity, as the solution of the Poisson's equation for specific boundary conditions. A minimum and maximum values of the parameter are set as boundary conditions, where the maximum value of μ is associated with the moving boundary, and the minimum value is assigned to the nonmoving boundary.

4 Discrete Shape Manifolds

Contents

As already mentioned in chapter 3, working with Riemannian manifolds for shape optimization has great advantages. For this reason we focus on shape manifolds of finite dimension for the rest of this thesis. A brief overview of the main notions on differential geometry is provided in appendix A. Considering shapes as elements of a manifold has been already exploited in the context of shape optimization, see e. g., Schulz, 2014; Schulz, Siebenborn, Welker, 2014; Schulz, Siebenborn, Welker, 2016. In this case, the authors consider shapes as elements of the infinite-dimensional manifold B_e, proposed in Michor, Mumford, 2007, and endow it with the so-called Steklov-Poincaré metric. Unfortunately, following this approach has certain gaps in the theoretical results when passing to the discrete problem, as the one remarked in Geiersbach, Loayza-Romero, Welker, 2021b, p. 365, where the regularity of the shape gradient obtained from the Steklov-Poincaré metric differs from the expected regularity of the tangent space of B_e. To overcome this drawback, we propose to study shape optimization problems posed on finite-dimensional Riemannian manifolds. This chapter is devoted to studying discrete shape manifolds, focusing on the proposal of the manifold of planar triangular meshes. In a few words, this manifold describes all the possible configurations of node positions that an admissible mesh of a certain connectivity can attain.

The chapter is organized as follows. Section 4.1 aims to show the state-of-the-art of discrete shape manifolds. We can also understand this overview as the discrete counterpart of the shape representations discussed in section 3.1, with a particular focus on shape spaces that constitute Riemannian manifolds. For the formal definition of the manifold of triangular meshes, we use the language of simplicial complexes. A summary of the required definitions and results about abstract and geometric simplicial complexes is presented in section 4.2. We end this chapter with the detailed construction of the manifold of planar triangular meshes and the proof of its main properties. This chapter is based on the submitted paper Herzog, Loayza-Romero, 2020.

4.1 Overview of Discrete Shape Manifolds

As in the continuous case, there is also no consensus on the discrete representation of shapes. There are many different ways of represent a shape with a finite number of parameters. In what follows, we describe some of them with special focus on triangular meshes.

Landmarks: Shapes can be represented by a finite number of salient points, also called landmarks, up to transformations which leave the shapes unchanged, like rigid rotations and translations and nonrigid uniform scaling. This space was introduced in Kendall, 1984 and proved that it is indeed a manifold. Advanced statistical analysis was performed on this manifold with the help of the Procrustean metric.

Immersions of simplicial complexes in $\mathbb{R}^3$: In Kilian, Mitra, Pottmann, 2007 the authors consider the space of all immersions in $\mathbb{R}^3$ for a fixed simplicial complex. These immersions are represented by one large vector concatenating the vertices positions (elements of $\mathbb{R}^3$) of the complex. Their main contribution is the proposal of two Riemannian metrics on this space based on isometric and rigid deformations. In fact, the conceptual definition of meshes is the same as the one presented in this thesis. However, the authors do not make any specific attempts to ensure that the obtained shapes are free from self-intersections. In fact, they mention that for widely varying poses, such intersections may be obtained.

Discrete thin shells: This space is formed by triangular meshes of fixed connectivity, representing a surface embedded into $\mathbb{R}^3$, assuming they are made of an (ideally infinitesimally) thin material like metal or paper, see Heeren et al., 2012. They propose a Riemannian metric for this shape space which considers viscous dissipation in terms of membrane and bending energies to reflect physical behavior.

Quad-meshes as nonlinear constraints: The authors in Yang, Chang, Chen, 2011 provide a framework to characterize quad meshes implicitly, prescribed by a collection of nonlinear constraints. They also propose ways to explore this shape space using tangent vectors and quadratically parametrized osculant surfaces.

Dihedral angles: In Amenta, Rojas, 2020 the authors consider the embedding of triangular surface meshes into $\mathbb{R}^3$ up to translations, rotations, and scalings, using the vector of dihedral angles.

Discrete exterior derivative: In Liu et al., 2010, the authors propose considering shapes in terms of the discrete exterior derivative, or coboundary operator, of parametrizations over a finite simplicial complex. They also propose constructing shape spaces equipped with Riemannian metrics to measure how costly it is to interpolate two shapes through elastic deformations.

Frölicher spaces: For the sake of completeness, we also mention the approach proposed in Magnot, 2016; 2020. The main idea is to consider the space of all possible triangulations of a given domain (with different connectivity information) as a Frölicher space. Briefly, in a Frölicher space one replaces the atlas of a classical manifold with other intrinsic objects, which enable to define smoothness of mappings safely. Under this framework, the author proved the smooth dependence on the set of (possibly refined) triangulations for the Dirichlet problem discretized using the finite element method of piecewise linear elements.

4.2 Fundamentals on Simplicial Complexes

As already mentioned, we consider discretized shapes as triangulations. The description of triangular meshes is done via two pieces of information: the position of its nodes and the connectivity information. This section aims to provide the appropriate language to formally establish when a given connectivity information and the position of the nodes render an admissible mesh. In particular, we work with the notions of simplicial complexes, known to be one of the essential concepts in algebraic topology. Their definitions can be found in any basic algebraic topology book. We use the books Edelsbrunner, Harer, 2010; Munkres, 2018, the monograph Misztal, 2010 and the journal paper Horak, Jost, 2013.

4.2.1 Geometric Simplicial Complexes

We will denote by X a finite-dimensional vector space. A simplex σ of dimension $k \in \mathbb{N}_0$ (or **k-simplex**) in X is the convex hull of $k+1$ affine independent points in X. A **face** of **dimension** m $(0 \leq m \leq k)$ (an **m-face**) of σ is the convex hull of a subset of $m+1$ of its

vertices. **Vertices** are 0-faces, **edges** are 1-faces and **triangles** are 2-faces of a simplex. In figure 4.1 we depict examples of simplices of dimension zero to three. We will use the terms vertices, edges, and triangles to refer to the 0, 1, 2-faces of a simplex. We will use the term face of a simplex only when the dimension is not specified.

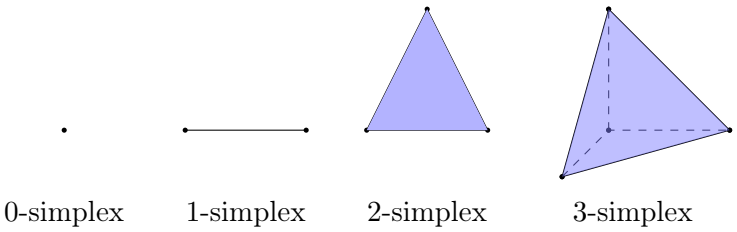


Figure 4.1. Examples of k-simplices.

Now, we present the definition geometric simplicial complex.

Definition 4.2.1. *A (finite)* ***geometric simplicial complex*** Σ *in* X *is a nonempty, finite set of simplices in* X *satisfying:*

(*i*) *Every face of a simplex* $\sigma \in \Sigma$ *also belongs to* Σ.

(*ii*) *The nonempty intersection of any two simplices* σ, σ' *in* Σ *is a face of both* σ *and* σ'.

Figure 4.2 depicts an example of a simplicial complex (blue) and a collection of simplices that do not form a simplicial complex (green).

We say that a geometric simplicial complex Σ is of **dimension** $k \in \mathbb{N}_0$ (or a **geometric simplicial** k**-complex**) if k is the largest dimension of any simplex in Σ.

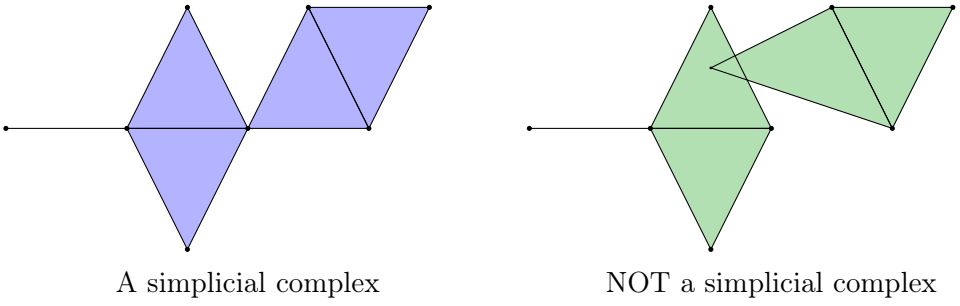


Figure 4.2. Examples of simplicial complex and not.

The **vertex set** of a geometric simplicial complex Σ is the union of the vertices of all of its faces.

We consider a special kind of geometric simplicial complexes which satisfy two additional properties. A geometric simplicial k-complex Σ is **pure** if all maximal elements of Σ (w.r.t. the partial order of set inclusion) have dimension k. Briefly, a geometric simplicial k-complex is pure if and only if every simplex in Σ is the face of some k-simplex in Σ. See figure 4.3 for an example of a simplicial complex which is pure (blue) and one which is not (green).

Furthermore, we study simplicial complexes which enjoy certain stiffness properties encoded in the so-called path connectedness. A geometric simplicial k-complex Σ in X is said to be m**-path connected** ($1 \leq m \leq k$) if for any two distinct m-faces σ, σ' in Σ, there exists a finite sequence of m-faces, starting in $\sigma_0 = \sigma$ and ending in $\sigma_n = \sigma'$, such that $\sigma_i \cap \sigma_{i+1}$ is

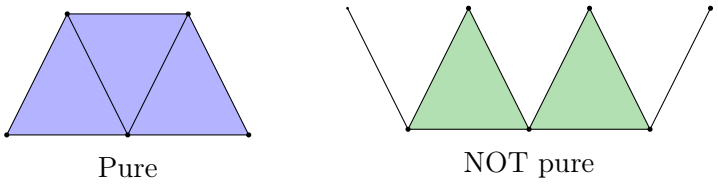


Figure 4.3. Examples of a pure simplicial complex and of one which is not pure.

an $(m-1)$-face for $i = 0, \ldots, n-1$. In figure 4.4 we show examples of a simplicial complex which is 2-path connected (blue) and which is not (green).

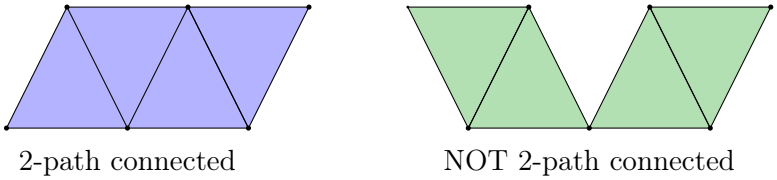


Figure 4.4. Examples of a simplicial complex which is 2-path connected and of one which is not 2-path connected.

Now, we introduce the definitions of the star, closed star, and link of a face $\sigma \in \Sigma$, which give local information about the simplicial complex. Let us consider a geometric simplicial complex Σ, denoting σ as one of its faces (of any dimension). The **star** of σ, is the union of the interior of those simplices of Σ which have σ as a face. We denote the star of σ as $\mathrm{St}(\sigma)$. For a given subset of faces of Σ, its **closure** is defined as the smallest subcomplex of Σ containing those faces. The closure of the star of σ, denoted by $\overline{\mathrm{St}(\sigma)}$, is called the **closed star**, and it can be understood as the union of all simplices of Σ having σ as a face, and it is a subcomplex of Σ. The **link** of σ in Σ is the set $\overline{\mathrm{St}(\sigma)} \setminus \mathrm{St}(\sigma)$ and it is denoted by $\mathrm{lk}(\sigma)$. Concisely, the link of σ is the set of all simplices in $\overline{\mathrm{St}(\sigma)}$ which do not have σ as a face. Figure 4.5 shows examples of the star, closed star, and link of a face σ, when σ is a vertex or an edge. It can also be proved that for any face σ of Σ, $\mathrm{lk}(\sigma)$ is a subcomplex of Σ. If Σ is pure and of dimension d, then $\overline{\mathrm{St}(\sigma)}$ is also pure and of dimension d. Furthermore, if σ is of dimension k, then $\mathrm{lk}(\sigma)$ is pure and of dimension $d-k-1$ (cf., Gallier, 2008, p. 100).

We clarify the distinction between boundary faces and interior faces when $X = \mathbb{R}^2$. Let Σ be a geometric simplicial 2-complex. We say that an edge $\sigma \in \Sigma$ is a **boundary edge** if it belongs to precisely one triangle. We denote the set of all boundary edges by E_∂. A triangle is said to be a **boundary triangle** if it contains at least one boundary edge. T_∂ will denote the set of all boundary triangles. Finally, a vertex is called a **boundary vertex** if it belongs to at least one boundary edge. We denote the set of all boundary vertices by V_∂.

On the other hand, an edge is said to be an **interior edge** if it belongs to exactly two triangles. A triangle is said to be an **interior triangle** if all of its edges are interior. Finally, a vertex is said to be an **interior vertex** if all edges it belongs to are interior. Notice that in a simplicial 2-complex, all vertices, edges or triangles are either boundary or interior.

Another concept that will be become handy in the following sections is the notion of polygonal chains. A **polygonal chain** is a connected series of edges. It can also be understood as a curve specified by a finite sequence of points called the vertices of the chain. Furthermore, a **simple polygonal chain** is such that only consecutive segments

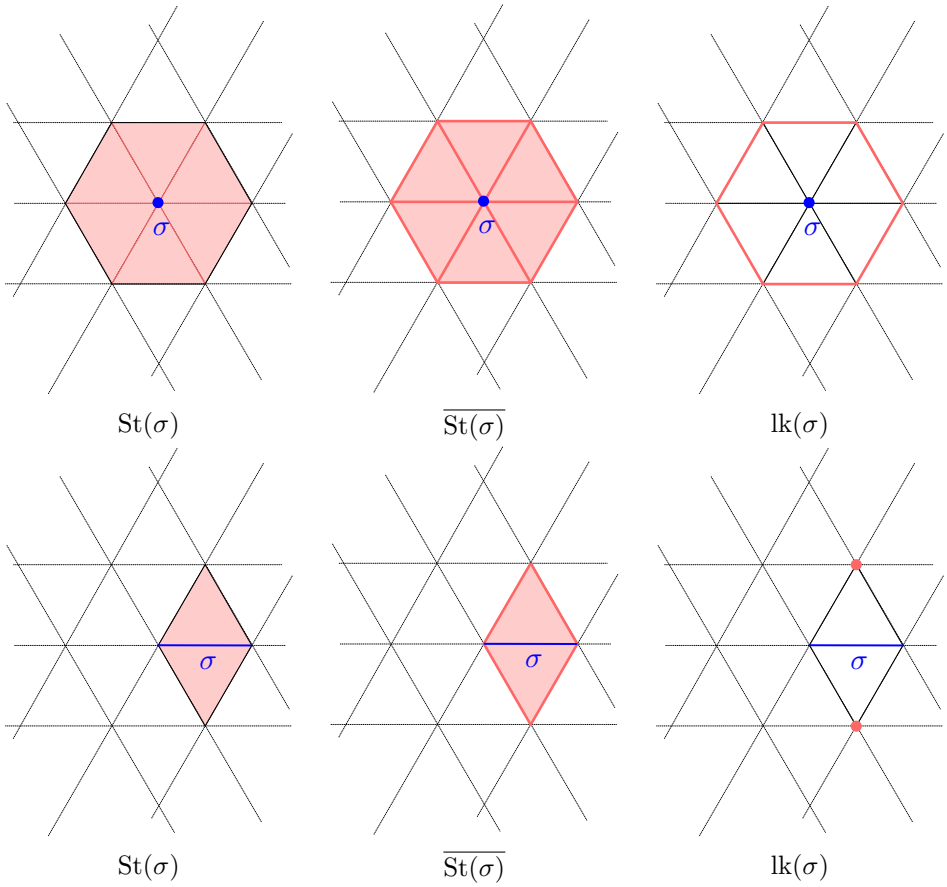
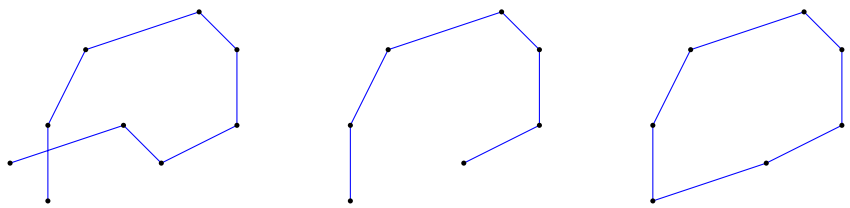


Figure 4.5. Examples of star, closed star and link for a vertex (top row) and an edge (bottom row) of a simplicial complex.



Polygonal chain | Simple polygonal chain | Closed and simple polygonal chain

Figure 4.6. Types of polygonal chains.

intersect, and they intersect only at their endpoints. In the same way, a **closed polygonal chain** is such that the first vertex coincides with the last one. See figure 4.6 for a illustration of these concepts.

The notions of link, and closed star of a face, together with the polygonal chains, allow us to present the following lemmas, which will be used in the next chapter. The main idea of these results is to characterize the link of interior vertices.

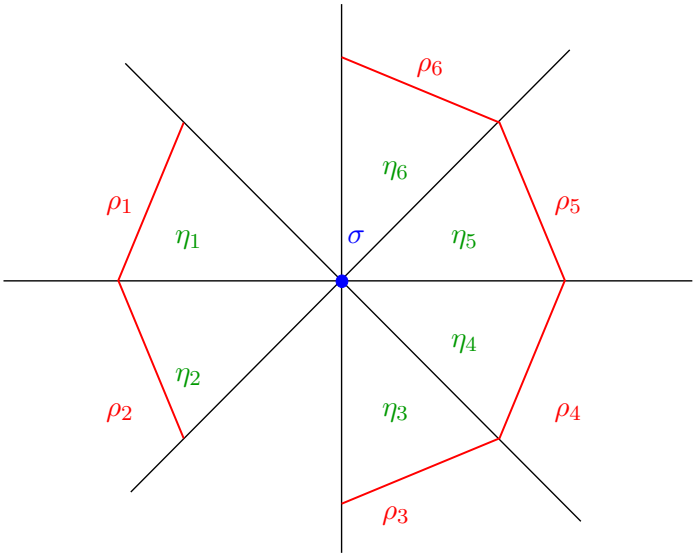


Figure 4.7. Illustration of the link (red) of a vertex (blue) which belongs to a geometric simplicial 2-complex which is not 2-path connected used in the contradiction argument of the proof of lemma 4.2.2.

Lemma 4.2.2. *Let Σ be a pure, 2-path connected, geometric simplicial 2-complex. If σ is a vertex of Σ, then $\mathrm{lk}(\sigma)$ is a simple polygonal chain.*

Proof. Since Σ is pure and has dimension two, we know the link of any face of Σ, is a subcomplex, which is pure, and its dimension is given by $d-k-1$, where d is the dimension of the simplicial complex, and k the dimension of the face. By direct application of this result when $d=2$ and $k=0$, we obtain that $\mathrm{lk}(\sigma)$ is a pure geometric simplicial complex of dimension one. In other words, it is a collection of edges which we denote by ρ_i. Moreover, these edges satisfy $\rho_i \cap \rho_j$ is either the empty set or a common vertex, when $i \neq j$. By direct comparison with the notion of polygonal chain, it remains to prove that this sequence of edges is connected. To this end, we proceed by contradiction and assume that the sequence of line segments is not connected (see figure 4.7 for an illustration of such a case).

This implies there exists at least two connected components of the set of edges. From the definition of $\mathrm{lk}(\sigma)$ we know that for every edge ρ_i in $\mathrm{lk}(\sigma)$, there exists a triangle η_i in $\overline{\mathrm{St}(\sigma)}$ such that $\rho_i \subset \eta_i$. Following the notation from figure 4.7, let us consider the triangle η_2, then it is easy to see that there is no sequence of triangles η_i which will connect η_2 with for example η_3, such that they are intersecting in one edge. This contradicts the assumption of Σ being 2-path connected. □

Lemma 4.2.3. *Let Σ and σ be as in lemma 4.2.2. Additionally, let us assume σ is an interior vertex. Then, the link of σ, $\mathrm{lk}(\sigma)$ is a closed and simple polygonal chain.*

Proof. From lemma 4.2.2, we know that $\mathrm{lk}(\sigma)$ is a simple polygonal chain; therefore, it remains to prove that it is closed. We proceed by contradiction and assume that the first vertex is different from the last. Considering the notation used in the proof of lemma 4.2.2, we denote by ρ_1 the first edge of the sequence, formed by the vertices $\sigma_{1,1}$ and $\sigma_{1,2}$, and w.l.o.g. we assume $\sigma_{1,1}$ is the first vertex of the polygonal chain. Then, by the definition of link, we know there exists a triangle $\eta_1 \in \overline{\mathrm{St}(\sigma)}$, uniquely defined by the vertex σ and the edge ρ_1, or in other words, by the vertices $\sigma, \sigma_{1,1}$ and $\sigma_{1,2}$. Now, we notice the edge

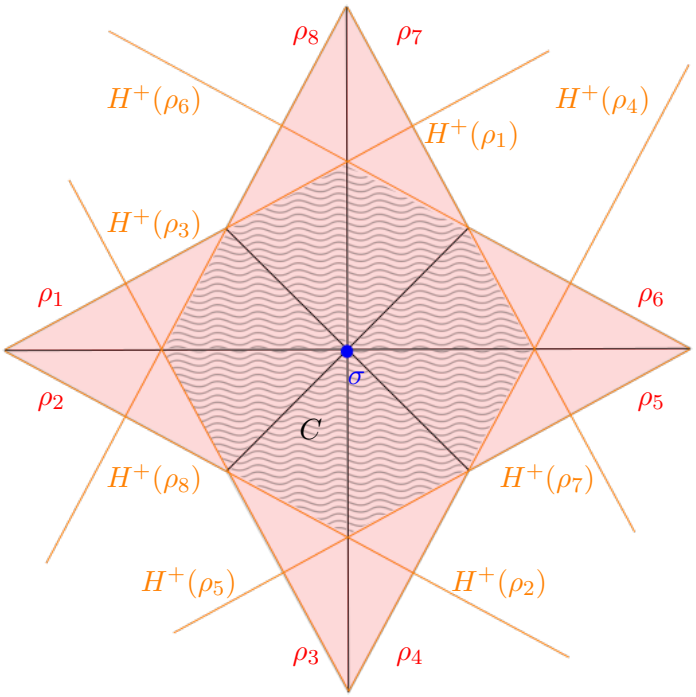


Figure 4.8. Illustration of the closed star (shaded red) of an interior vertex (blue) and the set C which is the intersection of half-spaces (texture) mentioned in lemma 4.2.4.

joining σ and $\sigma_{1,1}$ belongs to only one triangle (namely η_1), which by definition means it is a boundary edge and therefore σ is a boundary vertex, which produces a contradiction. □

This final lemma establishes a connection between the intersection of all half-spaces generated by the hyperplanes coming from the edges of the link of an interior vertex and its closed star.

Lemma 4.2.4. *Let Σ and σ be as in lemma 4.2.2, and σ interior. Let us denote by ρ_i the edges which belong to $\mathrm{lk}(\sigma)$. Finally, let us consider the set*

$$C = \bigcap_{i \in I} H^+(\rho_i),$$

where $I \coloneqq \{i \in \mathbb{N} \,|\, \rho_i \in \mathrm{lk}(\sigma)\}$ and $H^+(\rho_i)$ denotes the half-space generated by the edge ρ_i which contains σ as depicted in figure 4.8. Then, $C \subset \overline{\mathrm{St}(\sigma)}$ holds.

Proof. We proceed by contrapositive, i. e., we assume $x \notin \overline{\mathrm{St}(\sigma)}$ and we wish to prove $x \notin C$. By definition of the closed star, $x \notin \overline{\mathrm{St}(\sigma)}$ implies that $x \notin \eta_i$ for all $\eta_i \in \overline{\mathrm{St}(\sigma)}$. Now, since σ is an interior vertex and thanks to lemma 4.2.3, we know $\mathrm{lk}(\sigma)$ is a simple closed polygonal chain, which means there exists an edge, which we denote by $\rho_j \in \mathrm{lk}(\sigma)$, such that $x \in H^-(\rho_j)$, which implies $x \notin C$. □

4.2.2 Abstract Simplicial Complexes

Together with the definition of geometric simplicial complexes, we consider abstract simplicial complexes. A (finite) **abstract simplicial complex** Δ over a finite set Z, is a nonempty collection of subsets of Z such that, for all $\sigma \in \Delta$, every nonempty subset of σ also belongs to Δ. The elements of $\sigma \in \Delta$ are called the **faces** of Δ. A face σ is said to be of

dimension $m \in \mathbb{N}_0$ (an m-**face**) if $\#\sigma = m + 1$, where $\#\sigma$ stands for the cardinality of σ. We will keep the notation $0, 1, 2, \ldots, m$-faces to distinguish these with the ones associated with the geometric simplicial complexes.

An abstract simplicial complex Δ is said to be of **dimension** $k \in \mathbb{N}_0$ (an **abstract simplicial k-complex**) if k is the largest dimension of any of its faces.

The **vertex set** of an abstract simplicial complex Δ is the union of all of its faces, i. e., $\bigcup_{\sigma\in\Delta} \sigma \subset Z$. The elements of the vertex set are the **vertices**. An abstract simplicial complex over a finite set Z is also an abstract simplicial complex over its vertex set.

We have defined the notions of pureness and 2-path connectedness for geometric simplicial complexes to help us with the visualization. However, these are purely combinatorial properties that will be attributed, from now on, to the abstract simplicial complexes. In what follows, we introduce the kind of abstract simplicial complexes which we will use in what remains of this thesis.

Definition 4.2.5. *Suppose that Δ is an abstract simplicial 2-complex such that its vertex set is given by $\{1, \ldots, N_V\}$. We say that Δ is a* ***connectivity complex****, provided that*

(*i*) *Δ is pure,*
(*ii*) *Δ is 2-path connected.*

In example 4.2.6, we show three abstract simplicial complexes: the first one is a connectivity complex, the second one which does not satisfy item (*i*) and the third one which does not satisfy item (*ii*), from definition 4.2.5.

Example 4.2.6. *The abstract simplicial 2-complex*

$$\Delta = \big\{\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{2,3\},\{3,4\},\{4,1\},\{1,5\},\{2,5\},\\ \{3,5\},\{4,5\},\{1,2,5\},\{2,3,5\},\{3,4,5\},\{4,1,5\}\big\}$$

is pure and 2-path connected. By contrast, the following abstract simplicial 2-complexes violate items (i) and (ii), in definition 4.2.5, respectively:

$$\Delta = \big\{\{1\},\{2\},\{3\},\{4\},\{1,2\},\{2,3\},\{3,1\},\{1,4\},\{1,2,3\}\big\},$$
$$\Delta = \big\{\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{2,3\},\{3,1\},\{3,4\},\{4,5\},\{5,3\},\{1,2,3\},\{3,4,5\}\big\}.$$

Abstract simplicial complexes provide a purely combinatorial description of geometric simplicial complexes, disregarding actual vertex "positions". Now, we explain the way abstract and geometric simplicial complexes are related. Every geometric simplicial complex Σ defines an abstract simplicial complex Δ (unique up to homomorphisms, i. e., renaming vertices) of the same dimension, as follows: suppose that $q_1, q_2, \ldots, q_{N_V}$ are the vertices of Σ. Define Δ over the vertex set $\{1, \ldots, N_V\}$ as:

$$\Delta \coloneqq \big\{\sigma \subset \{1, \ldots, N_V\} \,\big|\, \operatorname{conv}\{q_i\}_{i\in\sigma} \in \Sigma\big\}.$$

We call Δ the **abstract simplicial complex associated** with Σ.

It is also worth mentioning that the notions of **boundary faces** and **interior faces** of dimensions 0, 1 and 2, of an abstract simplicial 2-complex, can be defined in the same way as in the geometric case.

Having established the notions of abstract and geometric simplicial complexes, now we focus on the orientability of them. **Orientation** of a simplicial complex can be established either in the geometric or abstract sense. Let us consider Δ an abstract simplicial 2-complex. An orientation of an abstract 2-face $\sigma = \{i_0, i_1, i_2\} \in \Delta$ is an equivalence class of orderings of the elements of σ, where two orderings i_0, i_1, i_2 and $i_{\pi(0)}$, $i_{\pi(1)}$ and $i_{\pi(2)}$ are equivalent if and only if π is an even permutation of $0, 1, 2$. We denote the equivalence class represented by the ordering i_0, i_1, i_2 by $[i_0, i_1, i_2]$. Each abstract 2-face has precisely two orientations. Moreover, the orientation $[i_0, i_1, i_2]$ of the abstract 2-face $\sigma = \{i_0, i_1, i_2\}$

induces an orientation on each 1-face contained in σ, namely $[i_1, i_2]$, $[i_2, i_0]$, and $[i_0, i_1]$, respectively.

We end this review with the notion of consistent orientation. Let Δ be an abstract simplicial 2-complex. Suppose that each of the 2-faces of Δ has an assigned orientation. We say that these orientations are consistent and that Δ is **consistently oriented** if and only if the orientations of any two 2-faces in Δ sharing a 1-face induce opposite orientations on that 1-face. Finally, we say that Δ is **orientable** if there exist orientations of all 2-faces in Δ which render Δ consistently oriented. Example 4.2.7 shows two instances of abstract simplicial complexes which are or are not orientable.

Example 4.2.7. *Let us consider the following simplicial 2-complexes*

$$\begin{aligned}\Delta_1 = \big\{&\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{2,3\},\{3,4\},\{4,1\},\{1,5\},\{2,5\},\\ &\{3,5\},\{4,5\},\{1,2,5\},\{2,3,5\},\{3,4,5\},\{4,1,5\}\big\},\\ \Delta_2 = \big\{&\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{2,3\},\{3,1\},\{1,4\},\{2,4\},\{1,5\},\{2,5\},\\ &\{1,2,3\},\{1,2,4\},\{1,2,5\}\big\}.\end{aligned}$$

If we endow its 2-faces with the orientations $[1,2,5]$, $[2,3,5]$, $[3,4,5]$ *and* $[4,1,5]$, *then* Δ_1 *is consistently oriented. By contrast,* Δ_2 *is nonorientable.*

4.2.3 Basic Notions about Triangles

This subsection is dedicated to fixing notation and collecting the essential properties and relations between triangles and their geometric measurements.

Let Δ be a connectivity complex as in definition 4.2.5, with vertex set $\{1, \ldots, N_V\}$, and $\{i_0, i_1, i_2\}$ a 2-face of Δ. Moreover let us consider a geometric simplicial 2-complex Σ with $q_1, \ldots, q_{N_V}$ its vertices from $\mathbb{R}^2$. The abstract and geometric simplicial complexes Δ an Σ are given such that Δ is the abstract simplicial complex associated to Σ. Let $\mathrm{conv}\{q_{i_0}, q_{i_1}, q_{i_2}\}$ be a triangle (2-face) of Σ. To simplify notation we collect all the vertex positions of Σ in a matrix $Q \in \mathbb{R}^{2\times N_V}$ and the faces of Σ are denoted by $\mathrm{conv}_Q(i_0, \ldots, i_m)$, when $m > 0$ and $\{q_{i_0}\}$ for $m = 0$.

The **edge lengths** of the triangle are denoted as $E^\ell_Q(i_0, i_1, i_2)$, for $\ell = 0, 1, 2$ and are defined by:

$$E^\ell_Q(i_0, i_1, i_2) \coloneqq \|q_{i_{\ell\oplus 1}} - q_{i_{\ell\oplus 2}}\|_2, \tag{4.1}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, and $\oplus$ the addition modulo 3. The ℓ-th edge is the one opposite the vertex $\{q_{i_\ell}\}$ for $\ell = 0, 1, 2$. The **interior angles** are denoted by $\theta^\ell_Q(i_0, i_1, i_2)$, where the angle numbered ℓ is the one at the vertex $\{q_{i_\ell}\}$ for $\ell = 0, 1, 2$. The ℓ-th **height** is denoted as $h^\ell_Q(i_0, i_1, i_2)$ and it is the length of the line segment perpendicular to the edge $\mathrm{conv}_Q(i_{\ell\oplus 1}, i_{\ell\oplus 2})$ which passes through the vertex $\{q_{i_\ell}\}$. The **inradius** is the radius of the largest circle that fits inside the triangle and we denote it as $r_Q(i_0, i_1, i_2)$. The **circumradius** is the radius of the smallest circle into which the triangle fits, and we denote it as $R_Q(i_0, i_1, i_2)$. Figure 4.9 illustrates all of these notions.

We list the main relations between the geometric measurements of a triangle. The **signed area** of a triangle is given by the expression:

$$A_Q(i_0, i_1, i_2) \coloneqq \frac{1}{2} \det\left[q_{i_1} - q_{i_0},\ q_{i_2} - q_{i_1}\right]. \tag{4.2}$$

Notice that this definition is independent of the particular ordering of vertices representing the orientation since $A_Q(i_0, i_1, i_2) = A_Q(i_1, i_2, i_0) = A_Q(i_2, i_0, i_1)$ holds. Specifically, $A_Q(i_0, i_1, i_2) > 0$ indicates that the vertices $\{q_{i_0}, q_{i_1}, q_{i_2}\}$ are in counterclockwise order. The opposite order leads to a change in the sign. Moreover, $A_Q(i_0, i_1, i_2) \neq 0$ holds if and only

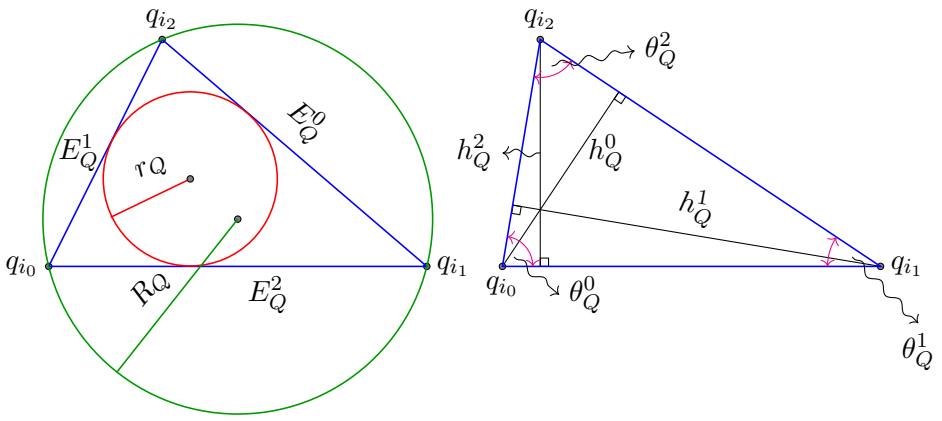


Figure 4.9. Illustration of the geometric measurements of a triangle: inradius r_Q (red), circumradius R_Q (green), edge lengths E_Q^ℓ (blue), heights h_Q^ℓ (black), and interior angles θ_Q^ℓ (magenta). To improve readability we omitted the indices (i_0, i_1, i_2).

if $\{q_{i_0}, q_{i_1}, q_{i_2}\}$ are affine independent, regardless of their orientation. Clearly, if the vertices are in counterclockwise order, then the signed area satisfies the identity

$$A_Q(i_0,i_1,i_2) = \frac{E_Q^\ell(i_0,i_1,i_2)\, h_Q^\ell(i_0,i_1,i_2)}{2} = s_Q(i_0,i_1,i_2)\, r_Q(i_0,i_1,i_2), \tag{4.3}$$

where $s_Q(i_0,i_1,i_2)$ is the semiperimeter, i. e.,

$$s_Q(i_0,i_1,i_2) \coloneqq \frac{1}{2}\left[E_Q^0(i_0,i_1,i_2) + E_Q^1(i_0,i_1,i_2) + E_Q^2(i_0,i_1,i_2)\right]. \tag{4.4}$$

Rearranging terms in (4.3), we have that the heights satisfy the following relation:

$$h_Q^\ell(i_0,i_1,i_2) = 2\, r_Q(i_0,i_1,i_2)\, \frac{s_Q(i_0,i_1,i_2)}{E_Q^\ell(i_0,i_1,i_2)}.$$

Since $s_Q(i_0,i_1,i_2) > \max\{E_Q^0(i_0,i_1,i_2), E_Q^1(i_0,i_1,i_2), E_Q^2(i_0,i_1,i_2)\}$ holds, we have

$$h_Q^\ell(i_0,i_1,i_2) > 2\, r_Q(i_0,i_1,i_2). \tag{4.5}$$

Notice moreover that the following equalities hold.

$$h_Q^\ell(i_0,i_1,i_2) = \sin\Big(\theta_Q^{\ell\oplus1}(i_0,i_1,i_2)\Big)\; E_Q^{\ell\oplus2}(i_0,i_1,i_2)\,,$$
$$h_Q^\ell(i_0,i_1,i_2) = \sin\Big(\theta_Q^{\ell\oplus2}(i_0,i_1,i_2)\Big)\; E_Q^{\ell\oplus1}(i_0,i_1,i_2)\,.$$

Since $\sin\Big(\theta_Q^\ell(i_0,i_1,i_2)\Big) \leq 1$ holds, the heights also satisfy

$$h_Q^\ell(i_0,i_1,i_2) \leq E_Q^{\ell\oplus1}(i_0,i_1,i_2)\,, \quad h_Q^\ell(i_0,i_1,i_2) \leq E_Q^{\ell\oplus2}(i_0,i_1,i_2)\,. \tag{4.6}$$

Every triangle satisfies the Weitzenböck's inequality; see Alsina, Nelsen, 2008

$$4\sqrt{3}\, A_Q(i_0,i_1,i_2) \leq \Big(E_Q^0(i_0,i_1,i_2)^2 + E_Q^1(i_0,i_1,i_2)^2 + E_Q^2(i_0,i_1,i_2)^2\Big)\,, \tag{4.7}$$

and the isoperimetrical inequality (cf. Agricola, Friedrich, 2008, Thm. 25, p. 42)

$$12\sqrt{3}\, A_Q(i_0,i_1,i_2) \leq \Big(E_Q^0(i_0,i_1,i_2) + E_Q^1(i_0,i_1,i_2) + E_Q^2(i_0,i_1,i_2)\Big)^2. \tag{4.8}$$

Both inequalities hold with equality if and only if the triangle $\mathrm{conv}_Q(i_0, i_1, i_2)$ is equilateral.

We will also use the well-known relation $R_Q(i_0, i_1, i_2) \geq 2\, r_Q(i_0, i_1, i_2)$ between the inradius and circumradius (see Svrtan, Veljan, 2012, p. 198). Another useful relation between the inradius and the heights is given by:

$$\frac{1}{r_Q(i_0, i_1, i_2)} = \frac{1}{h^0_Q(i_0, i_1, i_2)} + \frac{1}{h^1_Q(i_0, i_1, i_2)} + \frac{1}{h^2_Q(i_0, i_1, i_2)}, \tag{4.9}$$

see e. g., Kay, 2011, p. 353.

From Bırsan, 2015, Cor. 3 we obtain the following bounds on the interior angles,

$$\begin{aligned} \cos\Big(\theta^\ell_Q(i_0, i_1, i_2)\Big) &\geq \frac{r_Q(i_0, i_1, i_2)}{R_Q(i_0, i_1, i_2)} - \sqrt{1 - \frac{2\, r_Q(i_0, i_1, i_2)}{R_Q(i_0, i_1, i_2)}}, \\ \cos\Big(\theta^\ell_Q(i_0, i_1, i_2)\Big) &\leq \frac{r_Q(i_0, i_1, i_2)}{R_Q(i_0, i_1, i_2)} + \sqrt{1 - \frac{2\, r_Q(i_0, i_1, i_2)}{R_Q(i_0, i_1, i_2)}}. \end{aligned} \tag{4.10}$$

We also use the characterization of the circumradius given in Agricola, Friedrich, 2008, Thm. 27, p. 43,

$$R_Q(i_0, i_1, i_2) = \frac{E^0_Q(i_0, i_1, i_2)\, E^1_Q(i_0, i_1, i_2)\, E^2_Q(i_0, i_1, i_2)}{4 A_Q(i_0, i_1, i_2)}. \tag{4.11}$$

We end this review by introducing the different notions of distance between geometric objects. As already mentioned, the Euclidean distance between two vertices q and q' is going to be denoted by $\|q - q'\|_2$. When we wish to emphasize the vertex numbers in the simplicial complex and the dependence on the matrix Q of node positions, we shall use the alternative notation $d_Q(i_0; i_1) = \|q_{i_0} - q_{i_1}\|_2$ instead. Notice that if $[i_0, i_1, i_2] \in \Delta$ then $d_Q(i_0; i_1)$ coincides with $E^\ell_Q(i_0, i_1, i_2)$ for some $\ell = 0, 1, 2$.

The notation $d_Q(\cdot\,;\cdot)$ is also used to denote the Euclidean distance between higher-dimensional geometric objects, and is based on the following definition of distance between nonempty, convex and compact sets.

Definition 4.2.8. *Let $\mathcal{C}$ denote the collection of all nonempty, convex, compact subsets of $\mathbb{R}^2$. A distance on $\mathcal{C}$ is a mapping $d\colon \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ satisfying the following properties for all $A, B \in \mathcal{C}$.*

(a) $d(A, B) \geq 0$.
(b) $d(A, B) = d(B, A)$.
(c) $d(A, B) > 0$ *if and only if* $A \cap B = \emptyset$.

Note that this definition is more general than a metric, since we do not require it to satisfy the triangle inequality, and unlike metrics, this distance is zero even if the sets are not identical.

In this thesis, we consider the mapping

$$d(A; B) = \min\Big\{ \|a - b\| \;\Big|\; a \in A, \text{and } b \in B \Big\}, \tag{4.12}$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^2$.

Now, we will prove that $d(\cdot, \cdot)$ is indeed a distance in the sense of definition 4.2.8.

Proposition 4.2.9. *Let A, B be two nonempty, convex and compact sets. Then, the mapping $(A, B) \mapsto d(A, B)$ given in (4.12) is well-defined and is a distance in the sense of definition 4.2.8.*

Proof. Let A, B be two nonempty, convex and compact subsets of $\mathbb{R}^2$. Let us consider now the function defined from $A \times B$ to $\mathbb{R}^+$ given by $(a, b) \mapsto \|a - b\|$. This function is a

continuous, and defined on a compact set (namely $A \times B$); therefore, it attains its minimum and its maximum, which proves $d(A, B)$ is well-defined. Moreover, conditions (a) and (b) are immediately obtained from the properties of the norms on $\mathbb{R}^2$.

Now, we focus on the proof of condition (c). First, let us assume that $d(A, B) > 0$, then we wish to prove $A \cap B = \emptyset$. We proceed by contradiction and assume that there exists $c \in A \cap B$. Then, it holds $0 = \|c - c\|_2$. From the fact that $d(A, B)$ is the minimum, it follows $0 \geq d(A, B) > 0$, which clearly produces a contradiction.

Conversely, we assume $A \cap B = \emptyset$ and we aim to prove $d(A, B) > 0$. We proceed again by contradiction, and assume that $d(A, B) = 0$. From the definition of $d(A, B)$ given in (4.12), we know there exist $\bar{a} \in A$ and $\bar{b} \in B$ such that $0 = \|\bar{a} - \bar{b}\|_2$. Using the properties of the norms in $\mathbb{R}^2$, it follows immediately that $\bar{b} \in A$, which contradicts the fact that $A \cap B = \emptyset$. □

We can consider different distances between convex and compact sets associated to different norms. For example, the **Euclidean distance of a vertex $\{q_{i_0}\}$ to an edge** $\mathrm{conv}_Q(j_0, j_1)$ will be denoted by

$$d_Q(i_0; [j_0, j_1]) \coloneqq \min\big\{ \|q_{i_0} - q\|_2 \,\big|\, q \in \mathrm{conv}_Q(j_0, j_1) \big\}.$$

In the same way, one can also consider the **1-norm distance of a vertex to an edge**

$$D_Q(i_0; [j_0, j_1]) \coloneqq \min\big\{ \|q_{i_0} - q\|_1 \,\big|\, q \in \mathrm{conv}_Q(j_0, j_1) \big\},$$

where the 1-norm is based on an edge oriented coordinate system; see figure 4.10 for an illustration.

It can easily be shown that for the 1-norm distance of a vertex to an edge

$$D_Q(i_0; [j_0, j_1]) = \|q_{i_0} - q_{j_0} + t\,(q_{j_1} - q_{j_0})\|_1 \tag{4.13a}$$

holds, where

$$t = \max\left\{0, \min\left\{1, \frac{(q_{i_0} - q_{j_0}) \cdot (q_{j_1} - q_{j_0})}{\|q_{j_1} - q_{j_0}\|_2^2}\right\}\right\}. \tag{4.13b}$$

Clearly, the well-known equivalence of norms implies

$$d_Q(i_0; [j_0, j_1]) \leq D_Q(i_0; [j_0, j_1]) \leq \sqrt{2}\, d_Q(i_0; [j_0, j_1]). \tag{4.14}$$

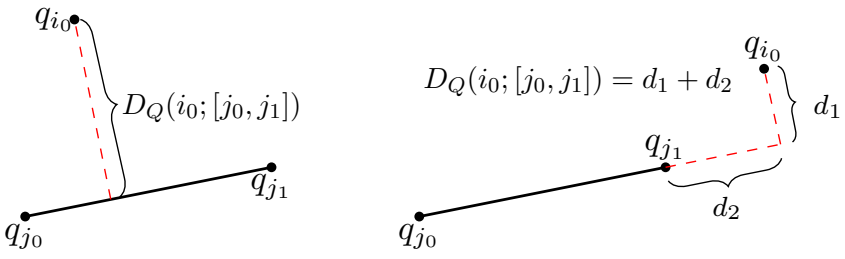


Figure 4.10. Illustration of the distance (4.13) of a vertex to an edge in an edge oriented coordinate system. The two cases shown are when the projection of the vertex onto the infinite line generated by the edge belongs to the edge (left), and when it does not (right).

Finally, we consider Euclidean distances between two faces of different dimensions. The next proposition allows us to express the distance between theses faces as the minimum of all the possible Euclidean distances of a vertex to an edge.

Proposition 4.2.10. *Let Δ be a connectivity complex over the vertex set $\{1,\dots,N_V\}$, and Σ a geometric simplicial 2-complex, with vertices $q_1,\dots,q_{N_V}$, stored in a matrix $Q \in \mathbb{R}^{2\times N_V}$. We assume Δ is the abstract simplicial complex associated with Σ. Moreover, let us consider two distinct faces of Δ, denoted by $\{i_0,\dots,i_m\}$ and $\{j_0,\dots,j_n\}$. If we assume the corresponding geometric faces $\mathrm{conv}_Q(i_0,\dots,i_m)$ and $\mathrm{conv}_Q(j_0,\dots,j_n)$ satisfy $\mathrm{conv}_Q(i_0,\dots,i_m)\cap\ \mathrm{conv}_Q(j_0,\dots,j_n)=\emptyset$, then, the Euclidean distance between them satisfies the following expression:*

$$d_Q([i_0,\dots,i_m];[j_0,\dots,j_n]) = \min_{\substack{\ell=0,\dots,m\\ \hat{\ell}=0,\dots,n}} \Big\{ d_Q(i_\ell;[j_{\hat{\ell}},j_{\hat{\ell}\oplus 1}]), d_Q(j_{\hat{\ell}};[i_\ell,i_{\ell\oplus 1}]) \Big\}, \tag{4.15}$$

for $m,n=\{0,1,2\}$, such that m,n do not equal zero simultaneously.

Proof. Recall the definition of the Euclidean distance between nonempty, convex and compact sets given in (4.12):

$$d_Q([i_0,\dots,i_m];[j_0,\dots,j_n]) = \min\Big\{ \|q-\tilde{q}\|_2 \,\Big|\, q\in \mathrm{conv}_Q(i_0,\dots,i_m), \tilde{q}\in \mathrm{conv}_Q(j_0,\dots,j_n)\Big\}.$$

Using the definition of the convex hull, we obtain the following equivalent expression for the Euclidean distance:

$$\begin{aligned} d_Q([i_0,\dots,i_m];[j_0,\dots,j_n]) = \min\Big\{ &\|[\eta_0 q_{i_0}+\ldots+\eta_m q_{i_m}]-[\lambda_0 q_{j_0}+\ldots+\lambda_n q_{j_n}]\|_2 | \\ &\eta_0,\dots,\eta_m\geq 0, \eta_0+\ldots+\eta_m=1, \\ &\lambda_0,\dots,\lambda_n\geq 0, \lambda_0+\ldots+\lambda_n=1\Big\}. \end{aligned}$$

We will start with the case $m=n=1$, i. e., we aim to express $d_Q([i_0,i_1];[j_0,j_1])$ as the minimum of the all possible Euclidean distances of a vertex to an edge. Precisely, we aim to solve the problem

$$\min\left\{ \big\| [\eta q_{i_0}+(1-\eta)q_{i_1}]-[\lambda q_{j_0}+(1-\lambda)q_{j_1}]\big\|_2 \,\big|\, \eta,\lambda\in[0,1]\right\}$$

Since $\mathrm{conv}_Q(i_0,\dots,i_m)\cap\mathrm{conv}_Q(j_0,\dots,j_n)=\emptyset$, the function $\|\cdot\|_2$ is continuous and differentiable for all $(\eta,\lambda)\in[0,1]^2$. This implies the optimal points (denoted by) $\bar{\eta},\bar{\lambda}$ may lie on the boundary $\{0,1\}$ of the compact set $[0,1]^2$ or in its interior. Now, we analyze the different cases.

$\bar{\eta}=0$**:** from the definition it immediately implies

$$d_Q([i_0,i_1];[j_0,j_1]) = \min\{\|q_{i_1}-[\lambda q_{j_0}+(1-\lambda)q_{j_1}]\|_2 \,|\, \lambda\in[0,1]\} = d_Q(i_1;[j_0,j_1]).$$

$\bar{\eta}=1$**:** under similar arguments

$$d_Q([i_0,i_1];[j_0,j_1]) = \min\{\|q_{i_0}-[\lambda q_{j_0}+(1-\lambda)q_{j_1}]\|_2 \,|\, \lambda\in[0,1]\} = d_Q(i_0;[j_0,j_1]).$$

$\bar{\lambda}=0$**:** $d_Q([i_0,i_1];[j_0,j_1]) = \min\{\|[\eta q_{i_0}+(1-\eta)q_{i_1}]-q_{j_1}\|_2 \,|\, \eta\in[0,1]\} = d_Q(j_1;[i_0,i_1])$.
$\bar{\lambda}=1$**:** $d_Q([i_0,i_1];[j_0,j_1]) = \min\{\|[\eta q_{i_0}+(1-\eta)q_{i_1}]-q_{j_0}\|_2 \,|\, \eta\in[0,1]\} = d_Q(j_0;[i_0,i_1])$.
$\bar{\eta},\bar{\lambda}\in(0,1)$**:** we will prove that this case cannot hold. Assume for now that the minimum is attained in the interior of the compact set $[0,1]^2$, then,

$$\begin{aligned} \frac{\partial}{\partial\eta}\big(\|\overline{\eta}q_{i_0}+(1-\overline{\eta})q_{i_1}-[\overline{\lambda}q_{j_0}+(1-\overline{\lambda})q_{j_1}]\|_2\big)=0, \\ \frac{\partial}{\partial\lambda}\big(\|\overline{\eta}q_{i_0}+(1-\overline{\eta})q_{i_1}-[\overline{\lambda}q_{j_0}+(1-\overline{\lambda})q_{j_1}]\|_2\big)=0, \end{aligned}$$

where $\partial/\partial\eta$, $\partial/\partial\lambda$ denote the partial derivatives of the norm w.r.t. η and λ, respectively. It can be proved that the following equality holds:

$$\begin{aligned}
&\begin{bmatrix} \frac{\partial}{\partial\eta}\left(\|\overline{\eta}q_{i_0} + (1-\overline{\eta})q_{i_1} - [\overline{\lambda}q_{j_0} + (1-\overline{\lambda})q_{j_1}]\|\right) \\ \frac{\partial}{\partial\eta}\left(\|\overline{\eta}q_{i_0} + (1-\overline{\eta})q_{i_1} - [\overline{\lambda}q_{j_0} + (1-\overline{\lambda})q_{j_1}]\|\right) \end{bmatrix} \\
&= c \begin{bmatrix} q^1_{i_1} - q^1_{i_0} & q^2_{i_1} - q^2_{i_0} \\ q^1_{j_1} - q^1_{j_0} & q^2_{j_1} - q^2_{j_0} \end{bmatrix} \begin{bmatrix} \overline{\eta}q^1_{i_0} + (1-\overline{\eta})q^1_{i_1} - [\overline{\lambda}q^1_{j_1} - (1-\overline{\lambda}q^1_{j_1}] \\ \overline{\eta}q^2_{i_0} + (1-\overline{\eta})q^2_{i_1} - [\overline{\lambda}q^2_{j_1} - (1-\overline{\lambda}q^2_{j_1}] \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 0 \end{bmatrix},
\end{aligned}$$

with $c = 1/\|\overline{\eta}q_{i_0} + (1-\overline{\eta})q_{i_1} - [\overline{\lambda}q_{j_0} + (1-\overline{\lambda})q_{j_1}]\|_2$. Since $\overline{\lambda}$ and $\overline{\eta}$ belong to $(0,1)^2$, the it holds: $\|\overline{\eta}q_{i_0} + (1-\overline{\eta})q_{i_1} - [\overline{\lambda}q_{j_0} + (1-\overline{\lambda})q_{j_1}]\|_2, \neq 0$, and

$$\begin{bmatrix} q^1_{i_1} - q^1_{i_0} & q^2_{i_1} - q^2_{i_0} \\ q^1_{j_1} - q^1_{j_0} & q^2_{j_1} - q^2_{j_0} \end{bmatrix} \neq 0.$$

Therefore, it must hold

$$\left[\overline{\eta}q_{i_0} + (1-\overline{\eta})q_{i_1} - [\overline{\lambda}q_{j_1} - (1-\overline{\lambda}q_{j_1}]\right] = 0.$$

This implies, $\overline{\lambda}q_{j_1} - (1-\overline{\lambda})q_{j_1} \in \mathrm{conv}_Q(i_0, i_1)$ which contradicts the fact that the intersection is empty.

Summarizing,

$$d_Q([i_0,i_1]; j_0, j_1) = \min\{d_Q(i_0;[j_0,j_1]), d_Q(i_1;[j_0,j_1]), d_Q(j_0;[i_0,i_1]), d_Q(j_1;[i_0,i_1])\}\,.$$

The same arguments can be used when $m = 0$ and $n = 2$, i. e., for the Euclidean distance of a vertex to a triangle. Using the definition of the convex hull we have:

$$d_Q(i_0;[j_0,j_1,j_2]) = \min\{\|q_{i_0} - [\lambda_1 q_{j_0} + \lambda_2 q_{j_1} + (1-\lambda_1-\lambda_2)q_{j_2}]\|_2 \,|\, \lambda_1,\lambda_2 \geq 0,\ \lambda_1+\lambda_2 \leq 1\}$$

Analyzing by cases again we obtain

$\bar{\lambda}_1 = 0$**:** $d_Q(i_0;[j_0,j_1,j_2]) = \min\{\|q_{i_0} - [\lambda_2 q_{j_1} + (1-\lambda_2)q_{j_2}\|_2 \,|\, \lambda_2 \in [0,1]\} = d_Q(i_0;[j_1,j_2])$.
$\bar{\lambda}_2 = 0$**:** $d_Q(i_0;[j_0,j_1,j_2]) = \min\{\|q_{i_0} - [\lambda_1 q_{j_0} + (1-\lambda_1)q_{j_2}\|_2 \,|\, \lambda_2 \in [0,1]\} = d_Q(i_0;[j_0,j_2])$.
$\bar{\lambda}_1 + \bar{\lambda}_2 = 1$**:** $d_Q(i_0;[j_0,j_1,j_2]) = \min\{\|q_{i_0} - [\lambda_1 q_{j_0} + \lambda_2 q_{j_1}\|_2 \,|\, \lambda_2 \in [0,1]\} = d_Q(i_0;[j_0,j_1])$.
$\bar{\lambda}_1 + \bar{\lambda}_2 < 1$ **and** $\bar{\lambda}_1, \bar{\lambda}_2 > 0$**:** in this case one can use similar arguments to show that this cannot hold.

Altogether implies

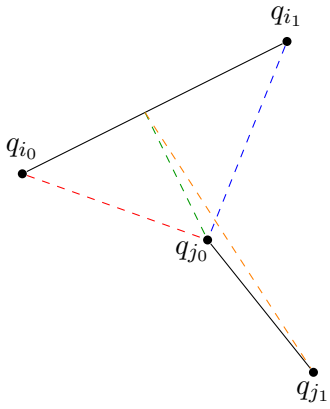
$$d_Q(i_0;[j_0,j_1,j_2]) = \min\{d_Q(i_0;[j_0,j_1]), d_Q(i_0;[j_1,j_2]), d_Q(i_0;[j_2,j_0]),\}$$

Finally, the Euclidean distance of an edge to a triangle and of a triangle to a triangle can be expressed in terms of Euclidean distance of a vertex to a triangle and/or the Euclidean distance of an edge to and edge. The expression showed in (4.15) can be obtained as a generalization of these results. See figure 4.11 for an illustration of the Euclidean distances between different faces of different dimensions. □

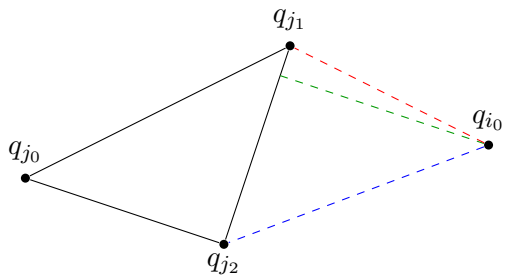
4.3 Construction of the Manifold of Planar Triangular Meshes

Having introduced all the required notions, we are ready to describe the step-by-step construction of the manifold of planar triangular meshes. This section is based on Herzog, Loayza-Romero, 2020, Sec. 3.

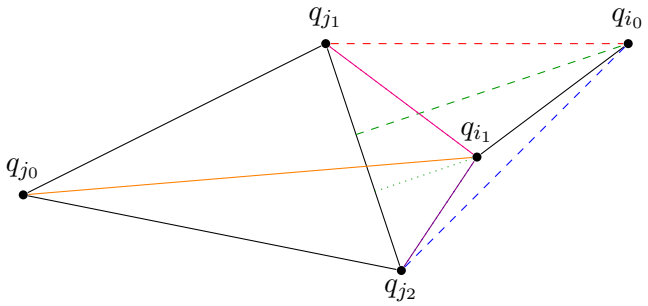
We consider discrete shapes as triangular meshes in $\mathbb{R}^2$. In the language of simplicial complexes, the connectivity information of such meshes is precisely a *pure, 2-path connected abstract simplicial 2-complex*, or as given in definition 4.2.5, a *connectivity complex*. This



(a) edge–edge distance is the minimum of the (four different) vertex-edge distances. The lengths of the lines depict the distances $d_Q(i_0; [j_0, j_1])$ (red), $d_Q(i_1; [j_0, j_1])$ (blue), $d_Q(j_0; [i_0, i_1])$ (green), and $d_Q(j_1; [i_0, i_1])$ (orange).



(b) vertex–triangle distance is the minimum of the (three different) vertex-edge distances. The lengths of the lines depict the distances $d_Q(i_0; [j_0, j_1])$ (red), $d_Q(i_0; [j_0, j_2])$ (blue), and $d_Q(i_0; [j_1, j_2])$ (green).



(c) edge–triangle distance is the minimum of the (nine different) vertex-edge distances. The lengths of the dashed lines depict the distances $d_Q(i_0; [j_0, j_1])$ (red), $d_Q(i_0; [j_1, j_2])$ (green) and $d_Q(i_0; [j_0, j_2])$ (blue). The lengths of the dotted lines depict the distances $d_Q(i_1; [j_0, j_1])$ (red), $d_Q(i_1; [j_1, j_2])$ (green) and $d_Q(i_1; [j_0, j_2])$ (blue). The lengths of the solid lines depict the distances $d_Q(j_0; [i_0, i_1])$ (orange), $d_Q(j_1; [i_0, i_1])$ (magenta) and $d_Q(j_2; [i_0, i_2])$ (violet).

Figure 4.11. Euclidean distance between objects of different dimensions.

implies that for a given distribution of vertices over $\mathbb{R}^2$, the union of all the resulting triangles forms a connected subset of $\mathbb{R}^2$. Two examples of such meshes are shown in figure 4.12, which are similar to the ones presented in Alexa, Cohen-Or, Levin, 2000, Fig. 4.

We are interested in all possible configurations of node positions a mesh of a given connectivity can attain. To this end, we need to formulate conditions which avoid triangles becoming degenerate and vertices entering triangles to which they are not incident; see figure 4.13.

As already motivated, to make these ideas formal, we utilize the concept of abstract simplicial complexes as well as geometric simplicial complexes; see section 4.2. For simplicity of notation and without loss of generality, the 0-faces of the connectivity complex Δ will be numbered $\{1, \ldots, N_V\}$. We emphasize that all geometric simplicial complexes throughout

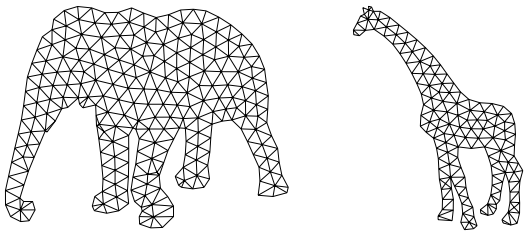


Figure 4.12. Two examples of triangular meshes.

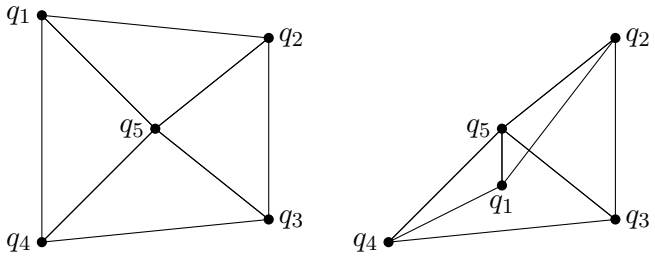


Figure 4.13. Admissible (left) and inadmissible (right) assignment of vertex coordinates for two meshes sharing the same connectivity.

this thesis have vertices in $\mathbb{R}^2$. We recall that the terms vertex, edge, and triangle are used in the context of geometric simplicial complexes, while $0, 1, 2$-faces are reserved for the abstract simplicial complexes. Moreover, we will use indistinctly the terms node of a mesh and vertices of a geometric simplicial complex.

A connectivity complex Δ is a purely combinatorial object, which we can think of as a recipe for constructing meshes. In order to achieve the latter, we need to assign node positions. These can be summarized in a matrix

$$Q = \begin{bmatrix} q_1, & q_2, & \dots, & q_{N_V} \end{bmatrix} \in \mathbb{R}^{2 \times N_V}. \tag{4.16}$$

As is illustrated in figure 4.13, not all assignments of node positions will give rise to an admissible mesh. In order to distinguish those which do from those which don't, we require further notation. Given a connectivity complex Δ and an assignment Q of its node positions, we define

$$\Sigma_\Delta(Q) \coloneqq \big\{ \mathrm{conv}_Q(i_0, \dots, i_k) \,\big|\, \{i_0, \dots, i_k\} \in \Delta \big\} \subset \mathcal{P}(\mathbb{R}^2), \tag{4.17}$$

where $\mathcal{P}(\mathbb{R}^2)$ denotes the power set of $\mathbb{R}^2$. In other words, $\Sigma_\Delta(Q)$ collects the convex hulls of the vertices of all 0-, 1- and 2-faces in Δ. To illustrate the set $\Sigma_\Delta(Q)$ we consider the following example.

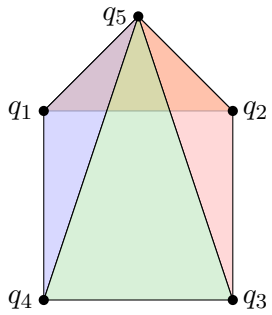
Example 4.3.1. *Let us revisit example 4.2.6, and consider the connectivity complex Δ. Additionally, we consider the node positions given by:*

$$Q = \begin{bmatrix} 0 & 1 & 1 & 0 & 0.5 \\ 1 & 1 & 0 & 0 & 1.5 \end{bmatrix}.$$

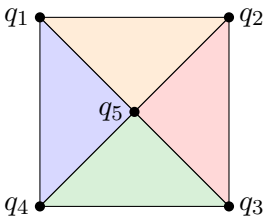
The collection of convex hulls $\Sigma_\Delta(Q)$ is

$$\begin{aligned} \Sigma_\Delta(Q) = \big\{ &\{q_1\}, \{q_2\}, \{q_3\}, \{q_4\}, \{q_5\}, \mathrm{conv}_Q(1,2), \mathrm{conv}_Q(2,3), \mathrm{conv}_Q(3,4), \\ &\mathrm{conv}_Q(4,1), \mathrm{conv}_Q(1,5), \mathrm{conv}_Q(2,5), \mathrm{conv}_Q(3,5), \mathrm{conv}_Q(4,5), \\ &\mathrm{conv}_Q(1,2,5), \mathrm{conv}_Q(2,3,5), \mathrm{conv}_Q(3,4,5), \mathrm{conv}_Q(4,5,1) \big\}. \end{aligned}$$

Figure 4.14a illustrates this collection of convex hulls.



(a) $\Sigma_\Delta(Q)$ described in example 4.3.1



(b) $\Sigma_\Delta(\widetilde{Q})$ described in example 4.3.3

Figure 4.14. Collection of convex hulls $\Sigma_\Delta(Q)$. The vertices are depicted by black circles, the edges in solid black lines. Moreover, the triangles are shaded $\operatorname{conv}_Q(1,2,5)$ (blue), $\operatorname{conv}_Q(2,3,5)$ (orange), $\operatorname{conv}_Q(3,4,5)$ (red), and $\operatorname{conv}_Q(4,5,1)$ (green).

We can now formalize the set of all admissible meshes with a given connectivity as follows.

Definition 4.3.2. *Suppose that Δ is a connectivity complex with vertex set $\{1,\dots,N_V\}$. Then we define the* ***set of admissible meshes with connectivity*** *Δ as*

$$\mathcal{M}_0(\Delta) \coloneqq \left\{ Q \in \mathbb{R}^{2\times N_V} \;\middle|\; \begin{array}{l} \Sigma_\Delta(Q) \text{ is a geometric simplicial 2-complex} \\ \text{whose associated abstract simplicial complex is } \Delta \end{array} \right\}. \tag{4.18}$$

It follows from definition 4.3.2 that if $Q \in \mathcal{M}_0(\Delta)$ then $\{q_{i_0}, q_{i_1}, q_{i_2}\}$ are affine independent for all $\{i_0, i_1, i_2\} \in \Delta$.

The conditions of $\Sigma_\Delta(Q)$ formalize the idea of an admissible mesh, i. e., that all triangles be nondegenerate and that any two intersecting triangles can only intersect in a common edge or a common vertex. For example, Q described in example 4.3.1 does not belong to $\mathcal{M}_0(\Delta)$, while the assignment of the nodes described in example 4.3.3 does.

Example 4.3.3. *Let us revisit the oriented connectivity complex described in example 4.3.1. Additional to Q, we consider the node positions given by $\widetilde{Q}$ as follows:*

$$Q = \begin{bmatrix} 0 & 2 & 1 & 0 & 0.5 \\ 1 & 1 & 0 & 0 & 1.5 \end{bmatrix}, \quad \text{and} \quad \widetilde{Q} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0.5 \\ 1 & 1 & 0 & 0 & 0.5 \end{bmatrix}.$$

It is easy to verify that Q does not belong to $\mathcal{M}_0(\Delta)$, while $\widetilde{Q}$ does. See figures 4.14a and 4.14b. for an illustration.

Another important aspect of this construction is that we need to insist in definition (4.18) that the abstract simplicial complex associated with $\Sigma_\Delta(Q)$ agrees with Δ. As an example, consider

$$\Delta = \big\{\{1\},\{2\},\{3\},\{4\},\{1,2\},\{2,3\},\{3,1\},\{1,4\},\{4,2\},\{2,1\},\{1,2,3\},\{2,1,4\}\big\}$$

and choose coordinates stored in $Q \in \mathbb{R}^{2\times 4}$ such that $\operatorname{conv}_Q(1,2,3)$ and $\operatorname{conv}_Q(2,1,4)$ are 2-simplices but $q_3 = q_4$ holds. Then $\Sigma_\Delta(Q)$ is a geometric simplicial complex but its abstract simplicial complex is smaller than Δ since the two triangles coincide.

It is easy to see that there exist connectivity complexes Δ for which $\mathcal{M}_0(\Delta)$ is empty. This is the case, for instance, when

$$\Delta = \big\{\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{2,3\},\{3,1\},\{1,4\},\{2,4\},\{1,5\},\{2,5\},\\ \{1,2,3\},\{1,2,4\},\{1,2,5\}\big\},$$

which was already considered in example 4.2.7.

Notice that there are three 2-faces with one common 1-face, representing an impossible configuration for a geometric simplicial 2-complex in $\mathbb{R}^2$.

The possible emptiness of $\mathcal{M}_0(\Delta)$ will not be a cause of concern in what follows.

Proposition 4.3.4. *For any given connectivity complex Δ with vertex set $\{1,\dots,N_V\}$, the set $\mathcal{M}_0(\Delta)$ is an open (possibly empty) subset of $\mathbb{R}^{2\times N_V}$.*

Proof. We can assume that $\mathcal{M}_0(\Delta)$ is nonempty . Let $Q \in \mathcal{M}_0(\Delta)$ be arbitrary. We need to prove that there exists $\delta > 0$ such that the open ball $B_\delta(Q) \subset \mathbb{R}^{2\times N_V}$, e. g., in the Frobenius norm, belongs to $\mathcal{M}_0(\Delta)$. We proceed in the following steps, selecting a suitable $\delta > 0$ along the way. We show that, for all $U \in B_\delta(Q)$,

(*i*) $\Sigma_\Delta(U)$ is a geometric simplicial 2-complex and
(*ii*) $\Sigma_\Delta(U)$ has associated abstract simplicial complex Δ.

We begin with statement (*i*). Suppose that $\{i_0, i_1, i_2\}$ is an arbitrary 2-face in Δ. Since $\{q_{i_0}, q_{i_1}, q_{i_2}\}$ is affine independent, $\det\left[q_{i_1} - q_{i_0},\ q_{i_2} - q_{i_1}\right] \neq 0$. By continuity of the determinant function, we can find $\delta > 0$ such that $\det\left[u_{i_1} - u_{i_0},\ u_{i_2} - u_{i_1}\right]$ has the same sign as before for all $U \in B_\delta(Q)$. Therefore, $\mathrm{conv}_U(i_0, i_1, i_2)$ is a 2-simplex, i. e., a collection of three affine independent points. Since the number of 2-faces in Δ is finite, a joint value of $\delta > 0$ can be found which is valid for all 2-faces in Δ. Clearly, the same reasoning also applies to the 1-faces and 0-faces. Consequently, for all $U \in B_\delta(Q)$, $\Sigma_\Delta(U)$ consists of a collection of simplices whose dimension agrees with the dimension of the corresponding face of Δ.

In the following, let σ and σ' be any two *distinct* faces in Δ. We denote by $\sigma(Q)$ and $\sigma'(Q)$ the corresponding simplices in $\Sigma_\Delta(Q)$. For instance, when $\sigma = \{i_0,\dots,i_k\}$, then $\sigma(Q) = \mathrm{conv}_Q(i_0,\dots,i_k)$. By construction, it is clear that when $\tau \subset \sigma$ holds, then $\tau(U)$ is a face of $\Sigma_\Delta(U)$, for all $U \in B_\delta(Q)$. We also know that $\tau(Q) \coloneqq \sigma(Q) \cap \sigma'(Q)$ is either empty or a face of both. In order to conclude statement (*i*), we now show that this property extends to all $U \in B_\delta(Q)$, possibly for a smaller value of $\delta > 0$ than previously chosen. We distinguish two cases. • When $\tau(Q) = \emptyset$, then since $\sigma(Q)$ and $\sigma'(Q)$ are compact and convex, there exists an affine linear functional φ such that $\varphi < 0$ on $\sigma(Q)$ and $\varphi > 0$ on $\sigma'(Q)$; see, e. g., Hiriart-Urruty, Lemaréchal, 2001, Cor. 4.1.3, p. 52. Possibly by making δ smaller, we retain $\varphi < 0$ on $\sigma(U)$ and $\varphi > 0$ on $\sigma'(U)$ for all $U \in B_\delta(Q)$. Consequently, $\tau(U) = \sigma(U) \cap \sigma'(U) = \emptyset$. • Suppose that $\tau(Q)$ is a face of both $\sigma(Q)$ and $\sigma'(Q)$, say, $\sigma(Q) = \mathrm{conv}_Q(i_0,\dots,i_k)$, $\sigma'(Q) = \mathrm{conv}_Q(j_0,\dots,j_\ell)$ and $\tau(Q) = \mathrm{conv}_Q(i_0,\dots,i_m) = \mathrm{conv}_Q(j_0,\dots,j_m)$ with some $0 \leq m \leq \min\{k,\ell\}$. We have already proved that $\sigma(U)$ and $\sigma'(U)$ are simplices of dimensions k and ℓ, respectively, for all $U \in B_\delta(Q)$. Therefore, the only concern is that $\tau(U)$ is larger than $\mathrm{conv}_U(i_0,\dots,i_m)$. In each case, however, we can construct a hyperplane, defined by two vertices of either σ or σ', which separates $\sigma(U) \setminus \sigma'(U)$ from $\sigma'(U) \setminus \sigma(U)$, for all $U \in B_\delta(Q)$, possibly for a smaller value of $\delta > 0$ than previously chosen. See figure 4.15 for an illustration.

Altogether, this confirms that $\tau(U) \coloneqq \sigma(U) \cap \sigma'(U)$ is either empty or a face of both for U in a suitable ball $B_\delta(Q)$. While looping over all pairs of distinct faces $\{\sigma, \sigma'\}$, δ needs to be reduced only finitely many times, therefore we have shown statement (*i*).

In order to show statement (*ii*), recall from subsection 4.2.2 that the abstract simplicial complex associated with $\Sigma_\Delta(Q)$ is defined as

$$\Delta(Q) = \big\{\{i_0, \dots, i_m\} \subset \{1, \dots, N_V\} \,\big|\, \mathrm{conv}_Q(i_0, \dots, i_m) \in \Sigma_\Delta(Q)\big\}.$$

Since $Q \in \mathcal{M}_0(\Delta)$, $\Delta(Q) = \Delta$ holds by definition. Moreover, by (4.17) we clearly have $\Delta \subset \Delta(U)$ for all $U \subset \mathbb{R}^{2\times N_V}$ and thus for all $U \in B_\delta(Q)$. However, the considerations above show that there can be no additional simplices in $\Sigma_\Delta(U)$ than those coming from (4.17). In other words, $\Delta = \Delta(U)$ holds for all $U \in B_\delta(Q)$, which is statement (*ii*). □

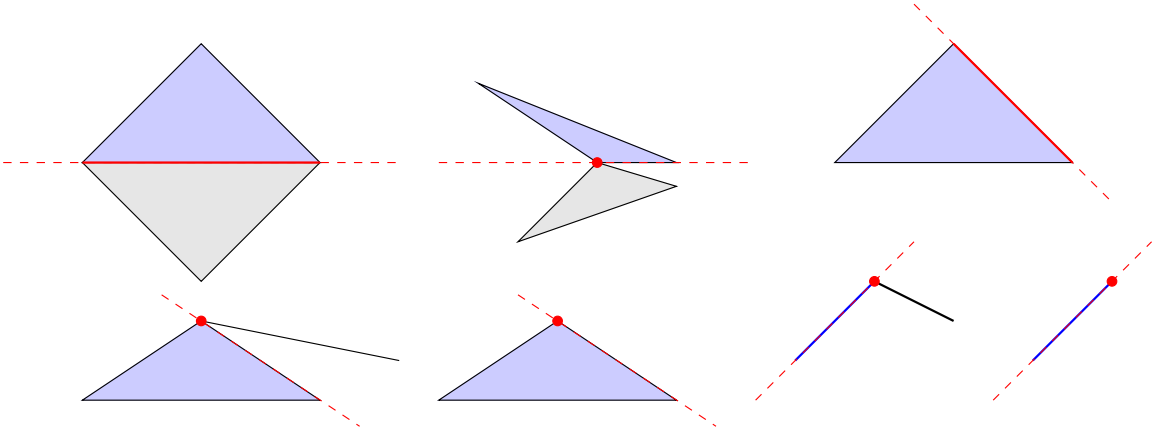


Figure 4.15. Illustration for the proof of proposition 4.3.4, showing (from top left to bottom right) the representative cases $(k, \ell, m) \in \{(2,2,1), (2,2,0), (2,1,1), (2,1,0), (2,0,0), (1,1,0), (1,0,0)\}$. $\sigma(Q)$ is shown in blue, $\sigma'(Q)$ is shown in black, and their intersection $\tau(Q)$ is shown in red. A possible separating hyperplane is displayed as a dashed red line.

Proposition 4.3.4 shows that $\mathcal{M}_0(\Delta)$ is a smooth open submanifold of $\mathbb{R}^{2\times N_V}$. This is obtained since an open subset of a manifold is can be endowed with the subspace topology, and with the atlas $(\mathbb{R}^{2\times N_V}, \mathrm{id})$ of $\mathbb{R}^{2\times N_V}$, then $(\mathbb{R}^{2\times N_V} \cap \mathcal{M}_0(\Delta), \mathrm{id})$ is an atlas for $\mathcal{M}_0(\Delta)$. In the same way, one can derive a smooth structure for $\mathcal{M}_0(\Delta)$ using the smooth structure from $\mathbb{R}^{2\times N_V}$. We refer the reader to theorem A.1.1 and remark A.1.3 for a formal statement of these results.

It is easy to see that any nonempty $\mathcal{M}_0(\Delta)$ is not path connected and, equivalently, not connected. In fact, $Q \in \mathcal{M}_0(\Delta)$ implies that $-Q$ lies in $\mathcal{M}_0(\Delta)$ as well, but in a different connected component. This is true even if Δ contains only a single 2-face. For example, the two meshes shown in figure 4.16 cannot be joined by a continuous path of node positions that remains inside $\mathcal{M}_0(\Delta)$. In order to resolve this issue, we need to consider orientations.

Recall that an orientation of an abstract 2-face $\sigma = \{i_0, i_1, i_2\}$ is an equivalence class of orderings of σ, where two orderings i_0, i_1, i_2 and $i_{\pi(0)}$, $i_{\pi(1)}$ and $i_{\pi(2)}$ are equivalent if and only if π is an even permutation of $0, 1, 2$, and we denote as $[i_0, i_1, i_2]$ the oriented 2-faces of Δ. Moreover, we say Δ is orientable if there exist orientations of all 2-simplices in Δ which render Δ consistently oriented, as introduced in section 4.2.

Remark 4.3.5. *It is easy to see that a connectivity complex Δ can be encoded as a* ***connectivity matrix*** *$C \in \mathbb{R}^{3\times N_T}$, where N_T is the number of 2-faces in Δ and each column of C*

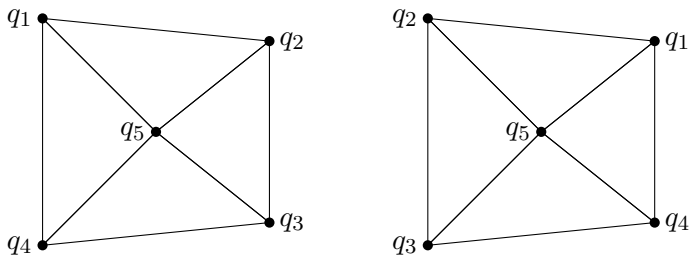


Figure 4.16. Two admissible meshes with the same connectivity complex Δ whose node position matrices Q lie in different connected components of $\mathcal{M}_0(\Delta)$.

lists the 0*-faces of one of the* 2*-faces. For instance, the connectivity complex*

$$\Delta = \big\{\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{2,3\},\{3,4\},\{4,1\},\{1,5\},\{2,5\},\\ \{3,5\},\{4,5\},\{1,2,5\},\{2,3,5\},\{3,4,5\},\{4,1,5\}\big\}$$

could be rewritten as:

$$C = \begin{bmatrix} 2 & 3 & 4 & 1 \\ 1 & 2 & 3 & 4 \\ 5 & 5 & 5 & 5 \end{bmatrix}.$$

Its consistent orientation is reflected in that 2*-faces sharing a* 1*-face have their two common* 0*-faces appearing in opposite orderings. Consider the third and fourth columns, where the shared* 0*-faces appear (after an even permutation) in the orders* $[5,4]$ *and* $[4,5]$*, respectively. Moreover, we would like to highlight that triangular mesh generators, including the one driving the* `initmesh` *function of* MATLAB*'s PDE toolbox usually provides this kind of connectivity matrices.*

We recall, we will say that "$[i_0,i_1,i_2]$ is a 2-face" instead of "$\{i_0,i_1,i_2\}$ is a 2-face with orientation $[i_0,i_1,i_2]$". Similarly, we write $[i_0,i_1]$ to denote a 1-face together with its orientation. For consistency of notation, we will also write $[i_0]$ instead of $\{i_0\}$ for a 0-face although orientation does not matter there. For the remainder of this thesis, we assume the following.

Assumption 4.3.6. *Let* Δ *be a consistently oriented connectivity complex with vertex set given by* $\{1,\dots,N_V\}$*. The matrix* $Q \in \mathbb{R}^{2\times N_V}$ *denotes an arbitrary assignment of its node positions.*

We recall the notion of signed area introduced in (4.2), subsection 4.2.3.

$$A_Q(i_0,i_1,i_2) \coloneqq \frac{1}{2}\det\left[q_{i_1}-q_{i_0},\ q_{i_2}-q_{i_1}\right].$$

The signed area allows us to introduce the set of admissible oriented meshes with connectivity Δ.

Definition 4.3.7. *We define the* ***set of admissible oriented meshes with connectivity*** Δ *as*

$$\mathcal{M}_+(\Delta) \coloneqq \big\{Q \in \mathcal{M}_0(\Delta) \,\big|\, A_Q(i_0,i_1,i_2) > 0 \text{ for all 2-faces } [i_0,i_1,i_2] \text{ of } \Delta\big\}. \tag{4.19}$$

Figure 4.17 illustrates two elements of $\mathcal{M}_+(\Delta)$.

Proposition 4.3.8. *The set* $\mathcal{M}_+(\Delta)$ *is an open (possibly empty) subset of* $\mathbb{R}^{2\times N_V}$*.*

PROOF. Due to the continuity of the determinant and proposition 4.3.4, the set described in (4.19) is a finite intersection of open subsets of $\mathbb{R}^{2\times N_V}$. □

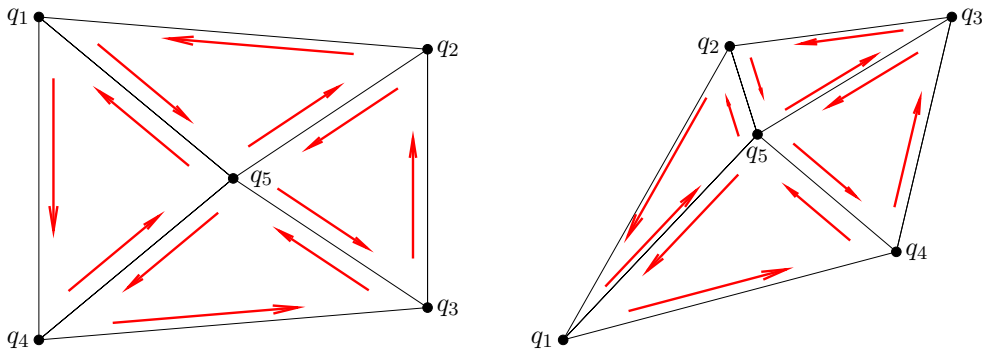


Figure 4.17. Two admissible oriented meshes, i. e., elements of $\mathcal{M}_+(\Delta)$, with the same consistently oriented connectivity complex Δ and different vertex positions Q. The orientation of the 2-faces is $[2,1,5]$, $[3,2,5]$, $[4,3,5]$, and $[1,4,5]$.

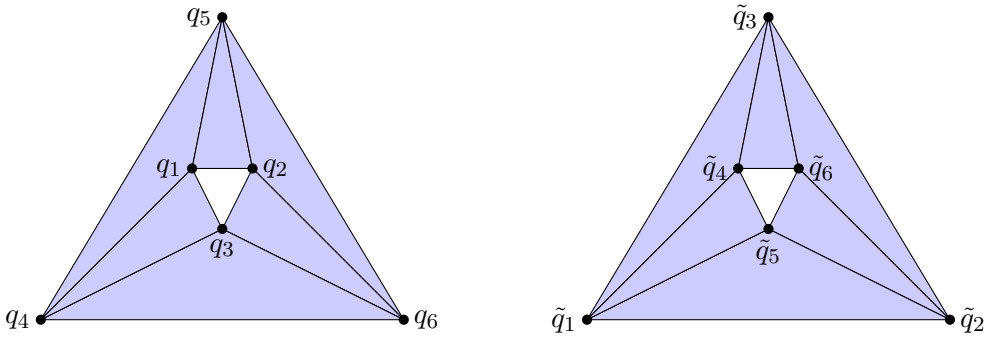


Figure 4.18. Illustration of the nonconnectedness of $\mathcal{M}_+(\Delta)$; see example 4.3.9 for details.

The set $\mathcal{M}_+(\Delta)$ in definition 4.3.7 formalizes our initial question, which node positions a mesh, or rather an oriented mesh, can attain? However, if we aim to endow it with a complete Riemannian metric, we are not quite done yet. Another important aspect is the connectedness of $\mathcal{M}_+(\Delta)$. The manifold $\mathcal{M}_+(\Delta)$ is in general not connected. Consider the following counterexample.

Example 4.3.9. *Let us consider the following connectivity complex*

$$\Delta = \big\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{1,2\},\{2,3\},\{1,3\},\{1,5\},\{2,5\},\{2,6\},\{3,6\},\{3,4\},\\ \{1,4\},\{4,5\},\{5,6\},\{4,6\},\{1,3,4\},\{1,2,5\},\{2,3,6\},\{3,4,6\},\{1,4,5\},\{2,5,6\}\big\},$$

with the following orientation on its 2-faces: $[1,4,3]$, $[1,2,5]$, $[2,3,6]$, $[3,4,6]$, $[1,5,4]$ *and* $[2,6,5]$. *Moreover, let us consider the node positions:*

$$Q = \begin{bmatrix} 0.75 & 1.25 & 1 & -0.5 & 1 & 2.5 \\ 1.25 & 1.25 & 0.75 & 0 & 2.5 & 0 \end{bmatrix}, \quad \widetilde{Q} = \begin{bmatrix} -0.5 & 2.5 & 1 & 0.75 & 1 & 1.25 \\ 0 & 0 & 2.5 & 1.25 & 0.75 & 1.25 \end{bmatrix}.$$

It is easy to verify that Q and $\widetilde{Q}$ belong to $\mathcal{M}_+(\Delta)$. However, there exists no continuous path, completely contained in $\mathcal{M}_+(\Delta)$, which connects Q and $\widetilde{Q}$.

One way to avoid this problem is by considering only triangulations without holes, but we do not pursue this direction, and conversely, we take a slightly more practical point

of view. Rather than asking which oriented meshes can be generated from an oriented connectivity complex, we start from a given oriented reference mesh (represented by vertex coordinates Q_{ref}) and ask which other oriented meshes can be obtained through continuous deformations within $\mathcal{M}_+(\Delta)$?

This leads us to state the following definition.

Definition 4.3.10. *Suppose that $Q_{\text{ref}} \in \mathcal{M}_+(\Delta)$ are the coordinates of a given reference mesh such that its associated abstract simplicial complex is Δ. We define the* ***manifold of planar triangular meshes*** *as*

$$\mathcal{M}_+(\Delta; Q_{\text{ref}}) \coloneqq \left\{ Q \in \mathcal{M}_+(\Delta) \;\middle|\; \begin{array}{l} \text{there exists a continuous path} \\ \text{in } \mathcal{M}_+(\Delta) \text{ from } Q_{\text{ref}} \text{ to } Q \end{array} \right\}. \tag{4.20}$$

The set $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is our primary object of interest. Let us summarize some essential properties.

Theorem 4.3.11. *The set $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is*

(a) a path component of $\mathcal{M}_+(\Delta)$, and thus path-connected,
(b) an open submanifold of $\mathbb{R}^{2\times N_V}$.

Proof. Statement (a) is immediate from the definition of $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. Moreover, since $\mathcal{M}_+(\Delta)$ is locally path connected, its path components are open. Since $\mathcal{M}_+(\Delta)$ is itself open in $\mathbb{R}^{2\times N_V}$ by proposition 4.3.8, statement (b) follows immediately. We refer the reader to Lee, 2011, Ch. 4 for a background on components and path components of manifolds. □

Remark 4.3.12. *Notice that as a submanifold, $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ inherits the Hausdorff and second countability properties of $\mathbb{R}^{2\times N_V}$.*

As in the case of the manifold $\mathcal{M}_0(\Delta)$, the manifold of triangular meshes $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ can also be endowed with the smooth structure inherited from $\mathbb{R}^{2\times N_V}$ as described in the following remark.

Remark 4.3.13. *Since $\mathbb{R}^{2\times N_V}$ is a smooth manifold, then $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ has a natural smooth structure consisting of the single chart $(\mathcal{M}_+(\Delta; Q_{\text{ref}}), \mathrm{id}|_{\mathcal{M}_+(\Delta;Q_{\text{ref}})})$ (cf., Lee, 2018, App. A, p. 375).*

Having proved $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is a smooth submanifold of $\mathbb{R}^{2\times N_V}$, we are ready to identify its tangent space. The tangent space at $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ will be denoted by $\mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$ and we remark that it is a vector space isomorphic to $\mathbb{R}^{2\times N_V}$, as a result of proposition A.1.4. Briefly, this proposition states that the differential of the inclusion map is an isomorphism. This relation allows us to identify the tangent space of the open submanifold with the tangent space of the smooth manifold. In our context, this means a tangent vector to an element of the manifold of planar triangular meshes $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ can be visualized as a collection of vectors in $\mathbb{R}^2$, attached to each node of the mesh. Analogously, the tangent vectors to $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ can be understood as discrete vector fields over the resulting triangulation, as described by Kilian, Mitra, Pottmann, 2007. See figure 4.19 for an illustration.

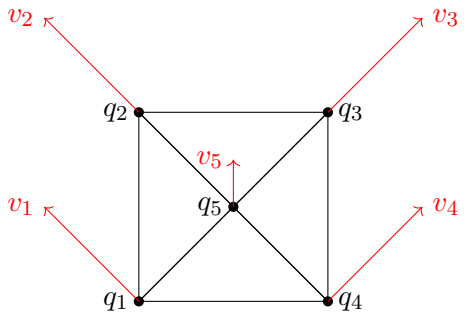


Figure 4.19. Example of a tangent vector $V = [v_1, v_2, v_3, v_4, v_5] \in \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$ at $Q = [q_1, q_2, q_3, q_4, q_5] \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$.

5 Complete Metrics for the Manifold of Triangular Meshes

Contents

Establishing the set of admissible shapes is not enough for the numerical solution of discrete PDE-constrained shape optimization problems. Additionally, one needs to endow this set with more structure, namely a Riemannian metric. In a few words, a Riemannian metric plays the role of a scalar product for the tangent space of a smooth manifold at every point. This in turn, allows us to naturally define notions like the length of tangent vectors, angles between tangent vectors, transform derivatives into gradients and more importantly, a Riemannian metric allows us to compute the equivalent of straight lines on a manifold, i. e., geodesics. From all the possible choices of Riemannian metrics which can be considered for a certain manifold, we are interested on the kind of metrics called complete, mainly because they render geodesics which can be extended infinitely without leaving the manifold.

The main aim of this chapter is the construction of two complete metrics on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. To this end, first, in section 5.1 we study two previously proposed Riemannian metrics: the Euclidean one and a metric associated to the linear elasticity problem. Section 5.2 aims to introduce complete Riemannian metric for $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ (taken from Herzog, Loayza-Romero, 2020) based on a function designed to avoid all possible kinds of self-intersections an element of $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ may be subjected to. In view of the applications to shape optimization, a second metric is proposed in section 5.3. This metric, besides of being complete, is invariant under uniform mesh refinements, and its presentation is based on the results obtained in Herzog, Loayza-Romero, 2021. Section 5.4 is devoted to the description of the theoretical and numerical methods used for the approximation of the geodesic equation's solution, exploiting the simple structure of the proposed complete metrics. We end this chapter by presenting some numerical investigations which aim to show how geodesics on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ behave. The results are collected in section 5.5.

Let us start by recalling that a Riemannian metric on the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is a correspondence which associates to each point $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ an inner product (symmetric, bilinear and positive-definite form) $(\cdot\,, \cdot)_Q \colon \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}}) \times \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}}) \to \mathbb{R}$ which varies smoothly from point to point. We refer the reader to definition A.2.1 for the formal introduction of this notion.

As previously mentioned, we are interested in geodesically complete metrics, which in a few words, are such that every maximal geodesic is defined for all $t \in \mathbb{R}$. Given a Riemannian metric, it is relatively simple to verify if it is complete or not. However, the construction of complete metrics on smooth manifolds is not as straightforward. Fortunately, Gordon,

1973, Thm. 1 describes a recipe on how to construct complete metrics on smooth manifolds. The main results of this chapter are based on this theorem.

Theorem (Gordon, 1973, Thm. 1). *Suppose that $\mathcal{M}$ is a connected manifold of class $\mathcal{C}^3$, endowed with a (not necessarily complete) Riemannian metric $\widetilde{g}$ with component functions $\widetilde{g}_{ab}$. If $f\colon \mathcal{M} \to \mathbb{R}$ is any proper function of class $\mathcal{C}^3$, then the Riemannian metric g defined by*

$$g_{ab} = \widetilde{g}_{ab} + \frac{\partial f}{\partial x^a}\frac{\partial f}{\partial x^b} \tag{5.1}$$

is geodesically complete.

We recall a function $f\colon \mathcal{M} \to \mathbb{R}$ is said to be proper if the preimages $f^{-1}(K)$ of compact sets $K \subset \mathbb{R}$ are compact in $\mathcal{M}$.

5.1 Previously Proposed Metrics

This section is devoted to the presentation of two Riemannian metrics on the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. They had been used in the literature before, even without the notion of the manifold of planar triangular meshes. In subsection 5.1.1 we introduce the Euclidean metric, by taking advantage of $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ being an open submanifold of $\mathbb{R}^{2\times N_V}$. We also present a numerical example that demonstrates the incompleteness of this metric. Subsection 5.1.2 aims to show that the bilinear form associated with the linear elasticity equation can be used to define a Riemannian metric for $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. The geodesic equation associated with this metric is not so easy to express, and thus to solve. Therefore, we cannot easily check whether it is a complete metric or not. This is definitely an interesting research question, which can be pursued in the future. However, it is outside the scope of this thesis.

Throughout, we consider the vectorization operation, which we denote as $\operatorname{vec}\colon \mathbb{R}^{2\times N_V} \to \mathbb{R}^{2\,N_V}$ stacks $Q, V \in \mathbb{R}^{2\times N_V}$ column by column.

5.1.1 Euclidean Metric

We start by recalling that $\mathbb{R}^{2\times N_V}$ can be endowed with the Euclidean Riemannian metric given by the following expression:

$$(V\,,\,\widetilde{V})_Q^{\text{Euc}} = (\operatorname{vec} V)\cdot(\operatorname{vec}\widetilde{V}) \tag{5.2}$$

for all $Q \in \mathbb{R}^{2\times N_V}$ and $V, \widetilde{V} \in \mathcal{T}_Q\mathbb{R}^{2\times N_V}$. Since every submanifold of a Riemannian manifold automatically inherits a Riemannian metric, (we refer to Lee, 2018, Ch. 2, p. 15 for more details), the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ automatically inherits the Riemannian metric $(\cdot\,,\,\cdot)_Q^{\text{Euc}}$, restricted to vectors which are tangent only to $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$.

Geodesic curves with respect to $(\cdot\,,\,\cdot)_Q^{\text{Euc}}$ are trivial to compute. They simply consist of a collection of straight lines, each one emanating from a vertex and traveling along with constant Euclidean velocity. More precisely, given a point $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ and a tangent vector $V \in \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$, the unique geodesic curve passing through Q with velocity V, at time zero, can be described as $Q + tV$. See example A.2.4 for a more detailed explanation. Clearly, this shows that geodesics w.r.t. the Euclidean metric generally cease to exist in $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ after a finite time. This allows us to conclude that the Euclidean metric is not complete for $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. Figure 5.2 shows three snapshot of the geodesic w.r.t. the Euclidean metric for a 5-nodes mesh, in which we observe that after $t = 0.8$, the mesh does no longer belong to $\mathcal{M}_+(\Delta; Q_{\text{ref}})$.

5.1.2 Linear Elasticity Metric

One can also choose the Riemannian metric whose matrix representation coincides with the bilinear form associated with the Lamé system of linear elasticity. For a point $Q \in$

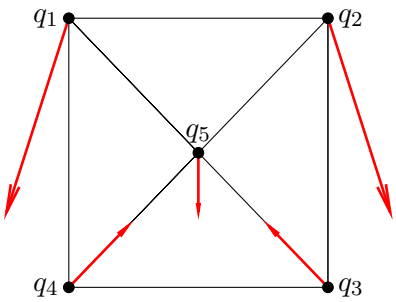
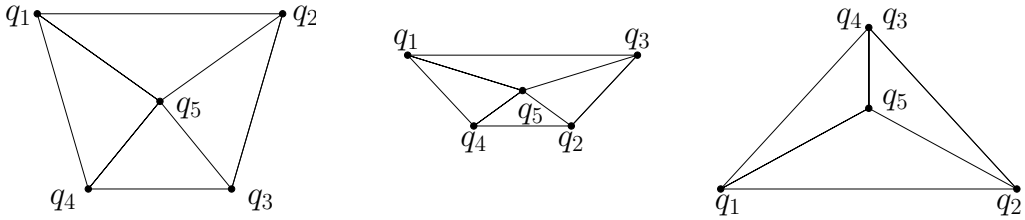


Figure 5.1. Initial tangent vector and mesh used as an example for the computation of geodesics on the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ of planar triangular meshes.



$t = 0.5$ $t = 0.8$ $t = 2.0$

Figure 5.2. Snapshots of a geodesic on the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ of planar triangular meshes with respect to the Euclidean Riemannian metric. After $t = 0.8$ the mesh degenerates.

$\mathcal{M}_+(\Delta; Q_{\text{ref}})$ and tangent vectors $V, \widetilde{V} \in \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$, the matrix representation of this metric w.r.t. the vec chart is as follows:

$$(V\,,\,\widetilde{V})_Q^{\text{elas}} \coloneqq (\operatorname{vec} V) \cdot \mathbb{K}\,(\operatorname{vec} \widetilde{V}) + \delta\,(\operatorname{vec} V) \cdot \mathbb{M}\,(\operatorname{vec} \widetilde{V}). \tag{5.3}$$

The matrix $\mathbb{K}$ is the finite element stiffness matrix, for piecewise linear elements over the mesh defined by Q, associated with the linear elasticity operator

$$2\mu \int_{\Omega_Q} \boldsymbol{\varepsilon}(\boldsymbol{v}) : \boldsymbol{\varepsilon}(\widetilde{\boldsymbol{v}})\,\mathrm{d}x + \lambda \int_{\Omega_Q} \operatorname{trace}(\boldsymbol{\varepsilon}(\boldsymbol{v}))\,\operatorname{trace}(\boldsymbol{\varepsilon}(\widetilde{\boldsymbol{v}}))\,\mathrm{d}x, \tag{5.4}$$

where $\boldsymbol{\varepsilon}(\boldsymbol{v}) = (D\boldsymbol{v} + D\boldsymbol{v}^{\mathrm{T}})/2$ is the linearized strain tensor, and D denotes the derivative (Jacobian) of a vector valued function. The constants λ, μ, are called the Lamé parameters. The parameter $\delta > 0$ is a damping parameter and it is required to ensure the metric is positive definite, since we do not have a clamping boundary. Moreover, $\mathbb{M}$ is the mass matrix.

This metric can be understood as the discrete counterpart of the Steklov-Poincaré metric proposed in Schulz, Siebenborn, Welker, 2016, for the infinite-dimensional manifold B_e introduced by Michor, Mumford, 2005.

In what follows, we present the result which guarantees that the bilinear form presented in (5.3) is indeed a Riemannian metric of $\mathcal{M}_+(\Delta; Q_{\text{ref}})$.

Impending interior self-intersections.

Impending exterior self-intersections.

Figure 5.3. Different kinds of impending self-intersections.

Theorem 5.1.1. *Let us consider $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$, and the bilinear form defined in (5.3), with Lamé parameters $\mu > 0$, $\lambda + \mu > 0$ and damping parameter $\delta > 0$. Then, $(\cdot\,,\cdot)_Q^{\text{elas}}$ is a Riemannian metric on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$.*

Proof. Recalling the definition of a Riemannian metric given in definition A.2.1, we need to prove that the local representation of$(\cdot\,,\cdot)_Q^{\text{elas}}$ (in the vec chart) is a symmetric, bilinear and positive definite form, which varies smoothly from point to point on the manifold. It is well-known that the matrix associated to the linear elasticity equation is bilinear, symmetric and positive definite, provided that $\mu > 0$, $\lambda + \mu > 0$ and $\delta > 0$. Therefore, it remains to be proved that the local representation of the metric is smooth, i. e., the matrices $\mathbb{K}$ and $\mathbb{M}$, as a functions of the node positions, are smooth for all $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$.

We recall that by means of the finite element method (see section 2.3), the stiffness and mass matrices can be expressed in terms of the *local stiffness and mass matrices*. These local contributions depend only on the affine transformation T_K given in (2.23), the substitution rule (2.29) and the node positions Q. Finally, we recall the connectivity information of the mesh, given by Δ, remains constant within $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. Thus, the mapping T_K is smooth for all $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ and this holds for all triangles of the mesh. □

Having proved the bilinear form of the linear elasticity operator is indeed a Riemannian metric, one could study its associated geodesic. This will involve the computation of the derivatives of the stiffness and mass matrices w.r.t. the node positions; however, we do not pursue this direction.

5.2 Quality Preserving Metrics

Now, we present the main topic of this chapter, i. e., the construction of a complete metric for $\mathcal{M}_+(\Delta; Q_{\text{ref}})$, based on Gordon, 1973, Thm. 1. By recalling the statement of the theorem, we need to find a $\mathcal{C}^3$ and proper function on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$.

Intuitively, it is clear that the local representations of the metric must be large whenever the mesh is close to a situation of self-intersection. These self-intersections can be of internal or external nature, see figure 5.3. Internal self-intersections are impending whenever a triangle is about to collapse to a line segment or even a point. Exterior self-intersections can be recognized by the distance of one boundary face to another boundary face becoming small (the definition of boundary and interior faces can be found in subsection 4.2.1). These situations, and the metric which prevents them, can be expressed in terms of the heights of the triangle (for internal self-intersections) and the distance of nonincident boundary vertices and edges (for external self-intersections).

We recall the different terminology introduced in section 4.2. Let Δ be a connectivity complex as in definition 4.2.5, and $Q \in \mathcal{M}_+(\Delta)$. Then, we refer as $0,1,2$-faces of Δ to the sets of indices $[i_0]$, $[i_0, i_1]$ and $[i_0, i_{1,2}]$ elements of Δ of cardinality $0,1,2$, respectively. On the other hand, according to the matrix Q the vertex associated with the 0-face $[i_0]$ is given by $\{q_{i_0}\}$. In the same way, the edge associated with the 1-face $[i_0, i_1]$ is given by $\mathrm{conv}_Q(i_0, i_1)$, and the triangle $\mathrm{conv}_Q(i_0, i_1, i_2)$.

From the definition of the signed area $A_Q(i_0, i_1, i_2)$ given in (4.2) and rearranging the terms of (4.3), we consider

$$h_Q^\ell(i_0, i_1, i_2) = \frac{2\, A_Q(i_0, i_1, i_2)}{E_Q^\ell(i_0, i_1, i_2)}, \quad \ell = 0, 1, 2, \tag{5.5}$$

be the **ℓ-th height** (the one through the ℓ-th vertex).

Notice that the signed areas A_Q are positive for $Q \in \mathcal{M}_+(\Delta)$ and thus all heights in (5.5) are positive as well.

We are now in the position to define a preliminary proper function f_1 which is going to help render the shape manifold $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ geodesically complete.

Definition 5.2.1. *We denote by V_∂ the set of the boundary 0-faces and by E_∂ the set of boundary 1-faces. Suppose that the 2-faces in Δ are numbered from 1 to N_T. We define the function $f_1 \colon \mathcal{M}_+(\Delta) \to \mathbb{R}$ by*

$$f_1(Q; Q_{\mathrm{ref}}) \coloneqq \sum_{k=1}^{N_T} \sum_{\ell=0}^{2} \frac{\alpha_1}{h_Q^\ell\big(i_0^k, i_1^k, i_2^k\big)} + \sum_{[j_0,j_1]\in E_\partial} \sum_{\substack{i_0 \in V_\partial \\ i_0 \neq j_0, j_1}} \frac{\alpha_2}{D_Q(i_0; [j_0, j_1])} + \frac{\alpha_3}{2} \|Q - Q_{\mathrm{ref}}\|_F^2. \tag{5.6}$$

Here $\alpha_1, \alpha_2, \alpha_3$ are nonnegative parameters, h_Q^ℓ are the heights given in (5.5) and D_Q is the 1-norm based distance of a vertex to an edge given in (4.13). Moreover, $Q_{\mathrm{ref}} \in \mathcal{M}_+(\Delta)$ serves as a reference configuration and $\|\cdot\|_F$ denotes the Frobenius norm.

We remark that the first term in (5.6) is designed to avoid interior self-intersections, which go along with at least one height in a triangle converging to zero. The second term avoids exterior self-intersections. The third term penalizes large deviations from the reference mesh. All three terms are required in order to show the properness of f_1 in theorem 5.2.6; but first, we study several properties of f_1.

Lemma 5.2.2. *For any choice of $\alpha_1, \alpha_2, \alpha_3 \geq 0$, the function f_1 defined in (5.6) is well-defined on $\mathcal{M}_+(\Delta)$ and continuous with values in $(0, \infty)$.*

Proof. From the definition of $\mathcal{M}_+(\Delta)$ it is clear that all areas and the lengths of the edges are strictly positive and thus the same is true for the heights in (5.5). In the same way, all the distances from a vertex to an edge given in (4.13) are positive. The continuity of f_1 w.r.t. Q on $\mathcal{M}_+(\Delta)$ is obvious. □

The major next step is to prove the properness of f_1 defined in (5.6). In order to make the proof more readable we present some intermediate results. The first one shows important bounds on the heights h_Q^ℓ (5.5), edge lengths E_Q^ℓ (4.1), signed area A_Q (4.2), inradius r_Q and circumradius R_Q (see subsection 4.2.3) of the triangles associated to the 2-faces of Δ in terms of the value of the function f_1. Moreover, we prove bounds on the interior angles θ_Q^ℓ of the triangles.

Lemma 5.2.3. *Let $[i_0, i_1, i_2]$ be an arbitrary 2-face of Δ. Suppose that $\alpha_1, \alpha_2 \geq 0$ and $\alpha_3 > 0$ holds and $Q \in \mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$. Then the following statements hold.*

(a) The heights satisfy $h_Q^\ell(i_0, i_1, i_2) \geq \dfrac{\alpha_1}{f_1(Q; Q_{\mathrm{ref}})}$, for all $\ell = 0, 1, 2$.

(b) The inradius satisfies $r_Q(i_0,i_1,i_2) \geq \dfrac{\alpha_1}{f_1(Q;Q_{\mathrm{ref}})}$.

(c) The edge lengths satisfy $\dfrac{2\alpha_1}{f_1(Q;Q_{\mathrm{ref}})} \leq E^\ell_Q(i_0,i_1,i_2) \leq 2\sqrt{\dfrac{f_1(Q;Q_{\mathrm{ref}})}{\alpha_3}} + \sqrt{2}\,\|Q_{\mathrm{ref}}\|_F$, *for all* $\ell = 0,1,2$.

(d) The area satisfies $A_Q(i_0,i_1,i_2) \geq \dfrac{\pi\alpha_1^2}{f_1^2(Q;Q_{\mathrm{ref}})}$.

(e) The inradius and the circumradius satisfy

$$\frac{r_Q(i_0,i_1,i_2)}{R_Q(i_0,i_1,i_2)} \geq \frac{4\pi\alpha_1^3}{\left(2\sqrt{\dfrac{f_1(Q;Q_{\mathrm{ref}})}{\alpha_3}} + \sqrt{2}\,\|Q_{\mathrm{ref}}\|_F\right)^3} \frac{1}{f_1(Q;Q_{\mathrm{ref}})^3}.$$

(f) There exists a function $\Psi\colon (0,\infty) \to \mathbb{R}$ *which is monotone increasing and takes values in* $[0,1)$ *such that*

$$\left|\cos\left(\theta^\ell_Q(i_0,i_1,i_2)\right)\right| \leq \Psi(f_1(Q;Q_{\mathrm{ref}})) \quad \text{for all } \ell = 0,1,2.$$

In particular, $\left|\cos\left(\theta^\ell_Q(i_0,i_1,i_2)\right)\right| \leq \Psi(b) < 1$ *holds whenever* $f_1(Q;Q_{\mathrm{ref}}) \leq b$.

Proof. We start by recalling that since $Q \in \mathcal{M}_+(\Delta;Q_{\mathrm{ref}})$, then all heights are strictly positive and we have

$$\frac{\alpha_1}{h^\ell_Q(i_0,i_1,i_2)} \leq \sum_{k=1}^{N_T}\sum_{\ell=0}^{2} \frac{\alpha_1}{h^\ell_Q\left(i^k_0,i^k_1,i^k_2\right)} \leq f_1(Q;Q_{\mathrm{ref}}),$$

which proves statement (a). Denoting by $r_Q(i_0,i_1,i_2)$ the inradius of a 2-face of Δ and using the relation between the inradius and the reciprocal of the heights given in (4.9), we obtain that

$$\frac{\alpha_1}{r_Q(i_0,i_1,i_2)} \leq \sum_{k=1}^{N_T}\sum_{\ell=0}^{2} \frac{\alpha_1}{h^\ell_Q\left(i^k_0,i^k_1,i^k_2\right)} \leq f_1(Q;Q_{\mathrm{ref}}),$$

which proves statement (b).

It is also well known that for any triangle the length of each of its edges is greater than twice the inradius, from where we obtain the first inequality of statement (c):

$$\frac{2\alpha_1}{f_1(Q;Q_{\mathrm{ref}})} \leq 2\,r_Q(i_0,i_1,i_2) \leq E^\ell_Q(i_0,i_1,i_2).$$

For the second inequality of statement (c), we use the triangle inequality and the definition of f_1 to obtain

$$\begin{aligned} E^\ell_Q(i^k_0,i^k_1,i^k_2) &= \left\|q_{i^k_{\ell\oplus1}} - q_{i^k_{\ell\oplus2}}\right\|_2 \leq \left\|q_{i^k_{\ell\oplus1}}\right\|_2 + \left\|q_{i^k_{\ell\oplus2}}\right\|_2 \leq \sqrt{2}\,\|Q\|_F \\ &\leq \sqrt{2}\,\|Q - Q_{\mathrm{ref}}\|_F + \sqrt{2}\,\|Q_{\mathrm{ref}}\|_F \leq 2\sqrt{\frac{f_1(Q;Q_{\mathrm{ref}})}{\alpha_3}} + \sqrt{2}\,\|Q_{\mathrm{ref}}\|_F, \end{aligned}$$

which completes the proof of statement (c). Next, we use that the area of each triangle is larger than the area of its incircle, i. e., $A_Q(i_0,i_1,i_2) \geq \pi r^2_Q(i_0,i_1,i_2)$. Using the bound in statement (b), we obtain statement (d). Statement (e) follows immediately from (4.11) and statements (b) to (d). To prove statement (f), we use (4.10) to obtain

$$\left|\cos\left(\theta^\ell(i_0,i_1,i_2)\right)\right| \leq \frac{r_Q(i_0,i_1,i_2)}{R_Q(i_0,i_1,i_2)} + \sqrt{1 - \frac{2\,r_Q(i_0,i_1,i_2)}{R_Q(i_0,i_1,i_2)}}$$

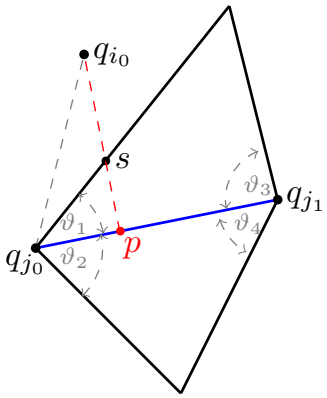


Figure 5.4. Illustration of the construction for the distance between the vertex $\{q_{i_0}\}$ and an interior edge $\mathrm{conv}_Q(j_0, j_1)$ (depicted in blue), used in lemma 5.2.4.

for any $\ell = 0, 1, 2$. Consider the function

$$\varphi(x) \coloneqq x + \sqrt{1 - 2x}$$

in the interval $[0, 1/2]$, where it is continuous and decreasing and takes values in $[0, 1]$. Its maximum value is $\varphi(0) = 1$. Thanks to statement (e), we know that

$$\frac{r_Q(i_0, i_1, i_2)}{R_Q(i_0, i_1, i_2)} \geq \frac{4\pi\alpha_1^3}{\left\{2\sqrt{\frac{f_1(Q;Q_{\mathrm{ref}})}{\alpha_3}} + \sqrt{2}\,\|Q_{\mathrm{ref}}\|_F\right\}^3} \frac{1}{f_1(Q;Q_{\mathrm{ref}})^3} =: \psi(f_1(Q;Q_{\mathrm{ref}})).$$

The function $\psi(x)\colon (0,\infty) \to (0,\infty)$ is continuous and decreasing. Since φ is also decreasing, we get

$$\left|\cos(\theta^\ell(i_0, i_1, i_2))\right| \leq \varphi\left(\frac{r_Q(i_0, i_1, i_2)}{R_Q(i_0, i_1, i_2)}\right) \leq \varphi(\psi(f_1(Q;Q_{\mathrm{ref}}))).$$

Now set $\Psi \coloneqq \varphi \circ \psi\colon (0,\infty) \to (0,1)$, which is increasing and continuous as claimed. □

In what follows, we present a lemma, which will become handy when we want to compute bounds on the 1-norm based distance of a vertex to an edge. Briefly recalling, a 1-face is interior if it is shared by two distinct 2-faces. Otherwise it belongs to only one 2-face and is referred to as a boundary 1-face. A 0-face is a called a boundary 0-face if it belongs to at least one boundary 1-face. Otherwise, all of its incident 1-faces are interior and the 0-face will be referred to as interior. These notations are also inherited by the vertices and edges associated to $\Sigma_\Delta(Q)$ when $Q \in \mathcal{M}_+(\Delta)$. With these notions in mind, lemma 5.2.4 shows that if $Q \in \mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$, the distance of a boundary 0-face $[i_0]$ from an interior 1-face $[j_0, j_1]$, when $\{q_{i_0}\} \cap \mathrm{conv}_Q(q_{j_0}, q_{j_1}) = \emptyset$, can be bounded below, provided that $\alpha_1, \alpha_2, \alpha_3 \geq 0$.

Lemma 5.2.4. *Suppose that $Q \in \mathcal{M}_+(\Delta)$. We consider a boundary 0-face, denoted by $[i_0]$ and an interior 1-face denoted by $[j_0, j_1]$ of Δ such that $\{q_{i_0}\} \cap \mathrm{conv}_Q(j_0, j_1) = \emptyset$. Suppose moreover, that $\alpha_1, \alpha_2, \alpha_3 \geq 0$, and that $\{q_{i_0}\} \cap \overline{\mathrm{St}(\mathrm{conv}_Q(j_0, j_1))} = \emptyset$, where $\overline{\mathrm{St}(\cdot)}$ denotes the closed star of a face. Then*

$$d_Q(i_0; [j_0, j_1]) \geq \min_{\vartheta \in \Theta}\left\{\sqrt{1 - \cos^2(\vartheta)}\right\} \min\{\|q_{i_0} - q_{j_0}\|_2, \|q_{i_0} - q_{j_1}\|_2\}, \tag{5.7}$$

where Θ is the set of four angles formed by the edge $\mathrm{conv}_Q(j_0, j_1)$ and the adjacent edges belonging to $\overline{\mathrm{St}(\mathrm{conv}_Q(j_0, j_1))}$; see figure 5.4.

PROOF. We start by noticing that thanks to the definition of interior edges, the closed star $\overline{\mathrm{St}(\mathrm{conv}\, Q[j_0, j_1])}$ contains exactly two triangles. Let us denote by p the orthogonal projection of q_{i_0} onto the infinite line defined by the edge $\mathrm{conv}_Q(j_0, j_1)$. We distinguish two cases. • If p does not belong to $\mathrm{conv}_Q(j_0, j_1)$, then by the definition of the distance of a vertex to an edge, we get $d_Q(i_0; [j_0, j_1]) = \min\{\|q_{i_0} - q_{j_0}\|_2, \|q_{i_0} - q_{j_1}\|_2\}$. • Conversely, if p does belong to $\mathrm{conv}_Q(j_0, j_1)$, then w.l.o.g. we assume $\|q_{i_0} - q_{j_0}\|_2 \leq \|q_{i_0} - q_{j_1}\|_2$.

Furthermore, we denote by s the unique point of intersection between the line segment joining $\{q_{i_0}\}$ and p with one of edges of $\overline{\mathrm{St}(\mathrm{conv}_Q(j_0, q_{j_1}))}$. See figure 5.4 for an illustration of this notation. It is clear that $d_Q(i_0; [j_0, j_1]) = \|q_{i_0} - p\|_2$. Now, by Pythagoras' Theorem we have that $\|q_{i_0} - p\|_2^2 = \|q_{i_0} - q_{j_0}\|_2^2 - \|q_{j_0} - p\|_2^2$. Moreover, we know that $\cos(\vartheta_1) = \|q_{j_0} - p\|_2 / \|q_{j_0} - s\|_2$, which implies $\|q_{i_0} - p\|_2^2 = \|q_{i_0} - q_{j_1}\|_2^2 - \cos^2(\vartheta_1)\, \|q_{j_0} - s\|_2^2$.

Since we have assumed $\{q_{i_0}\} \cap \overline{\mathrm{St}(\mathrm{conv}_Q(j_0, j_1))} = \emptyset$, we have $\|q_{i_0} - p\|_2 > \|s - p\|_2$, and using again Pythagoras' Theorem we obtain

$$\begin{aligned} \|q_{j_0} - s\|_2^2 &= \|q_{j_0} - p\|_2^2 + \|p - s\|_2^2 \\ &< \|q_{j_0} - p\|_2^2 + \|p - q_{i_0}\|_2^2 = \|q_{j_0} - q_{i_0}\|_2^2. \end{aligned}$$

Thus,

$$\begin{aligned} d_Q(i_0; [j_0, j_1])^2 &= \|q_{i_0} - p\|_2^2 > \|q_{j_0} - q_{i_0}\|_2^2 - \cos^2(\vartheta_1)\, \|q_{j_0} - q_{i_0}\|_2^2 \\ &= (1 - \cos^2(\vartheta_1))\, \|q_{j_0} - q_{i_0}\|_2^2. \end{aligned}$$

Thanks to the assumption $\|q_{i_0} - q_{j_0}\|_2 \leq \|q_{i_0} - q_{j_1}\|_2$ we can conclude that

$$\begin{aligned} d_Q(i_0; [j_0, j_1]) &> \sqrt{1 - \cos^2(\vartheta_1)}\, \min\{\|q_{i_0} - q_{j_0}\|_2, \|q_{i_0} - q_{j_1}\|_2\} \\ &\geq \min_{\vartheta \in \Theta} \left\{ \sqrt{1 - \cos^2(\vartheta)} \right\} \min\{\|q_{i_0} - q_{j_0}\|_2, \|q_{i_0} - q_{j_0}\|_2\}. \end{aligned}$$

Notice that $\sqrt{1 - \cos^2(\vartheta)} < 1$ holds for all $\vartheta \in \Theta$, whether or not the orthogonal projection of q_{i_0} onto the infinite line defined by $\mathrm{conv}_Q(j_0, j_1)$ belongs to $\mathrm{conv}_Q(j_0, j_1)$. We can thus conclude

$$d_Q(i_0; [j_0, j_1]) \geq \min_{\vartheta \in \Theta} \left\{ \sqrt{1 - \cos^2(\vartheta)} \right\} \min\{\|q_{i_0} - q_{j_0}\|_2, \|q_{i_0} - q_{j_1}\|_2\}. \qquad \square$$

As already anticipated, the next result establishes bounds on the distance of a vertex to an edge in terms of the function f_1. Here we distinguish between boundary and interior 0- and 1-faces, defined in section 4.2. Moreover, recall that the intersection of the half spaces generated by the edges which belong to the link of an interior vertex is completely contained in the closed start of the same vertex (cf., lemma 4.2.4). This result allows us to find a separating hyperplane which will keep the distance of a boundary vertex to an interior edge strictly positive.

Proposition 5.2.5. *Suppose that $Q \in \mathcal{M}_+(\Delta)$. We consider a 0-face $[i_0]$ and a 1-face $[j_0, j_1]$ of Δ such that $\{q_{i_0}\} \cap \mathrm{conv}_Q(j_0, j_1) = \emptyset$. Suppose $\alpha_1, \alpha_2, \alpha_3 \geq 0$. Then the following statements hold.*

(a) If $[i_0]$ and $[j_0, j_1]$ are boundary, then

$$D_Q(i_0; [j_0, j_1]) \geq \frac{\alpha_2}{f_1(Q; Q_{\mathrm{ref}})}. \tag{5.8}$$

(b) If $[i_0]$ is interior and $[j_0, j_1]$ is interior or boundary, then

$$D_Q(i_0; [j_0, j_1]) \geq \frac{\alpha_1}{f_1(Q; Q_{\mathrm{ref}})}. \tag{5.9}$$

(c) If $[i_0]$ *is boundary and* $[j_0, j_1]$ *is interior, then*

$$D_Q(i_0; [j_0, j_1]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{\sqrt{2} f_1(Q; Q_{\text{ref}})} \min_{\vartheta \in \Theta} \left\{ \sqrt{1 - \cos^2(\vartheta)} \right\}, \tag{5.10}$$

where Θ *is the set of four angles formed by the edge* $\text{conv}_Q(j_0, j_1)$ *and the adjacent edges belonging to* $\overline{\text{St}(\text{conv}_Q(j_0, j_1))}$*; see figure 5.4.*

PROOF. Let $Q \in \mathcal{M}_+(\Delta)$, we consider a 0-face $[i_0]$ and a 1-face $[j_0, j_1]$ of Δ such that $\{q_{i_0}\} \cap \text{conv}_Q(j_0, j_1) = \emptyset$. First of all, the distance $D_Q(i_0; [j_0, j_1])$ defined in (4.13) is strictly positive, thanks to the assumption $\{q_{i_0}\} \cap \text{conv}_Q(j_0, j_1) = \emptyset$. Along this proof, we are going to use the notions introduced in the end of subsection 4.2.1, specifically the link $\text{lk}(\cdot)$, and the closed star $\overline{\text{St}(\cdot)}$ of a 0- or 1-face.

We consider the cases following the statement of proposition 5.2.5.

$[i_0]$ **and** $[j_0, j_1]$ **are boundary faces:** We estimate

$$\frac{\alpha_2}{D_Q(i_0; [j_0, j_1])} \leq \sum_{[\ell_0, \ell_1] \in E_\partial} \sum_{\substack{k_0 \in V_\partial \\ k_0 \neq \ell_0, \ell_1}} \frac{\alpha_2}{D_Q(k_0; [\ell_0, \ell_1])} \leq f_1(Q; Q_{\text{ref}}),$$

which proves (5.8) and thus statement (a).

$[i_0]$ **is an interior 0-face and** $[j_0, j_1]$ **is interior or boundary 1-face:** • We first consider the case $\text{conv}_Q(j_0, j_1) \cap \overline{\text{St}(\{q_{i_0}\})} = \emptyset$. We denote by $\text{conv}_Q(j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})$ the ℓ-the edge belonging to $\text{lk}(\{q_{i_0}\})$. Moreover, $H^+\left(\text{conv}_Q(j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})\right)$ denotes the half-space generated by the edge $\text{conv}_Q(j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})$ which contains $\{q_{i_0}\}$. Using lemmas 4.2.3 and 4.2.4, we know there exists at least one edge $\text{conv}_Q(j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})$ which separates $\text{conv}_Q(j_0, j_1)$ from $\{q_{i_0}\}$ since $\text{conv}_Q(j_0, j_1) \cap H^+\left(\text{conv}_Q(j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})\right) = \emptyset$ (see figure 5.5 for an illustration).
Consider the triangle uniquely identified by $\text{conv}_Q(j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})$ and vertex $\{q_{i_0}\}$ and denote by $h_Q^{i_0}(i_0, j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})$ the height of this triangle passing through $\{q_{i_0}\}$. Then,

$$d_Q(i_0; [j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell}]) \geq h_Q^{i_0}(i_0, j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell}).$$

Since $\text{conv}_Q(j_0, j_1) \cap H^+\left(\text{conv}_Q(j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})\right) = \emptyset$ holds, this implies

$$\begin{aligned} D_Q(i_0; [j_0, j_1]) &\geq d_Q(i_0; [j_0, j_1]) > d_Q(i_0; [j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell}]) \\ &\geq h_Q^{i_0}(i_0, j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell}) \geq \frac{\alpha_1}{f_1(Q; Q_{\text{ref}})}, \end{aligned}$$

where the last inequality follows from lemma 5.2.3, statement (a). • We now consider the case when $\text{conv}_Q(j_0, q_{j_1}) \cap \overline{\text{St}(\{q_{i_0}\})} \neq \emptyset$. Since $\{q_{i_0}\}, \text{conv}_Q(j_0, j_1) \in \Sigma_\Delta$ and $\{q_{i_0}\} \cap \text{conv}_Q(j_0, j_1) = \emptyset$ by assumption, we have two possibilities. First, $\text{conv}_Q(j_0, j_1) \cap \overline{\text{St}(\{q_{i_0}\})}$ is a vertex. We assume w.l.o.g. this vertex is $\{q_{j_0}\}$, therefore it belongs to some $\text{conv}_Q(j_0, j_1^{\text{lk},\ell})$; see figure 5.6 for such a construction. Using Pythagoras' Theorem we get

$$d_Q(i_0; [j_0, j_1]) = \|q_{i_0} - q_{j_0^{\text{lk},\ell}}\|_2 \geq d_Q(i_0; [j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell}]) \geq h_Q^{i_0}(i_0, j_0, j_1^{\text{lk},\ell}),$$

which implies that $D_Q(i_0; [j_0, j_1]) \geq d_Q(i_0; [j_0, j_1]) \geq \frac{\alpha_1}{f_1(Q; Q_{\text{ref}})}$ by lemma 5.2.3, statement (a). • Second, $\text{conv}_Q(j_0, j_1) \cap \overline{\text{St}(\{q_{i_0}\})} = \text{conv}_Q(j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})$. Notice that $D_Q(i_0; [j_0, j_1]) \geq d_Q(i_0; [j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell}]) \geq h_Q^{i_0}(i_0, j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})$, where $h_Q^{i_0}(i_0, j_0^{\text{lk},\ell}, j_1^{\text{lk},\ell})$ is

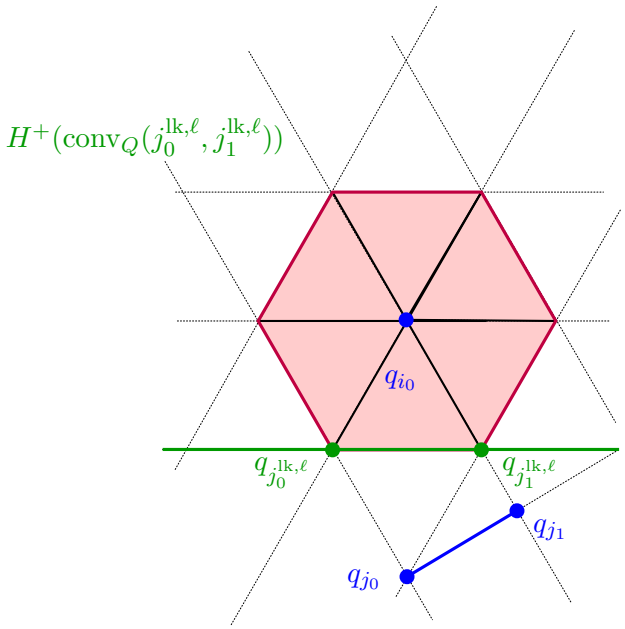


Figure 5.5. Illustration for the proof of proposition 5.2.5: interior vertex, arbitrary edge and empty intersection between closed star of vertex and edge. $\overline{\mathrm{St}(\{q_{i_0}\})}$ is shaded in red. Moreover, $\mathrm{lk}(\{q_{i_0}\})$ are shown by dark red lines.

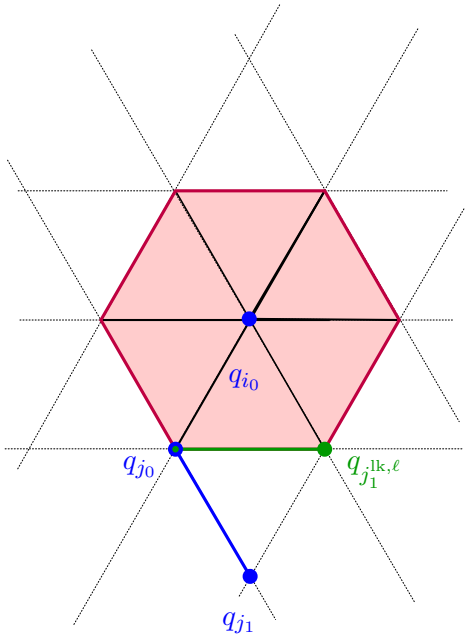


Figure 5.6. Illustration for the proof of proposition 5.2.5: interior vertex, arbitrary edge and nonempty intersection between closed star of vertex and edge. $\overline{\mathrm{St}(\{q_{i_0}\})}$ is shaded in red. Moreover, $\mathrm{lk}(\{q_{i_0}\})$ are shown by dark red lines.

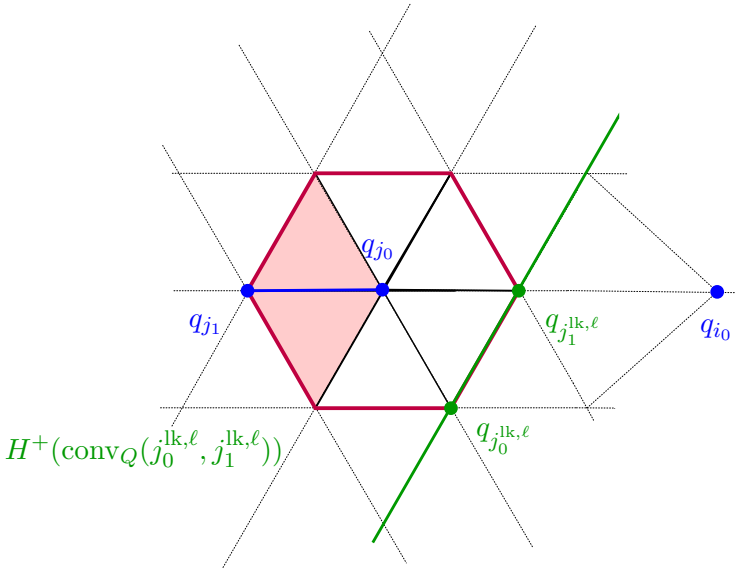


Figure 5.7. Illustration for the proof of proposition 5.2.5: interior edge, boundary vertex and nonempty intersection between vertex and closed start of one of the edge's vertices. $\overline{\mathrm{St}(\mathrm{conv}_Q(j_0 j_1))}$ is shaded in red. Moreover, $\mathrm{lk}(\{q_{j_1}\})$ are shown by dark red lines.

the height of the 2-face uniquely defined by $\{q_{i_0}\}$ and $\mathrm{conv}_Q(j_0^{\mathrm{lk},\ell}, j_1^{\mathrm{lk},\ell})$, which passes through $\{q_{i_0}\}$. Thanks to lemma 5.2.3, statement (a) we have

$$D_Q(i_0; [j_0, j_1]) \geq \frac{\alpha_1}{f_1(Q; Q_{\mathrm{ref}})}.$$

Altogether, we have proved (5.9) and thus statement (b).

$[i_0]$ **is a boundary 0-face and** $[j_0, j_1]$ **is an interior 1-face:** • We first consider the case $\{q_{i_0}\} \cap \overline{\mathrm{St}(\mathrm{conv}_Q(j_0, j_0))} = \emptyset$. In order to estimate $D_Q(i_0; [j_0, j_1])$, we are going to invoke lemma 5.2.4, which separates the task into the estimation of angles and the estimation of $\|q_{i_0} - q_{j_0}\|_2$ and $\|q_{i_0} - q_{j_1}\|_2$. We begin with the latter and assume, without loss of generality, that $\|q_{i_0} - q_{j_0}\|_2 \leq \|q_{i_0} - q_{j_1}\|_2$. Therefore, we focus on the estimation of $\|q_{i_0} - q_{j_0}\|_2$. We consider now two cases.

$\{q_{j_0}\}$ **is an interior vertex:** Since $\{q_{i_0}\} \cap \overline{\mathrm{St}(\mathrm{conv}_Q(j_0, j_1))} = \emptyset$, there exist two possibilities. • Suppose first that $\{q_{i_0}\} \cap \overline{\mathrm{St}(\{q_{j_0}\})} = \emptyset$ holds. Since $\{q_{j_0}\}$ is an interior vertex, then thanks to lemma 4.2.3 we know that $\mathrm{lk}(\{q_{j_0}\})$ is a closed polygonal chain. Therefore, there exists $\mathrm{conv}_Q(j_0^{\mathrm{lk},\ell}, j_1^{\mathrm{lk},\ell}) \in \mathrm{lk}(\{q_{j_0}\})$ such that $\{q_{i_0}\} \cap H^+\Big(\mathrm{conv}_Q(j_0^{\mathrm{lk},\ell}, j_1^{\mathrm{lk},\ell})\Big) = \emptyset$. We present in figure 5.7 an example of such construction.

This implies $\|q_{i_0} - q_{j_0}\|_2 > d_Q(j_0; [j_0^{\mathrm{lk},\ell}, j_1^{\mathrm{lk},\ell}]) \geq h_Q^{j_0}(j_0, j_0^{\mathrm{lk},\ell}, j_1^{\mathrm{lk},\ell})$. Using lemma 5.2.3, statement (a) we obtain $\|q_{i_0} - q_{j_0}\|_2 \geq \alpha_1 / f_1(Q; Q_{\mathrm{ref}})$. We denote as $h_Q^{j_0}(j_0, j_0^{\mathrm{lk},\ell}, j_1^{\mathrm{lk},\ell})$ the height of the 2-face uniquely defined by $\{j_0\}$ and $\mathrm{conv}_Q(j_0^{\mathrm{lk},\ell}, j_1^{\mathrm{lk},\ell})$, which passes through $\{q_{j_0}\}$. • Suppose now that $\{q_{i_0}\} \cap \overline{\mathrm{St}(\{q_{j_0}\})} \neq \emptyset$ holds. Since $\{q_{i_0}\}, \mathrm{conv}_Q(j_0, j_1) \in \Sigma_\Delta(Q)$, we conclude $\{q_{i_0}\} \in \mathrm{lk}(\{q_{j_0}\})$ and thus there exists an edge between $\{q_{i_0}\}$ and $\{q_{j_0}\}$. It thus follows from lemma 5.2.3, statement (c) that $\|q_{i_0} - q_{j_0}\|_2 \geq 2\alpha_1 / f_1(Q; Q_{\mathrm{ref}})$ holds. In both cases, we can conclude that $\|q_{i_0} - q_{j_0}\|_2 > \min\{\alpha_1, \alpha_2\} / \sqrt{2}\, f_1(Q; Q_{\mathrm{ref}})$.

This result, together with lemma 5.2.4, implies (5.10) as follows:

$$D_Q(i_0;[j_0,j_1]) \geq d_Q(i_0;[j_0,j_1]) > \frac{\min\{\alpha_1,\alpha_2\}}{\sqrt{2}f_1(Q;Q_{\text{ref}})} \min_{\vartheta\in\Theta}\left\{\sqrt{1-\cos^2(\vartheta)}\right\}.$$

Thus statement (c) for the case that $\{q_{j_0}\}$ is an interior vertex.

$\{q_{j_0}\}$ is a boundary 0-face: • We first consider the case $\{q_{i_0}\} \cap \overline{\text{St}(\{q_{j_0}\})} = \emptyset$. Thanks to the definition of boundary faces, we know that there exists a boundary 0-face, denoted as j_1^∂, such that $\text{conv}_Q(j_0, j_1^\partial)$ is a boundary edge and $\|q_{i_0} - q_{j_0}\|_2 \geq d_Q(i_0;[j_0,j_1^\partial])$. Using (4.14), and (5.8) we get $\|q_{i_0} - q_{j_0}\|_2 \geq D_Q(i_0;[j_0,j_1^\partial])/\sqrt{2} \geq \alpha_2/\sqrt{2}\, f_1(Q;Q_{\text{ref}})$. • If on the other hand, $\{q_{i_0}\} \cap \overline{\text{St}(\{q_{j_0}\})} \neq \emptyset$, then by definition of $\overline{\text{St}(\{q_{j_0}\})}$, we know that there exists an edge between $\{q_{i_0}\}$ and $\{q_{j_0}\}$. It thus follows from lemma 5.2.3, statement (c) that $\|q_{i_0} - q_{j_0}\|_2 \geq 2\alpha_1/f_1(Q;Q_{\text{ref}})$ holds. We can summarize both cases as $\|q_{i_0} - q_{j_0}\| \geq \min\{\alpha_1,\alpha_2\}/\sqrt{2}\, f_1(Q;Q_{\text{ref}})$. This result, together with lemma 5.2.4, implies (5.10) and thus statement (c) for the case that $\{q_{j_0}\}$ is an boundary vertex.

• We end this proof by considering the case when $\{q_{i_0}\} \cap \overline{\text{St}(\text{conv}_Q(j_0,j_1))} \neq \emptyset$. Since $\{q_{i_0}\}, \text{conv}_Q(j_0,j_1) \in \Sigma_\Delta(Q)$, the only possibility is that the vertex $\{q_{i_0}\}$ and the closed star of the edge $\text{conv}_Q(j_0,j_1)$ satisfy $\{q_{i_0}\} \cap \overline{\text{St}(\text{conv}_Q(j_0,j_1))} = \{q_{i_0}\}$. By considering the triangle uniquely defined by $\text{conv}_Q(j_0,j_1)$ and $\{q_{i_0}\}$, we have $D_Q(i_0;[j_0,j_1]) \geq d_Q(i_0;[j_0,j_1]) \geq h_Q^{i_0}(i_0,j_0,j_1)$, where $h_Q^{i_0}(i_0,j_0,j_1)$ is the height of the 2-face uniquely defined by $[i_0,j_0,j_1]$ which passes through $\{q_{i_0}\}$. Using lemma 5.2.3, statement (a) and the fact that $\sqrt{1-\cos^2(\vartheta_\ell)} < 1$ for all $\ell = 1,\dots,4$, we conclude

$$D_Q(i_0;[j_0,j_1]) \geq \frac{\alpha_1}{f_1(Q;Q_{\text{ref}})} > \frac{\min\{\alpha_1,\alpha_2\}}{\sqrt{2}f_1(Q;Q_{\text{ref}})} \min_{\vartheta\in\Theta}\{\sqrt{1-\cos^2(\vartheta)}\},$$

which proves (5.10) and thus statement (c) in this case. □

We are now in a position to present the main theorem of this chapter.

Theorem 5.2.6. *Suppose that $\alpha_1,\alpha_2,\alpha_3 > 0$ holds. Then the restriction of f_1 to the manifold of planar triangular meshes $\mathcal{M}_+(\Delta;Q_{\text{ref}})$ is proper.*

Proof. Suppose that $K \subset \mathbb{R}$ is an arbitrary, compact set. We can suppose that $f_1^{-1}(K)$ is nonempty. Consequently, we have $K \subset [a,b]$ for some $b > 0$. Since $\mathcal{M}_+(\Delta;Q_{\text{ref}})$ carries the metric subspace topology of $\mathbb{R}^{2\times N_V}$, the compactness of $f_1^{-1}(K)$ agrees with its sequential compactness. To verify the latter, suppose that $(Q_n) \subset f_1^{-1}(K) \subset \mathcal{M}_+(\Delta;Q_{\text{ref}})$ is an arbitrary sequence. We will show that it contains a subsequence which converges in $\mathcal{M}_+(\Delta;Q_{\text{ref}})$.

The definition of f_1 implies

$$\frac{\alpha_3}{2}\|Q_n - Q_{\text{ref}}\|_F^2 \leq f_1(Q_n;Q_{\text{ref}}) \leq b$$

for all $n \in \mathbb{N}$ and thus (Q_n) is bounded. Consequently, there exists a subsequence (which we do not relabel) converging to some Q^* in $\mathbb{R}^{2\times N_V}$. It remains to prove that Q^* belongs to $\mathcal{M}_+(\Delta;Q_{\text{ref}})$ and that $Q^* \in f_1^{-1}(K)$ holds. We proceed by proving the following results about the limit configuration Q^*:

(*i*) The signed area $A_{Q^*}(i_0^k, i_1^k, i_2^k)$ of each triangle $k = 1,\dots,N_T$ is strictly positive. In particular, the points $\{q^*_{i_0^k}, q^*_{i_1^k}, q^*_{i_2^k}\}$ are affine independent.

(*ii*) $\Sigma_\Delta(Q^*)$ is a simplicial 2-complex whose associated abstract simplicial complex is Δ.

Statements (*i*) and (*ii*) together prove that Q^* belongs to $\mathcal{M}_+(\Delta)$. From there we will proceed to show that

(*iii*) Q^* belongs to $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ and $Q^* \in f_1^{-1}(K)$ holds.

To show statement (*i*), fix an arbitrary oriented 2-face $[i_0, i_1, i_2]$ of Δ. Thanks to lemma 5.2.3, statement (d) we know

$$A_{Q_n}(i_0, i_1, i_2) \geq \frac{\pi\alpha_1^2}{f_1^2(Q_n; Q_{\text{ref}})} \geq \frac{\pi\alpha_1^2}{b^2}.$$

Since $A_Q(i_0, i_1, i_2)$ depends continuously on Q, we can pass to the limit and obtain statement (*i*).

The proof of statement (*ii*) is broken down into the following steps, according to the definition of $\Sigma_\Delta(Q^*)$, see (4.17), and the definition of simplicial complexes, see subsection 4.2.1.

(α) $\Sigma_\Delta(Q^*)$ is a nonempty, finite collection of simplices in $\mathbb{R}^2$.
(β) Every face of a simplex in $\Sigma_\Delta(Q^*)$ also belongs to $\Sigma_\Delta(Q^*)$.
(γ) The nonempty intersection of any two simplices σ_*, σ'_* in $\Sigma_\Delta(Q^*)$ is a face of both σ_* and σ'_*.
(δ) The abstract simplicial complex underlying $\Sigma_\Delta(Q^*)$ is Δ.

We proved in statement (*i*) that $A_{Q^*}(i_0, i_1, i_2) > 0$ holds for all 2-faces in Δ. Therefore, the node positions $\{q^*_{i_0}, q^*_{i_1}, q^*_{i_2}\}$ are affine independent so that their convex hulls are 2-simplices in $\mathbb{R}^2$. Since Δ is an abstract simplicial 2-complex, all other sets in $\Sigma_\Delta(Q^*)$ are the convex hulls of some subset of the vertices of a triangle, and these vertices are clearly affine independent as well. This shows statements (α) and (β).

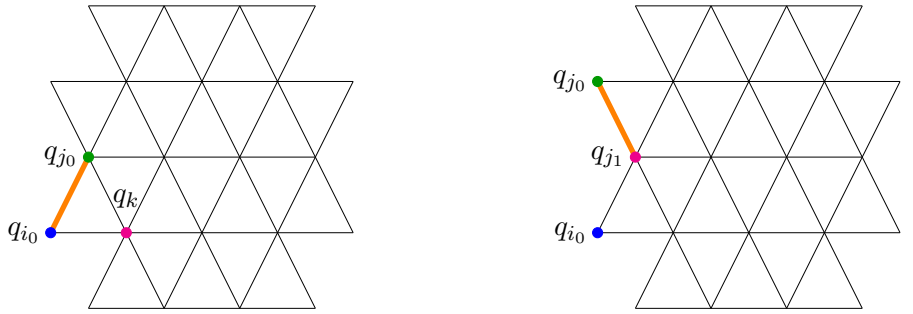
The proof of statements (γ) and (δ) is the most difficult part. In practical terms, we have to show that the structure of the simplicial complex describing the mesh does not change when passing to the limit $n \to \infty$ in the node positions.

The dimensions of the individual simplices in $\Sigma_\Delta(Q_n)$ are easily seen to be stable under this limit. Indeed, notice that 1-faces will not collapse at the limit thanks to $d_Q(i_0; i_1) = E^\ell_Q(i_0, i_1, i_2)$ and lemma 5.2.3, statement (c). In the same way, 2-faces do not collapse at the limit thanks to the bound on the heights given by lemma 5.2.3, statement (a). Therefore, the only concern is that unwanted intersections might appear. For instance, a vertex might converge to meet an edge which it is not supposed to intersect. We need to show that the properties of f_1 prevent this from happening.

In the following we are considering two arbitrary, *distinct* faces σ and σ' of Δ (the case of Δ consisting of a single 2-simplex is trivial). According to the dimension of σ and the vertices involved, we denote it by $[i_0]$ if it is a 0-face, by $[i_0, i_1]$ if it is a 1-face, and by $[i_0, i_1, i_2]$ in case of a 2-face. The corresponding vertex indices for σ' are j_0, j_1 and j_2. We denote the corresponding faces in $\Sigma_\Delta(Q_n)$ by σ_n and σ'_n, and those in $\Sigma_\Delta(Q^*)$ by σ_* and σ'_*, respectively. For instance, when $\sigma = [i_0, i_1]$, then $\sigma_n = \text{conv}_{Q_n}(i_0, i_1)$ and $\sigma_* = \text{conv}_{Q^*}(i_0, i_1)$.

We proceed by distinguishing cases, according to the dimensions of σ and σ'. Notice that $\dim\sigma = \dim\sigma_n = \dim\sigma_*$ and $\dim\sigma' = \dim\sigma'_n = \dim\sigma'_*$, which follows from the proof of statements (α) and (β) above. In each case, we need to verify that the intersection $\sigma_* \cap \sigma'_*$ is of the same type (empty set, a vertex, etc.) as $\sigma \cap \sigma'$ and $\sigma_n \cap \sigma'_n$. To this end, we argue that nonintersecting faces maintain a positive distance d_{Q^*} or D_{Q^*} also in the limit.

In addition to the dimensions of σ and σ', we need to distinguish whether they are interior or boundary faces. We will start assuming both faces σ and σ' are boundary, then



(a) $[i_0, j_0]$ belongs to Δ. (b) $[i_0, j_0]$ does not belong to Δ.

Figure 5.8. Illustration of the case when $[i_0]$ (green) and $[j_0]$ (blue) are boundary 0-faces, from proof of theorem 5.2.6.

we assume that both faces σ and σ' are interior faces. This part of the proof is complete once we have considered without loss of generality that σ is an interior and σ' is a boundary face. Obviously, the case with the reversed roles of σ and σ' will be identical.

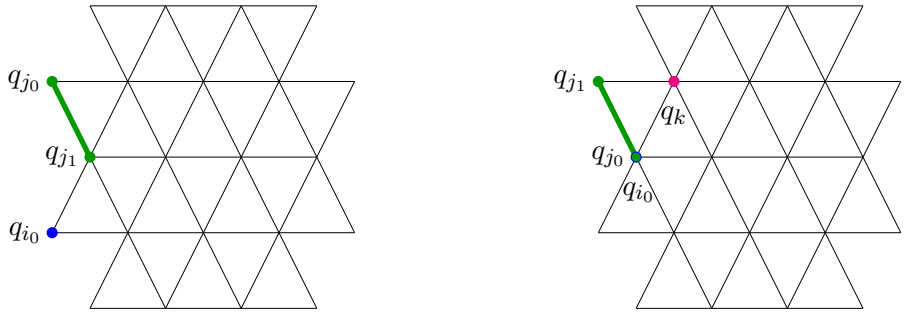
σ, σ' are boundary 0-faces: Since $\sigma = [i_0]$ and $\sigma' = [j_0]$ are distinct vertices, $\sigma \cap \sigma' = \emptyset$ holds. We have two cases to consider. • First, if $[i_0, j_0]$ is a 1-face in Δ, then there exists another 0-face $[k]$ such that $[i_0, j_0, k]$ is a 2-face of Δ. This is since Δ is pure (see figure 5.8a for an illustration). Therefore, $\|q_{i_0}^n - q_{j_0}^n\|$ agrees with the length $E_{Q_n}^\ell(i_0, j_0, k)$ of an edge for some $\ell \in \{0, 1, 2\}$. From lemma 5.2.3, statement (c), we know $E_{Q_n}^\ell(i_0, j_0, k) \geq (2\alpha_1)/f_1(Q_n; Q_{\text{ref}}) \geq (2\alpha_1)/b$ and thus $\|q_{i_0}^* - q_{j_0}^*\| \geq (2\alpha_1)/b$, i. e., $\sigma_* = \{q_{i_0}^*\}$ and $\sigma'_* = \{q_{j_0}^*\}$ do not intersect.
• Second, if $[i_0, j_0]$ is not a face in Δ, then there exists $[j_1]$ such that $[j_0, j_1]$ is a boundary 1-face, we refer to figure 5.8b for an illustration. Using proposition 5.2.5, statement (a) we know $D_{Q_n}(i_0; [j_0, j_1]) \geq \alpha_2/b$ and thus $D_{Q^*}(i_0; [j_0, j_1]) \geq \alpha_2/b$. In particular, $\|q_{i_0}^* - q_{j_0}^*\| \geq \alpha_2/b$ holds. Again, $\sigma_* \cap \sigma'_* = \emptyset$.

σ is a boundary 0-face and σ' is a boundary 1-face: • First assume that $\sigma = [i_0]$ and $\sigma' = [j_0, j_1]$ do not intersect. Using proposition 5.2.5, statement (a) we obtain $D_{Q_n}(i_0; [j_0, j_1]) \geq \alpha_2/b$ and thus $D_{Q^*}(i_0; [j_0, j_1]) \geq \alpha_2/b$, i. e., $\sigma_* \cap \sigma'_* = \emptyset$, as illustrated in figure 5.9a. • Second, if σ and σ' intersect, then necessarily $\sigma \cap \sigma' = \sigma$. Without loss of generality, suppose $[i_0] = [j_0] \neq [j_1]$. Moreover, since Δ is pure, there exists a vertex $[k]$ such that $[j_0, j_1, k]$ is a 2-face of Δ. (See figure 5.9b for an illustration). Therefore, $\|q_{i_0}^n - q_{j_1}^n\| = \|q_{j_0}^n - q_{j_1}^n\|$ agrees with the length $E_{Q_n}^\ell(j_0, j_1, k)$ of an edge for some $\ell \in \{0, 1, 2\}$. Thanks to lemma 5.2.3, statement (c), we have $E_{Q_n}^\ell(j_0, j_1, k) \geq (2\alpha_1)/f_1(Q_n; Q_{\text{ref}}) \geq (2\alpha_1)/b$ and thus $\|q_{i_0}^* - q_{j_1}^*\| \geq (2\alpha_1)/b$. Therefore, $\sigma_* \cap \sigma'_* = \sigma_*$ as desired.

σ is a boundary 0-face and σ' is a boundary 2-face: Thanks to proposition 4.2.10 we know that for all $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ it holds:

$$\begin{aligned} d_Q(i_0; [j_0, j_1, j_2]) & \qquad (5.11) \\ &= \min\{d_Q(i_0; [j_0, j_1]), d_Q(i_0; [j_1, j_2]), d_Q(i_0; [j_2, j_0])\} \\ &\geq \frac{1}{\sqrt{2}} \min\{D_Q(i_0; [j_0, j_1]), D_Q(i_0; [j_1, j_2]), D_Q(i_0; [j_0, j_2])\}\,. & (5.12) \end{aligned}$$

• Now, we suppose that $\sigma = [i_0]$ and $\sigma' = [j_0, j_1, j_2]$ do not intersect. Since σ' is a boundary face, at least one of its 1-faces is a boundary face. Without loss of



(a) $\{q_{i_0}\}$, $\operatorname{conv}_Q(j_0, j_1)$ do not intersect. (b) $\{q_{i_0}\}$, $\operatorname{conv}_Q(j_0, j_1)$ intersect in $\{q_{j_0}\}$.

Figure 5.9. Illustration of the case when $[i_0]$ (blue) and $[j_0, j_1]$ (green) are boundary 0-face and 1-face, respectively. See proof of theorem 5.2.6.

generality, let $[j_0, j_1]$ be a boundary 1-face; an illustration of this case can be found in figure 5.10a. As in the previous case, proposition 5.2.5, statement (a) gives a bound on the first term of (5.12). Moreover, w.l.o.g. we assume $[j_1, j_2]$ is and interior edge. Then, thanks to proposition 5.2.5, statement (c) the second term of (5.12) is bounded. In the same way, if $[j_0, j_2]$ is either interior or boundary can be bounded using one of the previous arguments. In summary, we get

$$d_{Q_n}(i_0; [j_0, j_1, j_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b} \min_{\vartheta \in \Theta} \left\{ \sqrt{1 - \cos^2(\vartheta)} \right\}.$$

Notice, that thanks to lemma 5.2.3, statement (f), $|\cos(\vartheta)| \leq \Psi(b) < 1$ holds for all $\vartheta \in \Theta$, which implies

$$d_{Q_n}(i_0; [j_0, j_1, j_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b} \left(\sqrt{1 - \Psi^2(b)} \right),$$

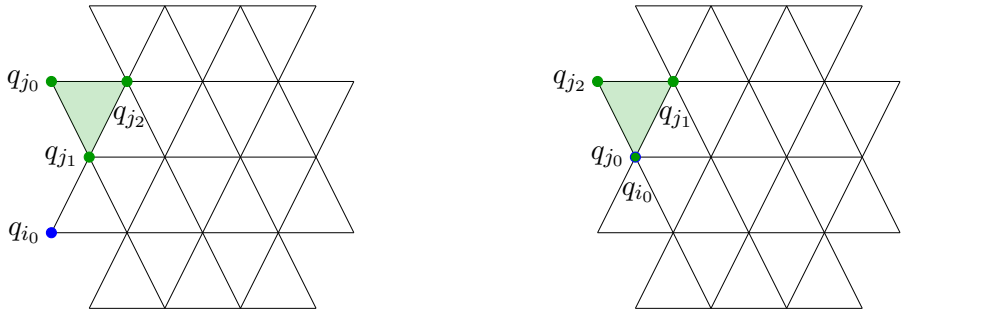
and thus the same bound is valid for $d_{Q^*}(i_0; [j_0, j_1, j_2])$. This shows that $\sigma_* \cap \sigma'_* = \emptyset$ holds.

• Second, if σ and σ' intersect, then necessarily $\sigma \cap \sigma' = \sigma$. Without loss of generality, suppose $[i_0] = [j_0]$. Therefore, $\|q^n_{i_0} - q^n_{j_1}\| = \|q^n_{j_0} - q^n_{j_1}\|$ agrees with the length $E^\ell_{Q_n}(j_0, j_1, j_2)$ of an edge for some $\ell \in \{0, 1, 2\}$. In the same way, the distance $d_{Q_n}(i_0; [j_1, j_2])$ agrees with $h^\ell_{Q_n}(j_0, j_1, j_2)$ for some $\ell = 0, 1, 2$. Thus, lemma 5.2.3, statements (a) and (c) show $\sigma_* \cap \sigma'_* = \sigma_*$ as desired. This case is depicted in figure 5.10b.

σ and σ' are boundary 1-faces: As in the previous case, proposition 4.2.10 guarantees that for any $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$, the distance between two edges $d_Q([i_0, i_1]; [j_0, j_1])$ satisfies:

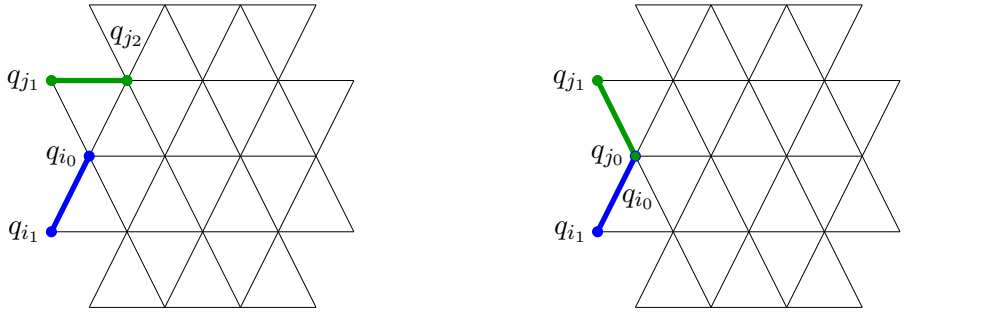
$$\begin{aligned} &d_Q([i_0, i_1]; [j_0, j_1]) \\ &\quad = \min\{d_Q(i_0; [j_0, j_1]), d_Q(i_1; [j_0, j_1]), d_Q(j_0; [i_0, i_1]), d_Q(j_1; [i_0, i_1])\} \\ &\quad \geq \frac{1}{\sqrt{2}} \min\{D_Q(i_0; [j_0, j_1]), D_Q(i_1; [j_0, j_1]), D_Q(j_0; [i_0, i_1]), D_Q(j_1; [i_0, i_1])\}. \end{aligned} \tag{5.13}$$

• First suppose that $\sigma = [i_0, i_1]$ and $\sigma' = [j_0, j_1]$ do not intersect. Since both σ and σ' are boundary 1-faces, their vertices are boundary vertices; see figure 5.11a for an illustration. Therefore, by proposition 5.2.5, statement (a), all four terms on the right-hand side of (5.13) with Q replaced by Q_n are bounded below by α_2/b. Consequently, we obtain $d_{Q^*}([i_0, i_1]; [j_0, j_1]) \geq \alpha_2/(\sqrt{2}b)$, i. e., $\sigma_* \cap \sigma'_* = \emptyset$. • Second,



(a) $\{q_{i_0}\}$, $\mathrm{conv}_Q(j_0, j_1, j_2)$ do not intersect. (b) $\{q_{i_0}\}$, $\mathrm{conv}_Q(j_0, j_1, j_2)$ intersect in $\{q_{j_0}\}$.

Figure 5.10. Illustration of the case when $[i_0]$ (blue) and $[j_0, j_1, j_2]$ (green) are boundary 0-face and 2-face, respectively. See proof of theorem 5.2.6.



(a) $\mathrm{conv}_Q(i_0, i_1)$ does not intersect $\mathrm{conv}_Q(j_0, j_1)$. (b) $\mathrm{conv}_Q(i_0, i_1)$ intersects $\mathrm{conv}_Q(j_0, j_1)$ at $\{q_{i_0}\} = \{q_{j_0}\}$.

Figure 5.11. Illustration of the case when $[i_0, i_1]$ (blue) and $[j_0, j_1]$ (green) are boundary 1-faces, from proof of theorem 5.2.6.

when $\sigma = [i_0, i_1]$ and $\sigma' = [j_0, j_1]$ intersect, they intersect in a vertex. Without loss of generality, $[i_0] = [j_0]$ holds. Thanks to proposition 5.2.5, statement (a), we find that $D_{Q_n}(i_1; [j_0, j_1]) \geq \alpha_2/b$ holds and thus $D_{Q^*}(i_1; [j_0, j_1]) \geq \alpha_2/b$ as well. The same argument also shows $D_{Q^*}(j_1; [i_0, i_1]) \geq \alpha_2/b$. As mentioned previously, the edge lengths $\|q^*_{i_0} - q^*_{i_1}\|$ and $\|q^*_{j_0} - q^*_{j_1}\|$ remain positive by lemma 5.2.3, statement (c). This implies $\sigma_* \cap \sigma'_* = \{q^*_{i_0}\} = \{q^*_{j_0}\}$; we refer to figure 5.11b for an illustration.

σ is a boundary 1-face and σ' is a boundary 2-face: In virtue of proposition 4.2.10 we know that for all $Q \in \mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$, the distance of an edge to a triangle

$d_Q([i_0,i_1];[j_0,j_1,j_2])$ satisfies

$$\begin{aligned} &d_Q([i_0,i_1];[j_0,j_1,j_2]) \\ &\quad = \min\{d_Q([i_0,i_1];[j_0,j_1]), d_Q([i_0,i_1];[j_1,j_2]), d_Q([i_0,i_1];[j_2,j_0])\} \\ &\quad \geq \frac{1}{\sqrt{2}} \min\big\{D_Q(i_0;[j_0,j_1]), D_Q(i_0;[j_1,j_2]), D_Q(i_0;[j_2,j_0]), \\ &\qquad\qquad D_Q(i_1;[j_0,j_1]), D_Q(i_1;[j_1,j_2]), D_Q(i_1;[j_2,j_0]), \\ &\qquad\qquad D_Q(j_0;[i_0,i_1]), D_Q(j_1;[i_0,i_1]), D_Q(j_2;[i_0,i_1])\big\}. \end{aligned} \tag{5.14}$$

• First suppose that $\sigma = [i_0,i_1]$ and $\sigma' = [j_0,j_1,j_2]$ do not intersect. Some of the terms on the right-hand side of (5.14) with Q replaced by Q_n are distances between boundary vertices and boundary edges, for which proposition 5.2.5, statement (a) provides the lower bound α_2/b. The remaining terms are distances between interior vertices and interior edges, or boundary vertices and interior edges, which can be bounded below by proposition 5.2.5, statement (b) or statement (c), respectively (see figure 5.12a for an illustration). Altogether, we obtain

$$d_{Q_n}([i_0,i_1];[j_0,j_1,j_2]) \geq \frac{\min\{\alpha_1,\alpha_2\}}{2b}\sqrt{1-\Psi^2(b)} > 0.$$

• Second, assume that σ and σ' intersect in a vertex. Without loss of generality, $[i_0] = [j_0]$ holds, as depicted in figure 5.12c. We recall (5.12):

$$\begin{aligned} &d_Q(i_1;[j_0,j_1,j_2]) \\ &\quad \geq \frac{1}{\sqrt{2}} \min\{D_Q(i_1;[j_0,j_1]), D_Q(i_1;[j_1,j_2]), D_Q(i_1;[j_2,j_0])\}. \end{aligned}$$

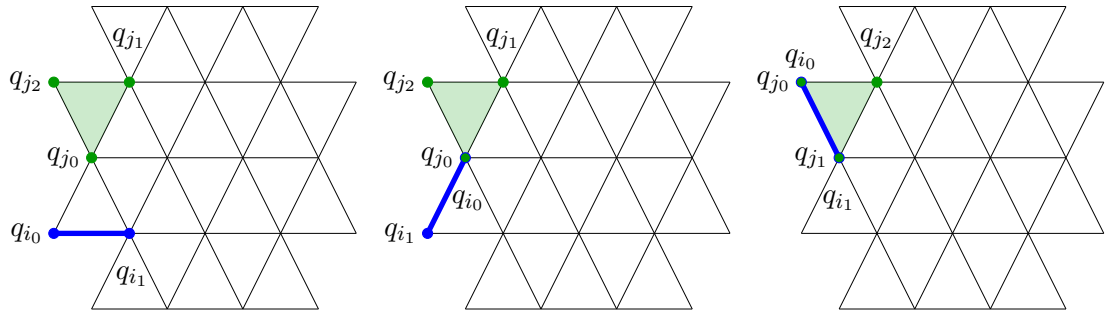
As above, some of the terms on the right-hand side of (5.12) with Q replaced by Q_n are distances between the boundary vertex $[i_1]$ and boundary edges, and the remaining terms are distances between $[i_1]$ and interior edges. An application of proposition 5.2.5, statement (a) and statement (c) yields a uniformly positive lower bound for $d_{Q_n}(i_1;[j_0,j_1,j_2])$ and thus for $d_{Q^*}(i_1;[j_0,j_1,j_2])$. This implies $\sigma_* \cap \sigma'_* = \{q^*_{i_0}\} = \{q^*_{j_0}\}$. • Third, assume that $\sigma \cap \sigma' = \sigma$. Without loss of generality, suppose $[i_0] = [j_0]$ and $[i_1] = [j_1]$, as illustrated in figure 5.12c. Then we can use $d_{Q_n}(j_2;[i_0,i_1]) \geq h^\ell_{Q_n}(j_0,j_1,j_2)$ for $\ell = 2$, and thus lemma 5.2.3, statement (a) yields a lower bound of α_1/b. As mentioned previously, the edge length $\|q_{i_0} - q_{i_1}\| = \|q_{j_0} - q_{j_1}\|$ remains positive by lemma 5.2.3, statement (c). From here we conclude $\sigma_* \cap \sigma'_* = \sigma_*$.

σ and σ' are boundary 2-faces: Thanks to proposition 4.2.10, it holds:

$$\begin{aligned} &d_Q([i_0,i_1,i_2];[j_0,j_1,j_2]) \\ &\quad \geq \frac{1}{\sqrt{2}} \min_{\substack{\ell=0,1,2 \\ \hat{\ell}=0,1,2}} \Big\{D_Q(i_\ell;[j_{\hat{\ell}},j_{\hat{\ell}\oplus 1}]), D_Q(j_{\hat{\ell}};[i_\ell,i_{\ell\oplus 1}])\Big\}. \end{aligned} \tag{5.15}$$

• First suppose that $\sigma = [i_0,i_1,i_2]$ and $\sigma' = [j_0,j_1,j_2]$ do not intersect, as depicted in figure 5.13a. Each term on the right-hand side of (5.15) with Q replaced by Q_n can be estimated below by proposition 5.2.5, statements (a) to (c). Thanks to lemma 5.2.3, statement (f), we obtain the uniform lower bound

$$d_{Q_n}([i_0,i_1,i_2];[j_0,j_1,j_2]) \geq \frac{\min\{\alpha_1,\alpha_2\}}{2b}\sqrt{1-\Psi^2(b)} > 0,$$



(a) $\mathrm{conv}_Q(i_0, i_1)$ does not intersect $\mathrm{conv}_Q(j_0, j_1, j_2)$.

(b) $\mathrm{conv}_Q(i_0, i_1)$ intersects $\mathrm{conv}_Q(j_0, j_1, j_2)$ at $\{q_{i_0}\} = \{q_{j_0}\}$.

(c) $\mathrm{conv}_Q(i_0, i_1)$ intersects $\mathrm{conv}_Q(j_0, j_1, j_2)$ at $\mathrm{conv}_Q(i_0, i_1) = \mathrm{conv}_Q(j_0, j_1)$.

Figure 5.12. Illustration of the case when $[i_0, i_1]$ (blue) and $[j_0, j_1, j_2]$ (green) are boundary 1-face and 2-face, respectively. See proof of theorem 5.2.6

and thus $\sigma_* \cap \sigma'_* = \emptyset$ holds.
• Second, suppose that σ and σ' intersect in a vertex. Without loss of generality, suppose $[i_0] = [j_0]$; an illustration of this case can be found in figure 5.13b. We need to consider

$$d_Q(i_1; [j_0, j_1, j_2]) \geq \frac{1}{\sqrt{2}} \min\{D_Q(i_1; [j_0, j_1]), D_Q(i_1; [j_1, j_2]), D_Q(i_1; [j_2, j_0])\}$$

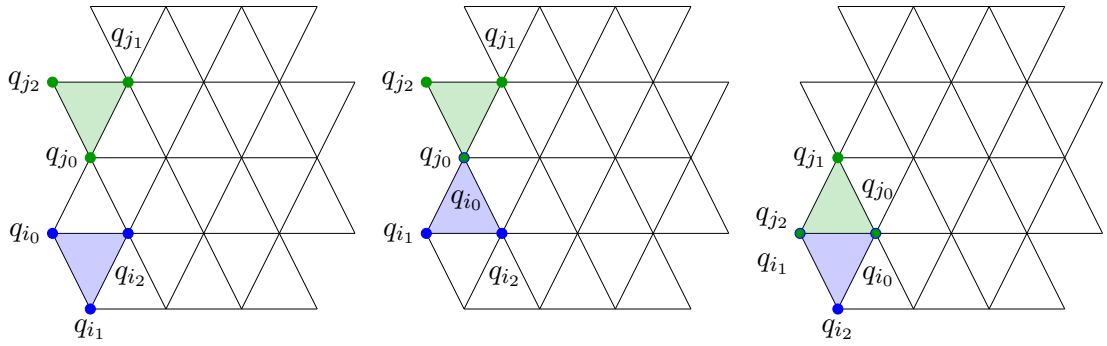
and the same with i_1 replaced by i_2 and show that these expressions are bounded away from zero for $Q = Q_n$. Regardless of whether $[i_1]$ and $[i_2]$, and $[j_0, j_1]$, $[j_1, j_2]$, $[j_2, j_0]$ are interior or boundary faces, in each case, one of proposition 5.2.5, statements (a) to (c) applies and provides this lower bound. The same argument applies with the roles of σ and σ' reversed. We can conclude that $\sigma_* \cap \sigma'_* = \{q^*_{i_0}\} = \{q^*_{j_0}\}$ as desired. • Third, suppose that σ and σ' intersect in a common edge, which is necessarily an interior 1-face. Without loss of generality, suppose that $[i_0] = [j_0]$ and $[i_1] = [j_2]$. We need to estimate only the distances from $[i_2]$ to $[j_0, j_1, j_2]$ and from $[j_1]$ to $[i_0, i_1, i_2]$, cf., figure 5.13c. To this end, we use:

$$d_{Q_n}(i_2; [j_0, j_1, j_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b} \sqrt{1 - \Psi^2(b)} > 0, \quad \text{and}$$
$$d_{Q_n}(j_1; [i_0, i_1, i_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b} \sqrt{1 - \Psi^2(b)} > 0,$$

which follow proposition 5.2.5, statements (a) to (c). Therefore, the intersection $\sigma_* \cap \sigma'_*$ equals to $\mathrm{conv}_{Q^*}(i_0, i_1)$ and to $\mathrm{conv}_{Q^*}(j_2, j_0)$ as desired. Moreover, $q^*_{i_0}$ and $q^*_{i_1}$ are going to remain distinct points since $\|q^*_{i_0} - q^*_{i_1}\| = E^\ell_{Q^*}(i_0, i_1, i_2)$, with $\ell = 2$ which remains bounded away from zero by lemma 5.2.3, statement (c).

Now, we focus in the cases when σ and σ' are interior faces.

σ, σ' are interior 0-faces: Since $\sigma = [i_0]$ and $\sigma' = [j_0]$ are distinct 0-faces, $\sigma \cap \sigma' = \emptyset$. We have to consider two cases. • First, if $[i_0, j_0]$ is a 1-face of Δ, then, there exists $[k]$ such that $[i_0, j_0, k] \in \Delta$ (since Δ is pure). Therefore, $\|q^n_{i_0} - q^n_{j_0}\|$ coincides with the edge length $E^\ell_{Q^n}(i_0, j_0, k)$ for some $\ell = 0, 1, 2$. From lemma 5.2.3, statement (c), we know $E^\ell_{Q^n}(i_0, j_0, k) \geq (2\alpha_1)/f_1(Q_n, Q_{\mathrm{ref}}) \geq (2\alpha_1)/b$ and thus

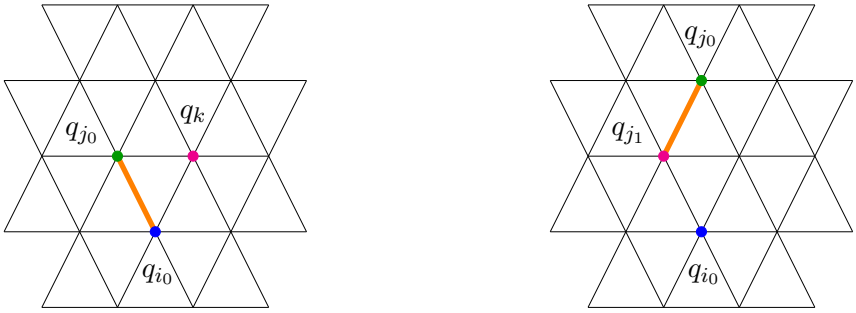


(a) $\mathrm{conv}_Q(i_0, i_1, i_2)$ does not intersect $\mathrm{conv}_Q(j_0, j_1, j_2)$.

(b) $\mathrm{conv}_Q(i_0, i_1, i_2)$ intersects $\mathrm{conv}_Q(j_0, j_1, j_2)$ at $\{q_{i_0}\} = \{q_{j_0}\}$.

(c) $\mathrm{conv}_Q(i_0, i_1, i_2)$ intersects $\mathrm{conv}_Q(j_0, j_1, j_2)$ at $\mathrm{conv}_Q(i_0, i_1) = \mathrm{conv}_Q(j_0, j_2)$.

Figure 5.13. Illustration of the case when $[i_0, i_1, i_2]$ (blue) and $[j_0, j_1, j_2]$ (green) are boundary 2-faces, from proof of theorem 5.2.6.



(a) $[i_0, j_0]$ belongs to Δ.

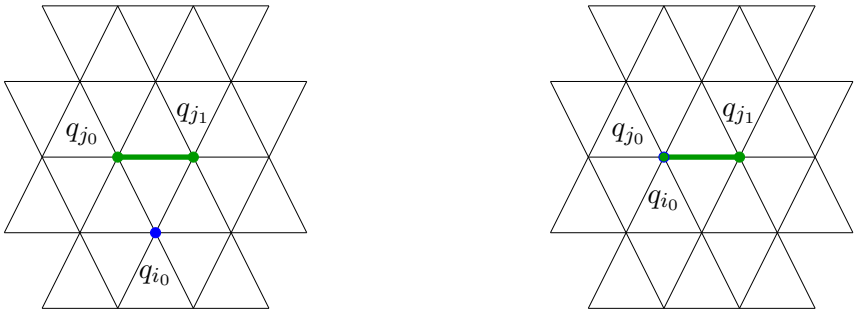
(b) $[i_0, j_0]$ does not belong to Δ.

Figure 5.14. Illustration of the case when $[i_0]$ (blue) and $[j_0]$ (green) are interior 0-faces, from proof of theorem 5.2.6.

$\|q^*_{i_0} - q^*_{j_0}\| \geq (2\alpha_1)/b > 0$, i. e., $\{q^*_{i_0}\}$ and $\{q^*_{j_0}\}$ do not intersect. See figure 5.14a for an illustration. • Second, if $[i_0, j_0]$ is not a 1-face of Δ, then there exists j_1 such that $[j_0, j_1]$ is an interior 1-face, as depicted for example in figure 5.14b. Using proposition 5.2.5, statement (b) we know $D_{Q_n}(i_0; [j_0, j_1]) \geq \alpha_1/b$ and thus $D_{Q^*}(i_0; [j_0, j_1]) \geq \alpha_1/b$. In particular, $\|q^*_{i_0} - q^*_{j_0}\| \geq \alpha_1/b$. Again, $\sigma_* \cap \sigma'_* = \emptyset$.

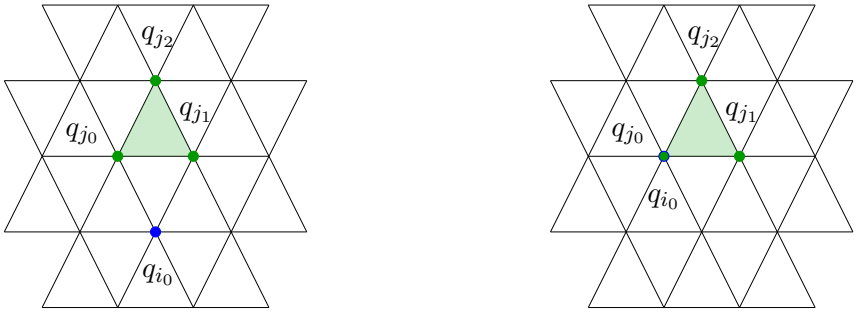
σ is an interior 0-face and σ' is an interior 1-face: • First assume that $\sigma = [i_0]$ and $\sigma' = [j_0, j_1]$ do not intersect. Using proposition 5.2.5, statement (b) we obtain $D_{Q_n}(i_0; [j_0, j_1]) \geq \alpha_1/b$ and thus $D_{Q^*}(i_0; [j_0, j_1]) \geq \alpha_1/b$, i. e., $\sigma_* \cap \sigma'_* = \emptyset$. See figure 5.15a for an illustration. • Second, if σ and σ' intersect, then necessarily $\sigma \cap \sigma' = \sigma$, as depicted in figure 5.15b. Now, we use lemma 5.2.3, statement (c) to prove that $\sigma_* \cap \sigma'_* = \sigma_*$.

σ is an interior 0-face and σ' is an interior 2-face: Recall the expression of the distance between a 0-face and a 2-face given in (5.12). • First, we suppose that $\sigma = [i_0]$ and $\sigma' = [j_0, j_1, j_2]$ do not intersect. Since σ' is an interior face, all of its 1-faces are interior faces (see figure 5.16a for an illustration), and by virtue of proposition 5.2.5, statement (b) we obtain that all the distances $D_{Q_n}(i_0; [j_0, j_1])$,$D_{Q_n}(i_0; [j_1, j_2])$, $D_{Q_n}(i_0; [j_0, j_2])$ are bounded away from zero by α_1/b. Thus, $d_{Q^*}(i_0; [j_0, j_1, j_2]) >$



(a) $\{q_{i_0}\}$ does not intersect $\mathrm{conv}_Q(j_0, j_1)$. (b) $\{q_{i_0}\}$ intersects $\mathrm{conv}_Q(j_0, j_1)$ at $\{q_{i_0}\} = \{q_{j_0}\}$.

Figure 5.15. Illustration of the case when $[i_0]$ (blue) and $[j_0, j_1]$ (green) are interior 0-face and 1-face, respectively. See proof of theorem 5.2.6.



(a) $\{q_{i_0}\}$ does not intersect $\mathrm{conv}_Q(j_0, j_1, j_2)$. (b) $\{q_{i_0}\}$ intersects $\mathrm{conv}_Q(j_0, j_1, j_2)$ at $\{q_{i_0}\} = \{q_{j_0}\}$.

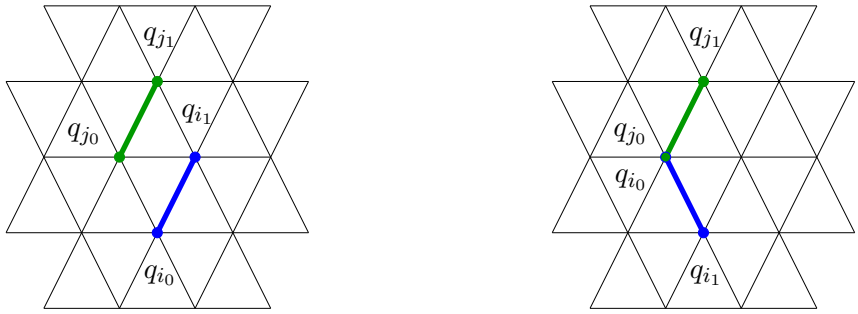
Figure 5.16. Illustration of the case when $[i_0]$ (blue) and $[j_0, j_1, j_2]$ (green) are interior 0-face and 2-face, respectively. See proof of theorem 5.2.6.

$\alpha_1/(\sqrt{2}b)$. Therefore, $\sigma_* \cap \sigma'_* = \emptyset$. • Second, if σ and σ' intersect, then necessarily $\sigma \cap \sigma' = \sigma$, as depicted in figure 5.16b. Using lemma 5.2.3, statements (a) and (c) one can prove that $\sigma_* \cap \sigma'_* = \sigma_*$.

σ and σ' are interior 1-faces: We recall distance between two edges $d_Q([i_0, i_1]; [j_0, j_1])$ given by (5.13) (cf., proposition 4.2.10). • First suppose that $\sigma = [i_0, i_1]$ and $\sigma' = [j_0, j_1]$ do not intersect. Since both σ and σ' are interior 1-faces, at least one of their 0-faces is interior, w.l.o.g. we assume $[i_0]$ and $[j_0]$ are the interior vertices. See figure 5.17a for an illustration. Therefore, by proposition 5.2.5, statement (b), the terms $D_{Q_n}(i_0; [j_0, j_1])$ and $D_{Q_n}(j_0; [i_0, i_1])$ appearing on the right-hand side of (5.13) are bounded below by α_1/b. If $[i_1]$ and/or $[j_1]$ are interior 0-faces we proceed in the same way. On the other hand, if $[i_1]$ and/or $[j_1]$ are boundary 0-faces, then we use proposition 5.2.5, statement (c), to bound the distances $D_{Q_n}(i_1; [j_0, j_1])$ and $D_{Q_n}(j_1; [i_0, i_1])$ by:

$$\frac{\min\{\alpha_1, \alpha_2\}}{\sqrt{2} f_1(Q_n; Q_{\mathrm{ref}})} \min_{\vartheta \in \Theta} \{\sqrt{1 - \cos^2(\vartheta)}\},$$

as given in (5.10). This implies $\sigma_* \cap \sigma'_* = \emptyset$, since the distances $D_{Q^*}(i_0; [j_0, j_1])$, $D_{Q^*}(i_1; [j_0, j_1])$, $D_{Q^*}(j_0; [i_0, i_1])$, $D_{Q^*}(j_1; [i_0, i_1])$ are strictly positive. • Second,



(a) $\mathrm{conv}_Q(i_0, i_1)$ does not intersect $\mathrm{conv}_Q(j_0, j_1)$.

(b) $\mathrm{conv}_Q(i_0, i_1)$ intersects $\mathrm{conv}_Q(j_0, j_1)$ at $\{q_{i_0}\} = \{q_{j_0}\}$.

Figure 5.17. Illustration of the case when $[i_0, i_1]$ (blue) and $[j_0, j_1]$ (green) are interior 1-faces. See proof of theorem 5.2.6.

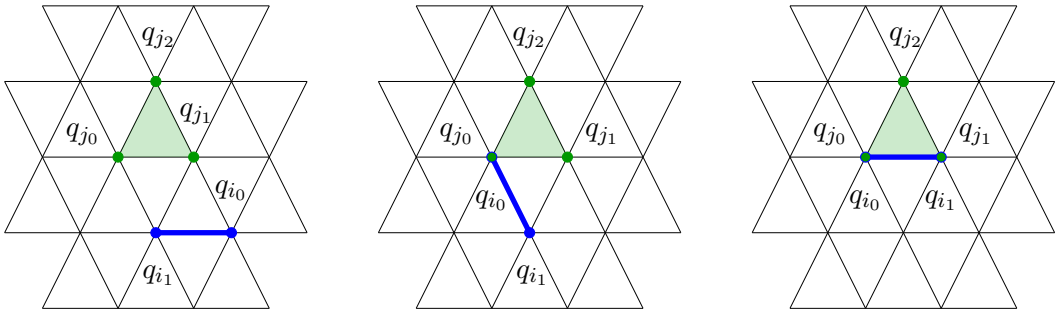
when $\sigma = [i_0, i_1]$ and $\sigma' = [j_0, j_1]$ intersect, they intersect in a vertex, as depicted in figure 5.17b. Without loss of generality, $[i_0] = [j_0]$ holds. If $[i_0]$ is boundary, then $[i_1]$ and $[j_1]$ are necessarily interior and therefore, one can use proposition 5.2.5, statement (b) to bound $D_{Q^*}(i_1; [j_0, j_1]) \geq \alpha_2/b$ and $D_{Q^*}(j_1; [i_0, i_1]) \geq \alpha_2/b$. If $[i_0]$ is interior, then $[i_1]$ and/or $[j_1]$ are boundary. In both cases, one can use proposition 5.2.5, statement (b), or statement (c) and find strictly positive bounds for the distances $D_{Q_n}(i_1; [j_0, j_1])$ and $D_{Q_n}(j_1; [i_0, i_1])$ and therefore for the distance when Q_n is replaced by Q^*. As previously stated, the edge lengths $\|q^*_{i_0} - q^*_{i_1}\|$ and $\|q^*_{j_0} - q^*_{j_1}\|$ remain positive by lemma 5.2.3, statement (c). This implies $\sigma_* \cap \sigma'_* = \{q^*_{i_0}\} = \{q^*_{j_0}\}$.

σ is a interior 1-face and σ' is a interior 2-face: We use again the distance of an edge to triangle $d_Q([i_0, i_1]; [j_0, j_1, j_2])$ given in (5.14). • First suppose that $\sigma = [i_0, i_1]$ and $\sigma' = [j_0, j_1, j_2]$ do not intersect, see figure 5.18a for an illustration. Some of the terms on the right-hand side of (5.14) with Q replaced by Q_n are distances between interior vertices and interior edges, for which proposition 5.2.5, statement (b) provides a lower bound α_1/b. The remaining terms are distances between boundary vertices and interior edges, which can be bounded below by proposition 5.2.5, statement (c). Altogether, we obtain again

$$d_{Q_n}([i_0, i_1]; [j_0, j_1, j_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b} \sqrt{1 - \Psi^2(b)} > 0.$$

• Second, assume that σ and σ' intersect in a vertex. Without loss of generality, $[i_0] = [j_0]$ holds, as illustrated in figure 5.18b. We recall the distance of a vertex to a triangle given in (5.12). As above, some of the terms on the right-hand side of (5.12) with Q replaced by Q_n are distances between the vertex $[i_1]$ which can be either boundary or interior and interior edges. An application of proposition 5.2.5, statement (b) or statement (c) yields a uniformly positive lower bound for $d_{Q_n}(i_1; [j_0, j_1, j_2])$ and thus for $d_{Q^*}(i_1; [j_0, j_1, j_2])$. This implies $\sigma_* \cap \sigma'_* = \{q^*_{i_0}\} = \{q^*_{j_0}\}$. • Third, assume that $\sigma \cap \sigma' = \sigma$. Without loss of generality, suppose $[i_0] = [j_0]$ and $[i_1] = [j_1]$, for example as shown in figure 5.18c. In this case it holds $d_{Q_n}(j_2; [i_0, i_1]) = h^{\ell}_{Q_n}(j_0, j_1, j_2)$, with $\ell = 2$, which we know is uniformly bounded away from zero in virtue of lemma 5.2.3, statement (a).

σ and σ' are interior 2-faces: We recall the definition of the distance between two triangles given in (5.15). • First suppose that $\sigma = [i_0, i_1, i_2]$ and $\sigma' = [j_0, j_1, j_2]$ do



(a) $\mathrm{conv}_Q(i_0, i_1)$ does not intersect $\mathrm{conv}_Q(j_0, j_1, j_2)$.

(b) $\mathrm{conv}_Q(i_0, i_1)$ intersects $\mathrm{conv}_Q(j_0, j_1, j_2)$ at $\{q_{i_0}\} = \{q_{j_0}\}$.

(c) $\mathrm{conv}_Q(i_0, i_1)$ intersects $\mathrm{conv}_Q(j_0, j_1, j_2)$ at $\mathrm{conv}_Q(i_0, i_1) = \mathrm{conv}_Q(j_0, j_1)$.

Figure 5.18. Illustration of the case when $[i_0, i_1]$ (blue) and $[j_0, j_1, j_2]$ (green) are interior 1-face and 2-face, respectively. See proof of theorem 5.2.6.

not intersect, as depicted in figure 5.19a. Each term on the right-hand side of (5.15) with Q replaced by Q_n can be estimated below by proposition 5.2.5, statements (b) and (c). These bounds, will involve the values of the cosines of interior angles, which thanks to lemma 5.2.3, statement (f) can be bounded away from zero, also for Q^*. • Second, if σ and σ' intersect in a vertex, which w.l.o.g. can be assumed to be $[i_0] = [j_0]$, as shown in figure 5.19b. Then, we need to consider the distances:

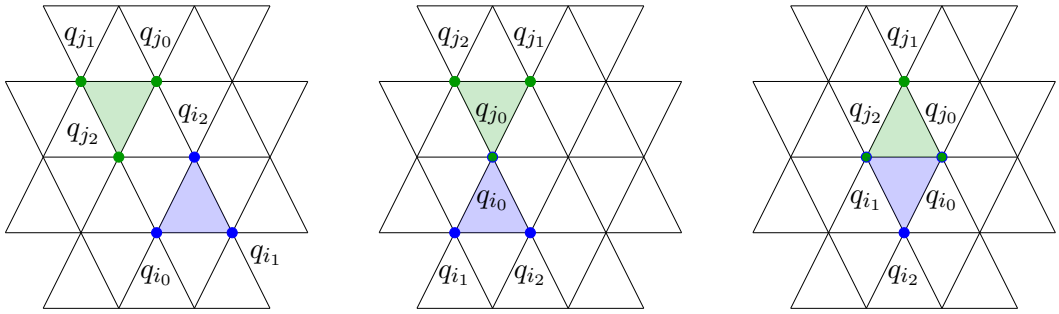
$$d_Q(i_1; [j_0, j_1, j_2]) \geq \frac{1}{\sqrt{2}} \min\{D_Q(i_1; [j_0, j_1]), D_Q(i_1; [j_1, j_2]), D_Q(i_1; [j_2, j_0])\}$$

and the same with i_1 replaced by i_2 and show that these expressions are bounded away from zero for $Q = Q_n$. Regardless of whether $[i_1]$ and $[i_2]$, and $[j_0, j_1]$, $[j_1, j_2]$, $[j_2, j_0]$ are interior or boundary faces, in each case, one of proposition 5.2.5, statements (a) to (c) applies and provides this lower bound. The same argument applies with the roles of σ and σ' reversed. We can conclude that $\sigma_* \cap \sigma'_* = \{q^*_{i_0}\} = \{q^*_{j_0}\}$ as desired. • Third, if σ and σ' intersect on a common edge, which is necessarily an interior 1-face. Without loss of generality, suppose that $[i_0] = [j_0]$ and $[i_1] = [j_2]$, see figure 5.19c for an illustration. We need to estimate only the distances from $[i_2]$ to $[j_0, j_1, j_2]$ and from $[j_1]$ to $[i_0, i_1, i_2]$. To this end, we use

$$d_{Q_n}(i_2; [j_0, j_1, j_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b} \sqrt{1 - \Psi^2(b)} > 0, \quad \text{and}$$
$$d_{Q_n}(j_1; [i_0, i_1, i_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b} \sqrt{1 - \Psi^2(b)} > 0,$$

which follow proposition 5.2.5, statements (a) to (c). Therefore, the intersection $\sigma_* \cap \sigma'_*$ equals to $\mathrm{conv}_{Q^*}(i_0, i_1)$ and to $\mathrm{conv}_{Q^*}(j_2, j_0)$ as desired. Moreover, $q^*_{i_0}$ and $q^*_{i_1}$ are going to remain distinct points since $\|q^*_{i_0} - q^*_{i_1}\| = E^2_{Q^*}(i_0, i_1, i_2)$, which remains bounded away from zero by lemma 5.2.3, statement (c).

Finally, we consider the case when one of the faces is interior and the other one is boundary. We assume, in what follows, σ is interior and σ' is boundary. Clearly, the proof is not affected if the roles of σ and σ' are reversed.



(a) $\mathrm{conv}_Q(i_0, i_1, i_2)$ does not intersect $\mathrm{conv}_Q(j_0, j_1, j_2)$.

(b) $\mathrm{conv}_Q(i_0, i_1, i_2)$ intersects $\mathrm{conv}_Q(j_0, j_1, j_2)$ at $\{q_{i_0}\} = \{q_{j_0}\}$.

(c) $\mathrm{conv}_Q(i_0, i_1, i_2)$ intersects $\mathrm{conv}_Q(j_0, j_1, j_2)$ at $\mathrm{conv}_Q(i_0, i_1) = \mathrm{conv}_Q(j_0, j_2)$.

Figure 5.19. Illustration of the case when $[i_0, i_1, i_2]$ (blue) and $[j_0, j_1, j_2]$ (green) are interior 2-faces, from proof of theorem 5.2.6

σ is interior 0-face and σ' is boundary 0-face: Since σ and σ' are distinct faces; therefore, $\sigma \cap \sigma' = \emptyset$. We consider two cases. • First, if $[i_0, j_0]$ is a 1-face of Δ, there exists $[k]$ such that $[i_0, j_0, k] \in \Delta$ (since Δ is pure). Therefore, $\|q^n_{i_0} - q^n_{j_0}\|$ coincides with some edge length $E^\ell_{Q_n}(i_0, j_0, k)$ with $\ell = 0, 1, 2$. From lemma 5.2.3 , statement (c), we know $E^\ell_{Q_n}(i_0, j_0, k) \geq (\alpha_1)/f_1(Q_n; Q_{\mathrm{ref}}) \geq (\alpha_1)/b$ and thus $\{q_{i_0}\}$ and $\{q_{j_0}\}$ do not intersect. • Second, if $[i_0, j_0]$ is not a 1-face of Δ, then there exists j_1 such that $[j_0, j_1]$ is a boundary 1-face. Using proposition 5.2.5, statement (b) we know $D_{Q_n}(i_0; [j_0, j_1]) \geq \alpha_1/b$ and thus $D_{Q^*}(i_0; [j_0, j_1]) \geq \alpha_1/b$. In particular, $\|q^*_{i_0} - q^*_{j_0}\| \geq \alpha_1/b$. Thus, $\sigma_* \cap \sigma'_* = \emptyset$.

σ is an interior 0-face and σ' is an boundary 1-face: • By definition of boundary 0-face, we know σ and σ' cannot intersect. Otherwise, it will contradict the fact that σ' is a boundary 1-face. Therefore, using proposition 5.2.5, statement (b) we obtain $D_{Q_n}(i_0; [j_0, j_1]) \geq \alpha_1/b$ and thus $D_{Q^*}(i_0; [j_0, j_1]) \geq \alpha_1/b$, i. e., $\sigma_* \cap \sigma'_* = \emptyset$.

σ is an interior 0-face and σ' is a boundary 2-face: Recall the expression of the distance of a vertex to a triangle given in (5.12). • Now, we suppose that $\sigma = [i_0]$ and $\sigma' = [j_0, j_1, j_2]$ do not intersect. Since σ is an interior face, and in virtue of proposition 5.2.5, statement (b) we obtain that all the distances D_{Q_n} between $[i_0]$ (interior 0-face) and the 1-faces of σ' are bounded away from zero by α_1/b. Thus, $d_{Q^*}(i_0; [j_0, j_1, j_2]) > \alpha_1/b$. Therefore, $\sigma_* \cap \sigma'_* = \emptyset$. • Second, if σ and σ' intersect, then necessarily $\sigma \cap \sigma' = \sigma$. We use lemma 5.2.3, statements (a) and (c) to prove $\sigma_* \cap \sigma'_* = \sigma_*$.

σ is an interior 1-face and σ' is boundary 1-face: We use again the distance between two edges $d_Q([i_0, i_1]; [j_0, j_1])$ given by (5.13). • First suppose that $\sigma = [i_0, i_1]$ and $\sigma' = [j_0, j_1]$ do not intersect. Since σ is an interior 1-face, at least one of its 0-face is interior, w.l.o.g. we assume $[i_0]$ is an interior 0-face. Therefore, by proposition 5.2.5, statement (b), the term $D_{Q_n}(i_0; [j_0, j_1])$ appearing on the right-hand side of (5.13) is bounded from below by α_1/b. The 0-face $[i_1]$ can be either interior or boundary, in both cases, one can use proposition 5.2.5, statement (a) or statement (b), to bound the term $D_{Q_n}(i_1; [j_0, j_1])$. On the other hand, since σ' is boundary then both of its 0-faces are boundary, and therefore using proposition 5.2.5, statement (c), we obtain bounds for $D_{Q_n}(j_0; [i_0, i_1])$ and $D_{Q_n}(j_1; [i_0, i_1])$, and thus for $D_{Q^*}(j_0; [i_0, i_1])$ and $D_{Q^*}(j_1; [i_0, i_1])$. Altogether, allows us to conclude $\sigma_* \cap \sigma'_* = \emptyset$ as desired. •

Second, when $\sigma = [i_0, i_1]$ and $\sigma' = [j_0, j_1]$ intersect, they intersect in a vertex. Without loss of generality, $[i_0] = [j_0]$ holds and necessarily it is a boundary 0-face. If $[i_0]$ is boundary, then $[i_1]$ is interior and therefore, one can use proposition 5.2.5, statement (b) to bound $D_{Q^*}(i_1; [j_0, j_1])$ from below by α_2/b. In the same way, to bound from below the term $D_{Q^*}(j_1; [i_0, i_1])$ by α_2/b, one can use proposition 5.2.5, statement (c). The edge lengths $\|q^*_{i_0} - q^*_{i_1}\|$ and $\|q^*_{j_0} - q^*_{j_1}\|$ remain positive by lemma 5.2.3, statement (c). This implies $\sigma_* \cap \sigma'_* = \{q^*_{i_0}\} = \{q^*_{j_0}\}$.

σ is an interior 1-face and σ' is a boundary 2-face: We use again the distance of an edge to a triangle $d_Q([i_0, i_1]; [j_0, j_1, j_2])$ given in (5.14). • First suppose that $\sigma = [i_0, i_1]$ and $\sigma' = [j_0, j_1, j_2]$ do not intersect. Then, some of the terms on the right-hand side of (5.14) with Q replaced by Q_n are distances between interior vertices and interior edges, for which proposition 5.2.5, statement (b) provides the lower bound α_1/b. Expression (5.14) also contains terms which are distances between interior vertices and boundary edges, and boundary vertices and interior edges, which can be bounded from below by proposition 5.2.5, statement (b), and statement (c), respectively. Altogether, we obtain again:

$$d_{Q_n}([i_0, i_1]; [j_0, j_1, j_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b} \sqrt{1 - \Psi^2(b)} > 0.$$

• Second, assume that σ and σ' intersect in a vertex. Without loss of generality, $[i_0] = [j_0]$ holds. We recall the distance of a vertex to a triangle given in (5.12). As above, some of the terms on the right-hand side of (5.12) with Q replaced by Q_n are distances between the vertex $[i_1]$ which can be either boundary or interior and interior edges. An application of proposition 5.2.5, statement (b) or statement (c) yields a uniformly positive lower bound for $d_{Q_n}(i_1; [j_0, j_1, j_2])$ and thus for $d_{Q^*}(i_1; [j_0, j_1, j_2])$. This implies $\sigma_* \cap \sigma'_* = \{q^*_{i_0}\} = \{q^*_{j_0}\}$. • Third, assume that $\sigma \cap \sigma' = \sigma$. Without loss of generality, suppose $[i_0] = [j_0]$ and $[i_1] = [j_1]$. We use the fact that $d_{Q_n}(j_2; [i_0, i_1]) \geq h^{\ell}_{Q_n}(j_0, j_1, j_2)$, for $\ell = 2$. Thus, lemma 5.2.3, statement (a) yields a lower bound of α_1/b. In the same way, the edge length $\|q_{i_0} - q_{i_1}\|$ remains positive by lemma 5.2.3, statement (c). Altogether, allows us to conclude $\sigma_* \cap \sigma'_* = \sigma_*$.

σ is an interior 2-face and σ' is boundary 2-face: We recall the definition of the distance between two triangles given in (5.15). • First suppose that $\sigma = [i_0, i_1, i_2]$ and $\sigma' = [j_0, j_1, j_2]$ do not intersect. Each term on the right-hand side of (5.15) with Q replaced by Q_n can be estimated below by proposition 5.2.5, statements (b) and (c). Since these bounds depend on the cosines of the interior angles, we use lemma 5.2.3, statement (f) to obtain a uniform positive lower bound, which in turn implies $\sigma_* \cap \sigma'_* = \emptyset$. • Second, when σ and σ' intersect in a vertex, which without loss of generality, can be assumed to be $[i_0] = [j_0]$. We consider the distance of a vertex to a triangle as follows:

$$\begin{aligned} &d_Q(i_1; [j_0, j_1, j_2]) \\ &\quad \geq \frac{1}{\sqrt{2}} \min\{D_Q(i_1; [j_0, j_1]), D_Q(i_1; [j_1, j_2]), D_Q(i_1; [j_2, j_0])\} \end{aligned}$$

and the same with i_1 replaced by i_2 and show that these expressions are bounded away from zero for $Q = Q_n$. Regardless of whether $[i_1]$ and $[i_2]$, and $[j_0, j_1]$, $[j_1, j_2]$, $[j_2, j_0]$ are interior or boundary faces, in each case, one of proposition 5.2.5, statements (a) to (c) applies and provides this lower bound. The same argument applies with the roles of σ and σ' reversed. We can conclude that $\sigma_* \cap \sigma'_* = \{q^*_{i_0}\} = \{q^*_{j_0}\}$ as desired. • Third, if σ and σ' intersect in a common edge, which

is necessarily an interior 1-face. Without loss of generality, suppose that $[i_0] = [j_0]$ and $[i_1] = [j_2]$. We need to estimate only the distances from $[i_2]$ to $[j_0, j_1, j_2]$ and from $[j_1]$ to $[i_0, i_1, i_2]$. To this end, we use

$$d_{Q_n}(i_2; [j_0, j_1, j_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b}\sqrt{1 - \Psi^2(b)} > 0, \quad \text{and}$$
$$d_{Q_n}(j_1; [i_0, i_1, i_2]) \geq \frac{\min\{\alpha_1, \alpha_2\}}{2b}\sqrt{1 - \Psi^2(b)} > 0,$$

which follow proposition 5.2.5, statements (a) to (c). Therefore, the intersection $\sigma_* \cap \sigma'_*$ equals to $\mathrm{conv}_{Q^*}(i_0, i_1)$ and to $\mathrm{conv}_{Q^*}(j_2, j_0)$ as desired. Moreover, $q^*_{i_0}$ and $q^*_{i_1}$ are going to remain distinct points since $\|q^*_{i_0} - q^*_{i_1}\| = E^{\ell}_{Q^*}(i_0, i_1, i_2)$, with $\ell = 2$, which remains bounded away from zero by lemma 5.2.3, statement (c).

To summarize, we conclude that the limiting simplices σ_*, σ'_* intersect in the same way as σ_n and σ'_n, and the dimension of this intersection is in turn dictated by the underlying abstract simplicial complex. We have thus shown that $\Sigma_\Delta(Q^*)$ is a simplicial 2-complex whose associated abstract simplicial complex is Δ, which concludes the proof of statement *(ii)*. Statements *(i)* and *(ii)* together show that Q^* belongs to $\mathcal{M}_+(\Delta)$. It remains to confirm statement *(iii)*. To this end, we will first argue that $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ is closed, as follows. Since $\mathcal{M}_+(\Delta)$ is locally connected, its connected components and path components agree. Since connected components are closed, and $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ is by definition a path component of $\mathcal{M}_+(\Delta)$, then $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ is closed in $\mathcal{M}_+(\Delta)$ (we refer the reader to Lee, 2011, Ch. 4 for more details.)

Having shown that Q_n converges to Q^* in $\mathcal{M}_+(\Delta)$ and $(Q_n) \subset \mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$, we can conclude that Q^* belongs to $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ as well. Finally, using the continuity of f_1 on $\mathcal{M}_+(\Delta)$, see lemma 5.2.2, and thus on $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$, we infer that $Q^* \in f_1^{-1}(K)$ holds. This confirms the sequential compactness of $f_1^{-1}(K)$ in $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ and concludes the proof. □

Remark 5.2.7. *For the proof of theorem 5.2.6, we used that Δ is a connectivity complex according to definition 4.2.5, i. e., Δ is a pure, abstract simplicial 2-complex, which is 2-path connected. The purity was used whenever we embedded a lower-dimensional face into a 2-face. The 2-path connectedness enters through proposition 5.2.5.*

Although theorem 5.2.6 shows the properness of f_1 on $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$, unfortunately we cannot directly use it in the construction of a complete metric as suggested in Gordon, 1973, Thm. 1 since f_1 lacks differentiability due to the occurrence of the distance of a vertex to and edge $Q \mapsto D_Q$ in (5.6), which is only Lipschitz. Therefore, we replace D_Q by a regularized function $Q \mapsto D^{\mu}_Q$ of class $\mathcal{C}^3$. This then gives rise to the following regularized function $f_1^{\mu} \colon \mathcal{M}_+(\Delta) \to \mathbb{R}$:

$$f_1^{\mu}(Q; Q_{\mathrm{ref}}) \coloneqq \sum_{k=1}^{N_T}\sum_{\ell=0}^{2} \frac{\alpha_1}{h^{\ell}_Q\left(i^k_0, i^k_1, i^k_2\right)} + \sum_{[j_0,j_1]\in E_\partial} \sum_{\substack{i_0 \in V_\partial \\ i_0 \neq j_0, j_1}} \frac{\alpha_2}{D^{\mu}_Q(i_0; [j_0, j_1])} + \frac{\alpha_3}{2}\|Q - Q_{\mathrm{ref}}\|_F^2, \tag{5.16}$$

which we can use in place of (5.6). Provided that we choose $f_1^{\mu} \geq f_1$, proposition 5.2.8 below implies that f_1^{μ} is proper as well.

Indeed, we are providing in appendix B an entire family of regularized functions f_1^{μ} with parameter μ which, in addition to satisfying $f_1^{\mu} \geq f_1$, approximate f_1 arbitrarily well.

Now, we focus on proving the properness of the regularized function f_1^{μ} given in (5.16).

Proposition 5.2.8. *Let X be a metric space and consider a function $f \colon X \to \mathbb{R}$ which is proper. Suppose that $\widehat{f} \colon X \to \mathbb{R}$ is continuous and it satisfies $0 \leq f \leq \widehat{f}$. Then $\widehat{f}$ is proper.*

PROOF. Let $K \subset \mathbb{R}$ be compact. We need to prove that $(\widehat{f})^{-1}(K)$ is compact. In case $(\widehat{f})^{-1}(K) = \emptyset$, nothing is to be shown. Otherwise, since $\widehat{f}$ is nonnegative, we can suppose that $K \subset [a, b]$ with $a \geq 0$. Since $(\widehat{f})^{-1}(K) \subset X$ and X is a metric space, the compactness of $(\widehat{f})^{-1}(K)$ is equivalent to its sequential compactness. To show the latter, consider a sequence $(x_n) \subset (\widehat{f})^{-1}(K)$, i. e., $\widehat{f}(x_n) \in K$. Using the assumption $0 \leq f \leq \widehat{f}$, we obtain $0 \leq f(x_n) \leq \widehat{f}(x_n) \in K \subset [a, b]$ and therefore $f(x_n) \in [0, b]$ for all $n \in \mathbb{N}$. Since f is proper and $[0, b]$ is compact, there exists a subsequence, still labeled (x_n), such that $x_n \to x^*$ in $[0, b]$. Thanks to the continuity of $\widehat{f}$, we have $\widehat{f}(x_n) \to \widehat{f}(x^*)$. Since all $\widehat{f}(x_n) \in K$, the limit $\widehat{f}(x^*)$ belongs to K as well. □

This allows us to present the following theorem.

Theorem 5.2.9. *Suppose that $\alpha_1, \alpha_2, \alpha_3 > 0$ holds. Consider a continuous function on $\mathcal{M}_+(\Delta)$, such that $Q \mapsto D_Q^\mu$ which satisfies $0 < D_Q^\mu \leq D_Q$. Then the following statements hold.*

(a) The restriction of f_1^μ to $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ defined in (5.16) is proper.

(b) Suppose in addition that $Q \mapsto D_Q^\mu$ is of class $\mathcal{C}^3$ on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. Then $\mathcal{M}_+(\Delta; Q_{\text{ref}})$, endowed with the Riemannian metric whose components (with respect to the vec *chart) are given by*

$$g_{ab}^{\text{complete}} = \delta_a^b + \frac{\partial f_1^\mu}{\partial (\operatorname{vec} Q)^a} \frac{\partial f_1^\mu}{\partial (\operatorname{vec} Q)^b}, \quad a, b = 1, \ldots, 2\,N_V, \tag{5.17}$$

is geodesically complete.

PROOF. The definition of f_1 in (5.6), the definition of f_1^μ in (5.16) and the assumption $0 < D_Q^\mu \leq D_Q$ imply $0 < f_1 \leq f_1^\mu$ on $\mathcal{M}_+(\Delta)$. Statement (a) now follows from proposition 5.2.8. Statement (b) follows immediately from Gordon, 1973, Thm. 1. □

Remark 5.2.10. *Theorem 5.2.9 remains valid when (5.16) is replaced by the slightly more general function which w.l.o.g. we denote in the same way.*

$$f_1^\mu(Q; Q_{\text{ref}}) := \sum_{k=1}^{N_T} \sum_{\ell=0}^{2} \chi_1\left(\frac{\alpha_1}{h_Q^\ell(i_0^k, i_1^k, i_2^k)}\right) + \sum_{[j_0,j_1] \in E_\partial} \sum_{\substack{i_0 \in V_\partial \\ i_0 \neq j_0, j_1}} \chi_2\left(\frac{\alpha_2}{D_Q^\mu(i_0; [j_0, j_1])}\right) + \frac{\alpha_3}{2} \|Q - Q_{\text{ref}}\|_F^2. \tag{5.18}$$

Here χ_1 is a cut-off function of class C^3 which satisfies $\chi_1(s) = 0$ on some interval $[0, \underline{s}]$ and $\chi_1(s) = s$ for $s \geq \overline{s}$. The same holds for χ_2. In other words, the first and second terms, which were seen to be responsible to avoid interior and exterior self-intersections, respectively, can safely be turned off when the heights, or the distances of boundary vertices to nonincident boundary edges, respectively, are larger than a threshold $1/\underline{s}$. We will exploit this in our numerical experiments.

Now, we study an invariance property of the proposed metric (5.17). Moreover, we show that this metric agrees with the Euclidean metric for tangent vectors representing translations in case $\alpha_3 = 0$.

Proposition 5.2.11. *Suppose that $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ is defined by $T(x) = R\,x + b$ with $R \in SO(2)$ and $b \in \mathbb{R}^2$. We extend R and T to $\mathbb{R}^{2 \times N_V}$, operating column by column. We denote by $g_Q\colon \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}}) \times \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}}) \to \mathbb{R}$ the metric (5.17) at an arbitrary point $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$. Moreover, we denote by $\overline{g}_Q$ the metric similar to (5.17) obtained by replacing*

Q_{ref} *by* $T(Q_{\text{ref}})$ *in* (5.16). *Then*

$$g_Q(V,W) = \overline{g}_{T(Q)}(RV, RW) \tag{5.19}$$

holds for all $V, W \in \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$.

Proof. Since T is a rotation and translation, $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ implies $T(Q) \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$. Moreover, the heights in the first term of (5.16) depend only on the relative positions of the vertices to each other, i. e.,

$$h^\ell_Q\big(i_0, i_1, i_2\big) = h^\ell_{T(Q)}\big(i_0, i_1, i_2\big)$$

holds for all 2-faces $[i_0, i_1, i_2]$. Similarly, the distances of a vertex to an edge are also invariant with respect to rotation/translation, i. e.,

$$D^\mu_Q(i_0; [j_0, j_1]) = D^\mu_{T(Q)}(i_0; [j_0, j_1])$$

holds for all boundary vertices i_0 and nonincident boundary edges $[j_0, j_1]$. This shows that the second sum in (5.16) is also invariant with respect to T. Finally, we have $\|Q - Q_{\text{ref}}\|_F = \|T(Q - Q_{\text{ref}})\|_F$. This shows that $f_1^\mu(Q; Q_{\text{ref}}) = f_1^\mu(T(Q); T(Q_{\text{ref}}))$ holds, where the right hand side term uses $T(Q_{\text{ref}})$ in place of Q_{ref} in (5.16). Equation (5.19) now follows easily from the chain rule. □

Proposition 5.2.12. *We denote by* g_Q *the metric* (5.17) *at an arbitrary point* Q *belonging to* $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. *Suppose that* $V \in \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}}) \cong \mathbb{R}^{2\times N_V}$ *is a tangent vector satisfying* $V = \big[V_0, V_0, \dots, V_0\big]$ *for some* $V_0 \in \mathbb{R}^2$, *representing a translation. Moreover,* $W = \big[W_1, W_2, \dots, W_{N_V}\big]$ *is an arbitrary vector* $\in \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$. *If* $\alpha_3 = 0$ *holds, then*

$$g_Q(V,W) = (\operatorname{vec} V) \cdot (\operatorname{vec} W) = \sum_{j=1}^{N_V} V_0 \cdot W_j \tag{5.20}$$

holds, i. e., the action of (5.17) *on* (V, W) *agrees with the action of the Euclidean metric.*

Proof. We consider a representative term for the first and second sum in (5.16). Suppose that $[i_0, i_1, i_2]$ is a 2-face of Δ. Since heights and distances are translation invariant, we have $h^\ell_Q(i_0, i_1, i_2) = h^\ell_{Q+tV}(i_0, i_1, i_2)$ as well as $D^\mu_Q(i_0; [j_0, j_1]) = D^\mu_{Q+tV}(i_0; [j_0, j_1])$ for all $t \in \mathbb{R}$. Consequently, the directional derivative of f_1^μ at Q, in the direction of V, is equal to zero. Taking into account the definition (5.17) of the metric, the claim follows. □

An immediate consequence of proposition 5.2.12 is that geodesics with respect to the metric (5.17), whose initial tangent vectors V represent a translation, will be identical to Euclidean geodesics, i. e., $\gamma(t) = Q + t\,V$ holds, provided that $\alpha_3 = 0$ holds.

Summarizing, by proving the completeness of the metric associated with f_1^μ given in (5.18), we could update any mesh, or being more precise, the distribution of the nodes of a mesh with a given connectivity, for any initial tangent vector and as long as we need without jeopardizing its quality. Considering that our main goal is to use complete metrics on the numerical solution of shape optimization problems, the following conditions are desirable on an augmentation function:

(a) invariant under rigid body motions (translations and rotations),
(b) invariant under uniform mesh refinements,

besides of the already known properness, and $\mathcal{C}^3$-regularity properties. The function f_1^μ defined in (5.16) already satisfies condition (a), obtained as an intermediate step in the proof of proposition 5.2.11. Unfortunately, f_1^μ does not satisfy condition (b); indeed, the values the function f_1^μ can attain depend directly on the heights of the triangles. Thus, the finer the mesh is the higher the values of f_1^μ are, even if no cell is pending to self-intersection.

This behavior can be misleading to the algorithm if our goal is to consider realistic meshes. For this reason, in the next section we propose a second proper function f_2^μ which besides of rendering a complete metric is invariant under uniform refinements of the mesh.

5.3 Metric Invariant under Uniform Mesh Refinements

As already mentioned this section aims to construct a second complete metric, which besides of being invariant under rigid body motions, is also invariant under uniform mesh refinements. As per Gordon, 1973, Thm. 1, the complete metric is generated from a function f_2^μ, which additionally to being proper and $\mathcal{C}^3$, satisfies conditions (a) and (b). The construction of this function is based on a well-known triangle quality measure

$$\frac{(E^0)^2 + (E^1)^2 + (E^2)^2}{4\sqrt{3}A} \tag{5.21}$$

for the cells in a finite element mesh, first introduced in Bhatia, Lawrence, 1990; see also Shewchuk, 2002, Tab. 6, Row 4. Here E^ℓ ($\ell = 0, 1, 2$) denotes the lengths of the edges, and A refers to the area of a triangular cell.

Our proposal for f_2 inherits the terms involving the coefficients α_2 and α_3 from f_1 in (5.16). However, the α_1-term, which penalizes small heights and serves to avoid interior self-intersections, is replaced by a term involving the triangle quality measure. Since the latter does not take into account the absolute size of a triangle but only its shape, we also add a term which avoids the total area of the mesh going to zero. Exterior self-intersections, on the other hand, are avoided by a term which agrees with the α_2-term in (5.16).

Definition 5.3.1. *Suppose that Δ and Q_{ref} are as in definition 4.3.10. Denote by V_∂ the set of the boundary 0-faces and by E_∂ the set of boundary 1-faces. Their cardinalities are denoted by $\#V_\partial$ and $\#E_\partial$, respectively. Suppose that the 2-faces in Δ are numbered from 1 to N_T and that the k-th triangle has vertices $\{q_{i_0^k}, q_{i_1^k}, q_{i_2^k}\}$. For parameters $\beta_j \geq 0$, for $j = 1, 2, 3, 4$, define $f_2^\mu \colon \mathcal{M}_+(\Delta; Q_{\mathrm{ref}}) \to \mathbb{R}$ as:*

$$\begin{aligned} f_2^\mu(Q; Q_{\mathrm{ref}}) \coloneqq & \sum_{k=1}^{N_T} \frac{1}{N_T} \frac{\beta_1}{\psi_Q(i_0^k, i_1^k, i_2^k)} + \frac{\beta_2}{\sum_{k=1}^{N_T} A_Q\big(i_0^k, i_1^k, i_2^k\big)} \\ & + \sum_{[j_0, j_1] \in E_\partial} \sum_{\substack{i_0 \in V_\partial \\ i_0 \neq j_0, j_1}} \frac{1}{\#E_\partial \# V_\partial} \frac{\beta_3}{D_Q^\mu(i_0; [j_0, j_1])} + \frac{\beta_4}{2N_V} \|Q - Q_{\mathrm{ref}}\|_F^2 \end{aligned} \tag{5.22}$$

with

$$\frac{1}{\psi_Q(i_0, i_1, i_2)} \coloneqq \frac{\big(E_Q^0(i_0, i_1, i_2)\big)^2 + \big(E_Q^1(i_0, i_1, i_2)\big)^2 + \big(E_Q^2(i_0, i_1, i_2)\big)^2}{4\sqrt{3}\, A_Q(i_0, i_1, i_2)}. \tag{5.23}$$

Recall that the distance of a vertex to an edge D_Q was defined in (4.13), and D_Q^μ, is then a $\mathcal{C}^3$-regularization of D_Q. Moreover, the edge lengths E_Q^ℓ are given in (4.1), the signed area A_Q can be found in (4.2) and $\|\cdot\|_F$ is the Frobenius norm.

In order to prove that f_2^μ is a proper function on $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$, the following result is essential. It shows that on any nonempty sublevel set of f_2^μ, the edge lengths E_Q^ℓ and the reciprocals of the heights $1/h_Q^\ell$ are uniformly bounded independently of the node positions $Q \in \mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$.

Proposition 5.3.2. *Suppose that Δ and Q_{ref} are as in definition 4.3.10. Consider f_2^μ defined in (5.22) with $\beta_1, \beta_2, \beta_3, \beta_4 > 0$. Let $\mathcal{N}_b$ be a nonempty sublevel set of f_2^μ, i. e.,*

$$\mathcal{N}_b \coloneqq \{Q \in \mathcal{M}_+(\Delta; Q_{\mathrm{ref}}) \mid f_2^\mu(Q; Q_{\mathrm{ref}}) \leq b\} = f_2^\mu(\cdot; Q_{\mathrm{ref}})^{-1}((-\infty, b]). \tag{5.24}$$

Then there exist constants $c, C, D > 0$ such that the edge lengths and heights satisfy

$$c \leq E_Q^\ell(i_0^k, i_1^k, i_2^k) \leq C, \tag{5.25}$$

$$\frac{1}{h_Q^\ell(i_0^k, i_1^k, i_2^k)} \leq D \tag{5.26}$$

for all $Q \in \mathcal{N}_b$, all $k = 1, \ldots, N_T$ and all $\ell = 0, 1, 2$. The constants c, C, D are independent from k and ℓ.

Proof. Let us consider $Q \in \mathcal{N}_b$, fixed but arbitrary. With this in mind, now we propose the following simplification in the notation. We will write $E_k^\ell := E_Q^\ell(i_0^k, i_1^k, i_2^k)$, $A_k := A_Q(i_0^k, i_1^k, i_2^k)$, and $\psi_k := \psi_Q(i_0^k, i_1^k, i_2^k)$ This is done, since $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ will be arbitrary but fixed and we want to highlight the dependence of these quantities on the triangle (indexed by k) and the vertex (indexed by ℓ). The proof is broken down into several steps:

(1) We will find upper and lower bounds for the edge length of one specific edge denoted as $E_{\bar{k}}^{\bar{\ell}}$ of the $\bar{k}$-th triangle.
(2) Using the bounds from step (1) we will find bounds for the remaining edges of the $\bar{k}$-th triangle, i. e., $E_{\bar{k}}^{\bar{\ell}\oplus 1}$ and $E_{\bar{k}}^{\bar{\ell}\oplus 2}$, where $\oplus$ denotes addition modulo 3.
(3) Compute the bounds for all the heights $h_{\bar{k}}^\ell$ of the $\bar{k}$-th triangle using step (2).
(4) We consider an arbitrary triangle k different from $\bar{k}$, and based on the 2-path connectedness of Δ we will use the bounds from steps (1) and (2) to bound all the edges of the k-th triangle.

Since $Q \in \mathcal{N}_b$, from the definition of f_2^μ given in (5.22) we immediately obtain

$$\sum_{k=1}^{N_T} A_k \geq \frac{\beta_2}{b}, \tag{5.27}$$

since $A_k \geq 0$ for all $k = 1, \ldots, N_T$, then, we know there exists at least one triangle $\bar{k}$ such that

$$A_{\bar{k}} \geq \frac{\beta_2}{N_T b}.$$

Using the so-called isoperimetric inequality for triangles given in (4.8) which states

$$A_{\bar{k}} \leq \frac{\left(E_{\bar{k}}^0 + E_{\bar{k}}^1 + E_{\bar{k}}^2\right)^2}{12\sqrt{3}}, \tag{5.28}$$

and by denoting $E_{\bar{k}}^{\bar{\ell}} := \max_{\ell=0,1,2}\{E_{\bar{k}}^\ell\}$, we obtain

$$E_{\bar{k}}^{\bar{\ell}} \geq \frac{2}{3^{1/4}}\left(\frac{\beta_2}{N_T b}\right)^{1/2} > 0. \tag{5.29}$$

Notice moreover, that since $Q \in \mathcal{N}_b$, then $\|Q - Q_{\text{ref}}\|_F^2 \leq {}^{2N_V b}/_{\beta_4}$, which implies $\|Q\|_F \leq \sqrt{{}^{2N_V b}/_{\beta_4}} + \|Q_{\text{ref}}\|_F$. We denote by $\bar{i}_0$ and $\bar{i}_1$ the vertices which form the edge whose edge length is $E_{\bar{k}}^{\bar{\ell}}$. Then, $E_{\bar{k}}^{\bar{\ell}} = \|q_{\bar{i}_0} - q_{\bar{i}_1}\| \leq \|q_{\bar{i}_0}\| + \|q_{\bar{i}_1}\| \leq \sqrt{2}\,\|Q\|_F$. Thus, $E_{\bar{k}}^{\bar{\ell}} \leq 2\sqrt{{}^{N_V b}/_{\beta_4}} + \sqrt{2}\|Q_{\text{ref}}\|_F$. Altogether implies

$$\frac{2}{3^{1/4}}\left(\frac{\beta_2}{N_T b}\right)^{1/2} \leq E_{\bar{k}}^{\bar{\ell}} \leq 2\sqrt{\frac{N_V b}{\beta_4}} + \sqrt{2}\|Q_{\text{ref}}\|_F. \tag{5.30}$$

This concludes step (1).

We proceed to find the bounds of $E_{\bar{k}}^{\bar{\ell}\oplus j}$ for $j = 1, 2$, i. e., step (2). Using again the fact that $Q \in \mathcal{N}_b$ and the definition of ψ_k given in (5.23), it follows

$$b \geq f_2^{\mu}(Q; Q_{\text{ref}}) \geq \frac{\beta_1}{N_T} \frac{\left(E_{\bar{k}}^{\bar{\ell}}\right)^2 + \left(E_{\bar{k}}^{\bar{\ell}\oplus 1}\right)^2 + \left(E_{\bar{k}}^{\bar{\ell}\oplus 2}\right)^2}{4\sqrt{3}A_{\bar{k}}}.$$

From the definition of the area we know $A_k = 0.5E_{\bar{k}}^{\bar{\ell}}h_{\bar{k}}^{\bar{\ell}}$. Moreover, we have mentioned in (4.6) that the inequalities $h_k^{\ell} \leq E_k^{\bar{\ell}\oplus 1}$ and $h_k^{\ell} \leq E_k^{\ell\oplus 2}$ hold for all $k = 1, \ldots, N_T$ and all $\ell = 0, 1, 2$. We will focus here on the bounds for the edge length $E_{\bar{k}}^{\bar{\ell}\oplus 1}$; however, the bounds for $E_{\bar{k}}^{\bar{\ell}\oplus 2}$, can be obtained using the same arguments. Therefore, $A_{\bar{k}} \leq 0.5\left(2\sqrt{N_V b/\beta_4} + \sqrt{2}\|Q_{\text{ref}}\|_F\right) E_{\bar{k}}^{\bar{\ell}\oplus 1}$. Thus,

$$b \geq f_2^{\mu}(Q; Q_{\text{ref}}) \geq \frac{\beta_1}{N_T} \frac{E_{\bar{k}}^{\bar{\ell}\oplus 1}}{2\sqrt{3}\left(2\sqrt{N_V b/\beta_4} + \sqrt{2}\|Q_{\text{ref}}\|_F\right)},$$

which implies $E_{\bar{k}}^{\bar{\ell}} \leq (2\sqrt{3}N_T b/\beta_1)\left(2\sqrt{N_V b/\beta_4} + \sqrt{2}\|Q_{\text{ref}}\|_F\right)$. In the same way,

$$b \geq f_2^{\mu}(Q; Q_{\text{ref}}) \geq \frac{\beta_1}{N_T} \frac{(2/3^{1/4})(\beta_2/N_T b)^{1/2}}{2\sqrt{3}E_{\bar{k}}^{\bar{\ell}\oplus 1}}$$

leads to $E_{\bar{k}}^{\bar{\ell}\oplus 1} \geq \beta_1\beta_2^{1/2}/3^{3/4}(N_T b)^{3/2}$, and thus, we have concluded step (2).

The bounds from step (3) are immediately obtained from noticing that $A_{\bar{k}} = (0.5)E_{\bar{k}}^{\bar{\ell}}h_{\bar{k}}^{\bar{\ell}} = (0.5)E_{\bar{k}}^{\bar{\ell}\oplus 1}h_{\bar{k}}^{\bar{\ell}\oplus 1}$ and using the bounds from steps (1) and (2). Thus, $1/h_{\bar{k}}^{\bar{\ell}} \leq 3^{3/4}(N_T b)^{3/2}/\beta_1\beta_2^{1/2}$, and $1/h_{\bar{k}}^{\bar{\ell}\oplus 1} \leq 2\,3^{5/4}(N_T b)^{5/2}/\beta_1^2\beta_2^{1/2}$.

Finally, we focus on step (4). Having found the constants for the $\bar{k}$-th triangle, we will use it as a pivot to compute the constants for the remaining triangles, based on the 2-path connectedness of Δ. To this end, we consider an arbitrary triangle k, different from the $\bar{k}$-th triangle. From all the possible paths joining the $\bar{k}$-th and k-th triangles, guaranteed by the 2-path connectedness, we choose a shortest one. Notice moreover, since Δ is a finite collection of simplices, the longest from the shortest paths has finite longitude and we denote it as L.

Suppose that the path joining $\bar{k}$-th triangles and k-th triangle, has m elements, and we denote them with the index i, such that when $i = 0$, the triangle coincides with the $\bar{k}$-th triangle, for which we know

$$\frac{1}{3^{3/4}(N_T b)^{3/2}} \leq E_0^{\ell_0} \leq \left(\frac{2\sqrt{3}N_T b}{\beta_1}\right)\left(2\sqrt{\frac{N_V b}{\beta_4}} + \sqrt{2}\|Q_{\text{ref}}\|_F\right). \tag{5.31}$$

Moreover, we know the triangles 0 and 1 share one edge, and we denote its length w.r.t. the first triangle as $E_1^{\ell_1}$, for which (5.31) also hold. Using the same techniques as before one can prove

$$\frac{\beta_1^2\beta_2^{1/2}}{2\,3^{5/4}(N_T b)^{5/2}} \leq E_1^{\ell_1\oplus 1} \leq \left(\frac{2\sqrt{3}N_T b}{\beta_1}\right)^2\left(2\sqrt{\frac{N_V b}{\beta_4}} + \sqrt{2}\|Q_{\text{ref}}\|_F\right). \tag{5.32}$$

The bounds of the heights can also be computed in the same manner, i. e., it holds:

$$\frac{1}{h_1^{\ell_1}} \leq \frac{2\,3^{5/4}(N_T b)^{5/2}}{\beta_1^2\beta_2^{1/2}}, \qquad \frac{1}{h_1^{\ell_1\oplus 1}} \leq \frac{2^2\,3^{7/4}(N_T b)^{7/2}}{\beta_1^3\beta_2^{1/2}},$$

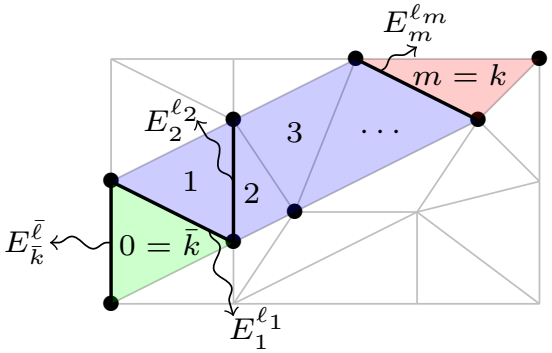


Figure 5.20. Illustration of a mesh associated to a 2-path connected Δ and $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$, used in the proof of proposition 5.3.2

and the same bounds will hold for $E_1^{\ell_1 \oplus 2}$, $1/h_1^{\ell_1 \oplus 2}$, respectively.

By repeating this process until we reach the m-th element of the path, i. e., the k-th triangle. We obtain the following bounds:

$$\frac{\beta_1^{m+1}\,\beta_2^{1/2}}{2^m\,3^{m/2+3/4}(N_T b)^{3/2+m}} \le E_m^{\ell_m} \le \left(\frac{2\sqrt{3}N_T b}{\beta_1}\right)^{m+1} \left(2\sqrt{\frac{N_V b}{\beta_4}} + \sqrt{2}\|Q_{\text{ref}}\|_F\right),$$
$$\frac{1}{h_m^{\ell_m}} \le \frac{2^m\,3^{m/2+3/4}(N_T b)^{m+3/2}}{\beta_1^{m+1}\,\beta_2^{1/2}},$$
$$\frac{1}{h_m^{\ell_m \oplus 1}} \le \frac{2^{m+1}\,3^{m/2+5/4}(N_T b)^{m+3/2}}{\beta_1^{m+2}\,\beta_2^{1/2}}.$$

Since all the constants are greater than one, these bounds hold not only for the m-th triangle of the path, but the whole sequence. Moreover, since we have denoted as L the length of the longest from the shortest paths joining any pair of triangles in Δ, it holds, for all $k = 1, \dots, N_T$ and all $\ell = 0, 1, 2$ that

$$\frac{\beta_1^{L+1}\,\beta_2^{1/2}}{2^L\,3^{L/2+3/4}(N_T b)^{3/2+L}} \le E_k^{\ell} \le \left(\frac{2\sqrt{3}N_T b}{\beta_1}\right)^{L+1} \left(2\sqrt{\frac{N_V b}{\beta_4}} + \sqrt{2}\|Q_{\text{ref}}\|_F\right),$$
$$\frac{1}{h_k^{\ell}} \le \frac{2^L\,3^{L/2+3/4}(N_T b)^{L+3/2}}{\beta_1^{L+1}\,\beta_2^{1/2}},$$
$$\frac{1}{h_k^{\ell \oplus 1}} \le \frac{2^{L+1}\,3^{L/2+5/4}(N_T b)^{L+3/2}}{\beta_1^{L+2}\,\beta_2^{1/2}}.$$

We recall that the constants do neither depend on Q, nor on the chosen pivot triangle $\bar{k}$ and edge $\bar{\ell}$. □

Figure 5.20, shows an illustration of how the path joining the triangles $\bar{k}$ and k will look like, and all the quantities involved in the proof of proposition 5.3.2.

We want to highlight the proof of proposition 5.3.2 builds on the fact that Δ is a connectivity complex in the sense of definition 4.2.5, and in particular it uses the 2-path connectedness of Δ.

Now, we are ready to prove the properness of f_2^μ from (5.22), by relating it to the function f_1^μ given in (5.16), for which properness has already been proved; see theorem 5.2.9.

Theorem 5.3.3. *Suppose that Δ and Q_{ref} are as in definition 4.3.10. Consider the functions f_1^μ from (5.16) and f_2^μ from (5.22) with all coefficients α_j and β_j strictly positive. Then for any sublevel set $\mathcal{N}_b$ of f_2^μ as in (5.24), there exists a constant $B > 0$ such that $\mathcal{N}_b \subset f_1^\mu(\cdot; Q_{\text{ref}})^{-1}([0, B])$. Therefore, f_2^μ is proper.*

Proof. Let us consider node positions $Q \in \mathcal{N}_b$. From proposition 5.3.2 and the definition of f_2^μ, we obtain the following estimates:

$$\sum_{k=1}^{N_T} \sum_{\ell=0}^{2} \frac{1}{h^\ell_{Q^n}(i_0^k, i_1^k, i_2^l)} \leq \frac{3N_T D}{\beta_1},$$
$$\sum_{[j_0,j_1]\in E_\partial} \sum_{\substack{i_0\in V_\partial \\ i_0\neq j_0,j_1}} \frac{1}{D^\mu_{Q^n}(i_0; [j_0, j_1])} \leq \frac{b\#E_\partial\#V_\partial}{\beta_3},$$
$$\frac{1}{2}\|Q^n - Q_{\text{ref}}\|_F^2 \leq \frac{N_V b}{\beta_4}.$$

Recalling the definition of f_1^μ from (5.16), we have:

$$f_1^\mu(Q; Q_{\text{ref}}) \leq 3N_T D\,\frac{\alpha_1}{\beta_1} + b\,\#E_\partial\,\#V_\partial\,\frac{\alpha_2}{\beta_3} + b\,\frac{\alpha_3 N_V}{\beta_4} =: B.$$

Since $Q \in \mathcal{N}_b \subset \mathcal{M}_+(\Delta; Q_{\text{ref}})$ holds, we also know $f_1^\mu(Q; Q_{\text{ref}}) \geq 0$, which in turn implies $Q \in f_1^\mu(\cdot; Q_{\text{ref}})^{-1}([0, B])$.

To show the properness of f_2^μ, consider any compact subset K of $\mathbb{R}$. We need to verify that $f_2^\mu(\cdot; Q_{\text{ref}})^{-1}(K)$ is compact in $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. In case $f_2^\mu(\cdot; Q_{\text{ref}})^{-1}(K)$ is empty, nothing is to be shown. Otherwise, we can find an interval $(-\infty, b]$ such that $f_2^\mu(\cdot; Q_{\text{ref}})^{-1}(K) \subset \mathcal{N}_b = f_2^\mu(\cdot; Q_{\text{ref}})^{-1}((-\infty, b])$ holds. In the rest of the proof we are going to show that $\mathcal{N}_b$ is compact. Since f_2^μ is continuous on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$, this then implies that $f_2^\mu(\cdot; Q_{\text{ref}})^{-1}(K)$ is a closed subset of a compact set, and thus also compact.

Let us now prove that $\mathcal{N}_b$ is compact in $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. Since the latter is a metric space (endowed here with the Euclidean metric of $\mathbb{R}^{2\times N_V}$), compactness is equivalent to sequential compactness. Hence, we consider a sequence $\{Q^n\} \subset \mathcal{N}_b$. Thanks to the first part of the proof, Q^n also belongs to $f_1^\mu(\cdot; Q_{\text{ref}})^{-1}([0, B])$. Owing to the properness of f_1^μ (theorem 5.2.9), we know that $f_1^\mu(\cdot; Q_{\text{ref}})^{-1}([0, B])$ is sequentially compact. Therefore, we can extract a subsequence from $\{Q^n\}$, denoted again by $\{Q^n\}$, which converges to some Q^* in $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. Thanks to the continuity of f_2^μ on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$, $Q^* \in \mathcal{N}_b$ holds, which shows the desired sequential compactness of $\mathcal{N}_b$. □

Remark 5.3.4. *Similar to remark 5.2.10, we can add $\mathcal{C}^3$ cut-off functions to various terms in f_2^μ while maintaining the properness of the function. For instance, theorem 5.3.3 remains true when the function f_2^μ given in (5.22) is replaced by*

$$\begin{aligned} f_2^\mu(Q; Q_{\text{ref}}) \coloneqq \sum_{k=1}^{N_T} \frac{1}{N_T} \frac{\beta_1}{\psi_Q(i_0^k, i_1^k, i_2^k)} + \frac{\beta_2}{\sum_{k=1}^{N_T} A_Q(i_0^k, i_1^k, i_2^k)} \\ + \sum_{[j_0,j_1]\in E_\partial} \sum_{\substack{i_0\in V_\partial \\ i_0\neq j_0,j_1}} \frac{\beta_3}{\#E_\partial\#V_\partial} \chi\left(\frac{1}{D^\mu_Q(i_0; [j_0, j_1])}\right) + \frac{\beta_4}{2}\|Q - Q_{\text{ref}}\|_F^2. \end{aligned} \tag{5.33}$$

Here χ is a cut-off function of class $\mathcal{C}^3$ which satisfies $\chi(s) = 0$ on some interval $[0, \underline{s}]$ and $\chi = s$ for $s > \underline{s}$. Similar cut-off functions could be added to any of the three remaining terms in (5.33) as well.

Since f_2^μ is proper by theorem 5.3.3, the following result is a direct consequence of Gordon, 1973, Thm. 1.

Proposition 5.3.5. *Suppose that $\beta_1, \beta_2, \beta_3, \beta_4 > 0$ holds. Then the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$, endowed with the Riemannian metric whose components (w.r.t. the* vec *chart) are given by*

$$g_{ab}^{\text{invariant}} = \delta_a^b + \frac{\partial f_2^\mu}{\partial (\operatorname{vec} Q)^a} \frac{\partial f_2^\mu}{\partial (\operatorname{vec} Q)^b}, \quad a, b = 1, \dots, 2\,N_V, \tag{5.34}$$

is geodesically complete.

Finally, we comment of the further properties of the function f_2^μ.

Remark 5.3.6. *The function f_2^μ is invariant under rigid body motions and uniform mesh refinements. Indeed,*

- *For any triangle, the function $1/\psi_Q$ is bounded below by 1, and this bound is attained if and only if the triangle is equilateral. This is due to the so-called Weitzenböck inequality; see Alsina, Nelsen, 2008.*
- *The invariance of f_2^μ under rigid body motions follows directly from its definition, and can be proved using the same arguments as in the proof of proposition 5.2.11.*
- *The scaling by $N_T, \#E_\partial, \#V_\partial$ and N_V is chosen so as to achieve invariance of f_2^μ under uniform mesh refinement.*

From now on, we will drop out the μ superindex, and work only with the $\mathcal{C}^3$-regularizations of the proposed functions. In the same way, we will use the same notation, at least in the formal sense, no matter if the cut-off functions are active or not.

5.4 Geodesic Equation

Up until now, we have constructed two different complete Riemannian metrics for the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. The natural next step is to describe how we are going to use them. As already mentioned, Riemannian metrics can be used for the computation of the lengths of tangent vectors, transformations of cotangent vectors into tangent vectors, and to navigate the manifold following geodesics. In this section we focus on the study of geodesics for the geodesically complete manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$.

Geodesics are uniquely defined by a choice of an initial point $\gamma(0) \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ and initial velocity $\dot\gamma(0) \in \mathcal{T}_{\gamma(0)}\mathcal{M}_+(\Delta; Q_{\text{ref}})$ as follows.

The curve γ on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is said to be a **geodesic** if and only if its coordinate curves solve the following system of second-order nonlinear ordinary differential equations:

$$\frac{\mathrm{d}^2\gamma^c}{\mathrm{d}t^2} + \sum_{a,b=1}^{d} \Gamma_{ab}^c \frac{\mathrm{d}\gamma^a}{\mathrm{d}t} \frac{\mathrm{d}\gamma^b}{\mathrm{d}t} = 0, \quad c = 1, \dots, 2\,N_V, \tag{5.35}$$

where Γ_{ab}^c are evaluated at $\gamma(t)$. We refer to Γ_{ab}^c as the **Christoffel symbols**, defined by the following expression:

$$\Gamma_{ab}^c = \frac{1}{2} \sum_{e=1}^{d} g_s^{ce} \left(\frac{\partial g_{ea}^s}{\partial x^b} + \frac{\partial g_{eb}^s}{\partial x^a} - \frac{\partial g_{ab}^s}{\partial x^e} \right), \tag{5.36}$$

where g_{ab}^s are the components of the metric tensor and g_s^{ab} the component of its inverse, with $s \in \{\text{complete}, \text{invartiant}\}$.

Unfortunately, the computation of the Christoffel symbols associated to the metrics g_{ab}^{complete} given in (5.17) and $g_{ab}^{\text{invariant}}$ given in (5.34) is intricate. Even if one would be able to compute such quantities explicitly, the resulting geodesic equation will not have a closed expression. For this reason, we decided to numerically approximate its solution. This section is devoted to the presentation of the computational aspects used to this end.

In view of using the Störmer-Verlet scheme, we study the Hamiltonian formulation for geodesics.

5.4.1 Hamiltonian Formulation for Geodesics

The main idea behind this formulation is that geodesics can be equivalently defined as the shortest path joining two points on a manifold. In other words, the geodesic equation is the Euler–Lagrange equation of the following problem:

$$\begin{aligned} \text{Minimize} \quad & \int_a^b \|\dot\gamma(t)\|_g \,\mathrm{d}t \quad \text{w.r.t. } \gamma\colon [a,b] \to \mathcal{M}_+(\Delta; Q_{\text{ref}}), \\ & \text{s.t. } \gamma(a) = Q_1 \text{ and } \gamma(b) = Q_2, \end{aligned} \tag{5.37}$$

where $Q_1, Q_2 \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$, and $\|V\|_g = \sqrt{(V\,,\,V)_g}$. Note that this definition depends on the chosen Riemannian metric. In the language of calculus of variations, the function $L(\gamma, \dot\gamma) = (\dot\gamma(t)\,,\,\dot\gamma(t))_{\gamma(t)}$ is known as the *Lagrangian*. Since this term may be confused with the Lagrangian defined in subsection 2.1.3 for the computation of the optimality conditions of a PDE-constrained problem, we do not use this terminology. However, we will exploit its interpretation within the Hamiltonian formalism. This section is based on the book Dubrovin, Fomenko, Novikov, 1992, particularly sections §31, §32, §33.

In the following we denote by Q^a, $a = 1, \dots, 2\,N_V$, the coordinates of a point $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ with respect the chart $\mathrm{vec}\colon \mathbb{R}^{2\times N_V} \to \mathbb{R}^{2\,N_V}$, which stacks $Q \in \mathbb{R}^{2\times N_V}$ column by column. Similarly, we denote by V^a the components of a tangent vector $V \in \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$ in the chart induced basis of the tangent space. Finally, the components of a cotangent vector $P \in \mathcal{T}^*_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$ are denoted by P_a. Note the difference of notation between sub and superindices. Using the notation $L(Q,V) = \sum_{a,b} g_{ab} Q^a V^b$, then the Euler–Lagrange equation of problem (5.37), is given by the following expression.

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L(Q,V)}{\partial V^a}\right) - \frac{\partial L(Q,V)}{\partial Q^a} = 0. \tag{5.38}$$

Note that the function L is defined on the tangent bundle $\mathcal{T}\mathcal{M}_+(\Delta; Q_{\text{ref}})$.

Furthermore, a Lagrangian L (in the sense of calculus of variations) is said to be nonsingular on a neighborhood of $(Q,V) \in \mathcal{T}\mathcal{M}_+(\Delta; Q_{\text{ref}})$ if

$$\det\left(\frac{\partial^2 L}{\partial V^a \partial V^b}\right) \neq 0.$$

By considering the momentum $P_a = \partial L/\partial V^a$, then L is called strongly nonsingular if for all $a = 1, \dots, N_V$, the equation which defines the momentum, determines a unique smooth function $V^a(Q,P)$ on the given neighborhood. It is common to refer to the space with coordinates (Q,V) as the phase space associated to L. If we assume L is strongly nonsingular, this implies the change of variables (Q,V) to the coordinates (Q,P) is smoothly invertible. In terms of these new variables the function L can be replaced by the Hamiltonian which is now given in terms of the position Q and momentum $P_a = \sum_{a,b} g_{ab}(Q)\, V^b$:

$$H(Q,P) \coloneqq PV - L(Q,V) = \frac{1}{2}\sum_{a,b} g^{ab}(Q)\, P_a\, P_b. \tag{5.39}$$

Note that the Hamiltonian is a function defined on the cotangent bundle $\mathcal{T}^*\mathcal{M}_+(\Delta; Q_{\text{ref}})$.

In our case, $L = \sum_{a,b} g_{ab} Q^a V^b$ is strongly nonsingular because g_{ab} is smooth and invertible. This is obtained immediately from the definition of Riemannian metric. Therefore, the Euler–Lagrange equation can be equivalently reformulated as the following system of

equations:

$$\dot{P}_a = -\frac{\partial H(Q,P)}{\partial Q^a} = \frac{1}{2}\sum_{b,d}\frac{\partial g_{bd}(Q)}{\partial Q^a}\left(\sum_c g^{bc}(Q)\,P_c\right)\left(\sum_e g^{de}(Q)\,P_e\right), \tag{5.40a}$$

$$\dot{Q}^a = +\frac{\partial H(Q,P)}{\partial P_a} = \sum_b g^{ab}(Q)\,P_b, \tag{5.40b}$$

which is usually referred as the *Hamiltonian system*. In (5.40a) we used the rule of the derivative of the inverse function.

One advantage of the Hamiltonian structure (5.40) is that efficient energy preserving integrators had been developed for this kind of systems, such as the Störmer–Verlet scheme.

5.4.2 Störmer–Verlet Scheme for Hamiltonian Systems

The flow of a Hamiltonian system as in (5.40) on a Riemannian manifold $\mathcal{M}$ is defined as $\varphi_t \colon \mathcal{T}^*\mathcal{M} \to \mathcal{T}^*\mathcal{M}$, mapping the initial condition (Q_0, P_0) to the solution at time t. In other words, a geodesic can be interpreted as the Hamiltonian flow associated to the length-squared function given by the metric on the tangent bundle, which is also known as the first fundamental form.

An important geometric property of a Hamiltonian systems is that its flow φ_t is *symplectic*, i. e., the derivative $\varphi_t' = \frac{\partial \varphi_t}{\partial(Q,P)}$ of the flow satisfies

$$\left(\varphi_t'\right)^{\mathrm{T}} J\,\varphi_t' = J \quad \text{with } J = \begin{bmatrix} 0 & \mathrm{id} \\ -\mathrm{id} & 0 \end{bmatrix},$$

where id denotes the identity matrix of the dimension of $\mathcal{M}$. A numerical integrator which preserves this property is said to be *symplectic*. A prominent example is the Störmer–Verlet scheme, see for instance Hairer, Lubich, Wanner, 2003, Eq. (2.10). We assume Einstein's summation convention for the presentation of the algorithm.

Its application to the Hamiltonian system (5.40) is given in algorithm 1. The description of the computational aspects used for the approximation of the geodesics on $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ is shown in what follows.

We need to efficiently implement the following functions:

- Compute the inverse of the matrix representation of the Riemannian metric, which is needed to evaluate the term $g^{bc}(Q)\,P_{c,n}$ (using Einstein's summation convention) in lines 3,4,5, of algorithm 1. We also require the inverse of the matrix in line 6 to evaluate of the current tangent vector as $V^a_{n+1} = g^{ab}(Q_{n+1})\,P_{b,n+1}$.
- Evaluate the derivative of the matrix representation of the Riemannian metric, which is required to assemble the right-hand side term of lines 3 and 5.
- Assemble the functions:

  (1) from line 3, of algorithm 1

$$F_1(P_{n+1/2}) = P_{a,n+1/2} - P_{a,n} + \frac{\Delta t}{4}\left\{\frac{\partial g_{bd}(Q_n)}{\partial Q^a}\,g^{bc}(Q_n)\,P_{c,n+1/2}\,g^{de}(Q_n)\,P_{e,n+1/2}\right\}. \tag{5.41}$$

  (2) From line 4, of algorithm 1

$$F_2(Q_{n+1}) = Q^a_{n+1} - Q^a_n - \frac{\Delta t}{2}\left\{g^{ab}(Q_n)\,P_{a,n+1/2} + g^{ab}(Q_{n+1})\,P_{a,n+1/2}\right\}. \tag{5.42}$$

Algorithm 1: Störmer–Verlet scheme for the geodesic equation (5.35) in Hamiltonian form (5.40).

Data: abstract simplicial complex Δ with N_V vertices
Data: matrix $Q_{\text{ref}} \in \mathcal{M}_+(\Delta) \subset \mathbb{R}^{2\times N_V}$ of reference node positions
Data: matrix $Q_0 \in \mathcal{M}_+(\Delta; Q_{\text{ref}}) \subset \mathbb{R}^{2\times N_V}$ of initial node positions
Data: initial tangent vector $V_0 \in \mathbb{R}^{2\times N_V}$
Data: final time T; number of time steps N; time step size $\Delta t := \frac{T}{N}$
Result: approximate solution of the geodesic equation (5.35) with initial conditions $\gamma(0) = Q_0$ and $\dot\gamma(0) = V_0$ on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ at times $t_n = n\,\Delta t$, $n = 0, \dots, N$

1 set initial momentum $P_{a,0} = g_{ab}(Q_0)\, V_0^b$;
2 **for** $n \leftarrow 0$ **to** $N-1$ **do**
3 Solve $P_{a,n+1/2} = P_{a,n} - \frac{\Delta t}{4}\left\{ \frac{\partial g_{bd}(Q_n)}{\partial Q^a}\, g^{bc}(Q_n)\, P_{c,n+1/2}\, g^{de}(Q_n)\, P_{e,n+1/2} \right\}$, $a = 1, \dots, 2N_V$;
4 Solve $Q^a_{n+1} = Q^a_n + \frac{\Delta t}{2}\{ g^{ab}(Q_n)\, P_{a,n+1/2} + g^{ab}(Q_{n+1})\, P_{a,n+1/2} \}$, $a = 1, \dots, 2N_V$;
5 Set $P_{a,n+1} = P_{a,n+1/2} - \frac{\Delta t}{4}\left\{ \frac{\partial g_{bd}(Q_{n+1})}{\partial Q^a}\, g^{bc}(Q_{n+1})\, P_{c,n+1/2}\, g^{de}(Q_{n+1})\, P_{e,n+1/2} \right\}$, $a = 1, \dots, 2N_V$;
6 Solve $V^a_{n+1} = g^{ab}(Q_{n+1})\, P_{b,n+1}, \quad a = 1, \dots, 2N_V$;
7 **end**
8 **return** $Q^a_0, \dots, Q^a_N$, $a = 1, \dots, 2N_V$, *approximating* $\gamma(t_0), \dots, \gamma(t_N)$
9 **return** $V^a_0, \dots, V^a_N$, $a = 1, \dots, 2N_V$, *approximating* $\dot\gamma(t_n), \dots, \gamma(t_N)$
10 **return** $P_{a,0}, \dots, P_{a,N}$, $a = 1, \dots, 2N_V$, *approximating the momentum* $P_a = g_{ab}(Q)\, V^b$ *at* $t_0, \dots, t_N$

(3) from line 5, of algorithm 1

$$F_3(P_{n+1/2}) = P_{a,n+1/2} - \frac{\Delta t}{4}\left\{ \frac{\partial g_{bd}(Q_{n+1})}{\partial Q^a}\, g^{bc}(Q_{n+1})\, P_{c,n+1/2}\, g^{de}(Q_{n+1})\, P_{e,n+1/2} \right\}. \tag{5.43}$$

– Solve the nonlinear systems

$$F_1(P_{n+1/2}) = 0 \quad \text{and} \quad F_2(Q_{n+1}) = 0, \tag{5.44}$$

where $F_1(\cdot)$ is defined in (5.41), and $F_2(\cdot)$ is defined in (5.42).

Let us start by fixing the notation. The complete metrics given in equations (5.17) and (5.34), are denoted by $g^s_{ab} = \delta^b_a + (\partial f_{1,2}/\partial Q^a)(\partial f_{1,2}/\partial Q^b)$. The function $f_{1,2}$ is to be understood as the specific choice of augmentation function which defines each metric, and $s \in \{\text{complete}, \text{invariant}\}$.

Inverse of the matrix representation: One of the main advantages of using the construction of complete metrics proposed in Gordon, 1973, Thm. 1, is that the resulting matrix representation of the metric is a rank-one perturbation of a known matrix. In our case, since the metrics g^{complete}_{ab} and $g^{\text{invariant}}_{ab}$ use as the base metric the Euclidean metric, they constitute a rank-one perturbation of the identity matrix. This has two main implications. First, the inverse matrix can be directly

computed using the simplified version of Sherman–Morrison formula (cf., Nocedal, Wright, 2006, Eq. (A.27), p. 612), i. e.,

$$g_s^{ab} = \delta_a^b - \frac{1}{1 + \sum_{a=1}^{N_V} (\partial f_{1,2}/\partial Q^a)^2} \frac{\partial f_{1,2}}{\partial Q^a} \frac{\partial f_{1,2}}{\partial Q^b}, \tag{5.45}$$

for all $a, b = 1, 2, \ldots, N_V$, where $s \in \{\text{complete}, \text{invariant}\}$.
Second, for a given $P \in \mathcal{T}_Q^* \mathcal{M}_+(\Delta; Q_{\text{ref}})$, the solution of the problem:

$$\begin{aligned} &\text{Find} \quad V \in \mathcal{T}_Q \mathcal{M}_+(\Delta; Q_{\text{ref}}) \\ &\text{such that} \quad g_{ab}^s V^a \widetilde{V}^b = P_b \widetilde{V}^b, \quad \text{for all} \ \ b = 1, \ldots, 2N_V, \end{aligned}$$

and all $\widetilde{V} \in \mathcal{T}_Q \mathcal{M}_+(\Delta; Q_{\text{ref}})$, with $s \in \{\text{complete}, \text{invariant}\}$, can be solved with two iterations of the conjugate gradient.
This result can be obtained since the convergence of the conjugate gradient method for a problem $Ax = b$, can be estimated as:

$$\|e^{(k)}\|_A \leq \min_{p_k \in \Pi_k, p_k(0)=1} \max_j |p_k(\lambda_j)| \|e^{(0)}\|_A,$$

where $\|x\|_A = x \cdot Ax$, $e^{(k)} = x - x^{(k)}$ and $x^{(k)}$ is the k-th iteration of the conjugate gradient and Π_k is the set of real polynomials of degree k, with $p(0) = 1$ (cf., Elman, Silvester, Wathen, 2014, Eq. (2.11), p. 76). In our case, since the matrix A is a rank-one perturbation of the identity, one can construct a second degree polynomial p_2 such that $p_2(1) = 0$ and $p_2\left(1 + \sum_{a=1}^{N_V} (\partial f_{1,2}/\partial Q^a)^2\right) = 0$, and obtain convergence in two iterations.

Derivatives of the matrix representation: Recall that the only terms that depend on the node positions Q from the definition of the metric are the functions f_1, f_2, which means we need to compute their second derivatives. Explicitly, the expression:

$$\left\{ \frac{\partial g_{bd}(Q_n)}{\partial Q^a} \, g^{bc}(Q_n) \, P_{c,n+1/2} \, g^{de}(Q_n) \, P_{e,n+1/2} \right\},$$

from lines 3 and 5 can be rewritten in terms of the second derivatives of f_1, f_2 as follows:

$$-\frac{1}{2} \frac{\partial^2 f_{1,2}}{\partial Q^a \partial Q^b} \left(g^{bc} P_c\right) \left(g^{bc} P_c \frac{\partial f_{1,2}}{\partial Q^b}\right), \tag{5.46}$$

where we recall we use Einstein's summation convention.

Assembly of functions given in (5.41), (5.42), and (5.43): To reduce the computational costs of the algorithm, we never assemble the matrices for the metric neither for the inverse metric. We only consider matrix-vector multiplications.

Solution of nonlinear equations: To approximate the solution of the nonlinear systems, given in (5.44), we use a fixed point iteration. We use a tolerance of 10^{-8} and the maximum number of iterations is fixed to 10. The initial value used is $P_{a,n}$ for line 3 and Q_n^a for line 4. Recalling the theory of convergence of the fixed point iteration, we know that if the initial value is chosen properly and the derivative of the function which defines the fixed point iteration is a contraction (its derivative is bounded above by 1); then, the algorithm converges. In our case, since the functions which define the fixed point iteration depend on Δt, the convergence of the fixed point iteration is closely linked to the amount of time steps. In other words, for a given time window, values of the parameters α_j or β_j, respectively, and a given initial tangent vector, we need to choose a large enough number of time steps, such that we achieve convergence of the fixed point scheme at each iteration of the

Störmer–Verlet algorithm. Unfortunately, this restriction significantly increases the computational cost of the approximation of geodesics for $\mathcal{M}_+(\Delta; Q_{\text{ref}})$.

5.5 Numerical Approximation of Geodesics

The purpose of this section is to numerically investigate how planar triangular meshes deform under the complete Riemannian metric (5.17), and to compare it with the Euclidean metric. We will reserve the complete metric introduced in (5.34) for the numerical solution of discrete shape optimization problems. We implemented the Störmer–Verlet scheme (algorithm 1) to integrate the geodesic equation numerically.

The experiments on this section are taken from Herzog, Loayza-Romero, 2020, Sec. 6, and they are structured as follows. In subsection 5.5.1, we investigate elementary transformations (translation, shearing, scaling, and rotation) of a square mesh with crossed diagonals. In subsection 5.5.2 we revisit the example depicted in figure 5.2, whose initial tangent vector gives rise to a more complex transformation dynamic. The aforementioned experiments confirm that the proposed metric successfully avoids self-intersections of the mesh due to its completeness. To study this also quantitatively, we conduct in subsection 5.5.3 an experiment using a slightly more complex mesh, where we evaluate a mesh quality measure along the geodesics.

To be precise, most of the experiments in this section are based on the function

$$f_1(Q; Q_{\text{ref}}) \coloneqq \sum_{k=1}^{N_T} \sum_{\ell=0}^{2} \frac{\alpha_1}{h_Q^\ell(i_0^k, i_1^k, i_2^k)} + \frac{\alpha_3}{2} \|Q - Q_{\text{ref}}\|_F^2, \tag{5.47}$$

which is used to construct the metric:

$$g_{ab}^{\text{complete}} = \delta_a^b + \frac{\partial f_1}{\partial (\operatorname{vec} Q)^a} \frac{\partial f_1}{\partial (\operatorname{vec} Q)^b}, \quad a, b = 1, \dots, 2\,N_V, \tag{5.48}$$

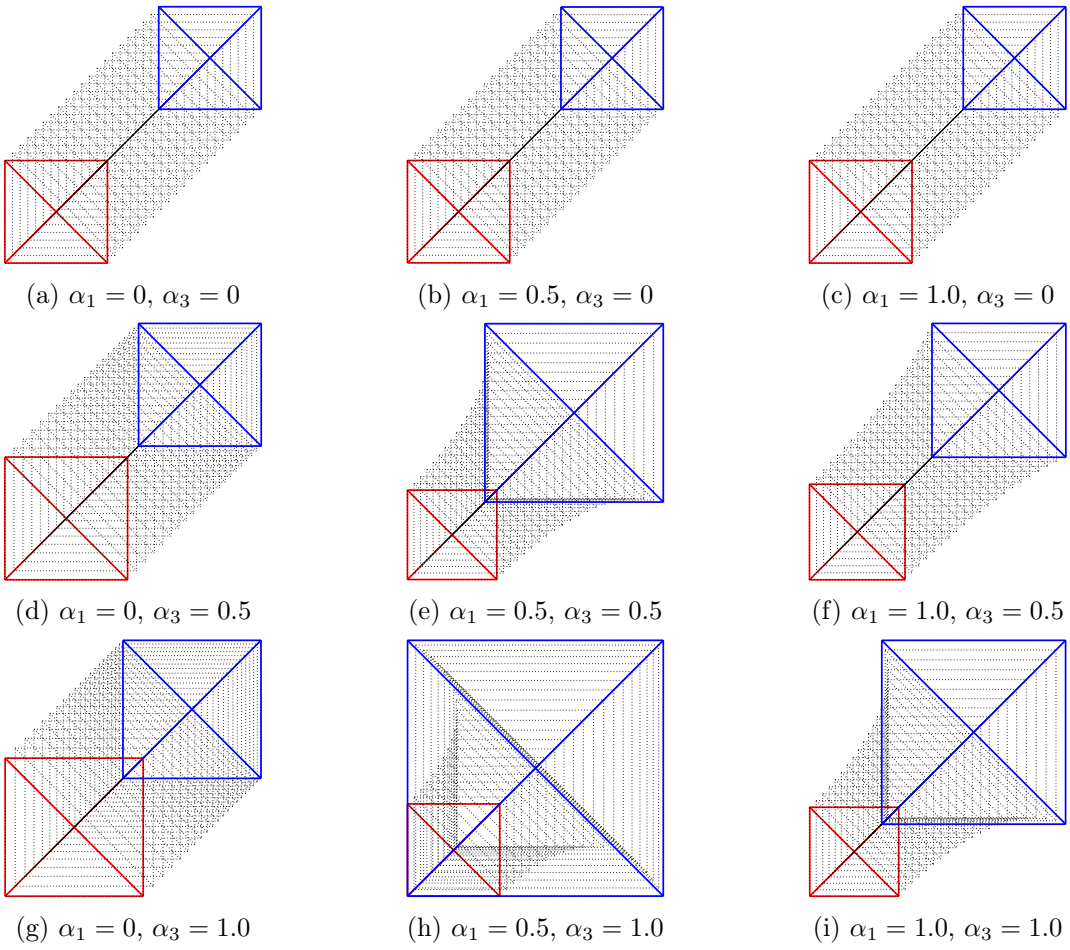
as in (5.17). The omission of the α_2-term in (5.47) preventing exterior self-intersections is justified by remark 5.2.10. For the experiments shown in this section, exterior self-intersections are never a factor, which we verified a posteriori, and thus we can choose a cut-off function χ_2 which effectively removes the second term in (5.16). The choice of parameters $\alpha_1, \alpha_3 > 0$, which control the relative importance of each term in f_1 in relation to the Euclidean base metric in (5.48), is described below for each experiment individually.

5.5.1 Elementary Mesh Transformations

In this section we showcase the deformation of a simple mesh under a number of elementary transformations. To be precise, we consider initial tangent vectors, which would produce a translation, shearing, scaling, or rotation, respectively, of the mesh in the Euclidean setting ($\alpha_1 = \alpha_3 = 0$ in (5.47)). In particular, we numerically study the influence of the parameters α_1 and α_3.

Each of the figures 5.21 to 5.24 shows 20 snapshots of a geodesic on the interval $[0, 3]$, obtained using values $\alpha_1, \alpha_3 \in \{0, 0.5, 1.0\}$. The initial tangent vector is the same for all plots in a figure. Although the initial tangent vectors are not shown, they easily can be recognized by the displacements they induce in the first time step. Notice that the scaling of the plots within a figure may vary to make better use of the available space. In each case, Q_{ref} is chosen to be the initial mesh.

The initial vector in figure 5.21 represents a translation of the mesh. As predicted by proposition 5.2.12, when $\alpha_3 = 0$ holds, geodesics w.r.t. (5.48) coincide with Euclidan geodesics, i. e., they remain translations, as can be seen from the first row in Figure 5.21. In case $\alpha_1 = 0$, the mesh is also merely translated, albeit with a speed (in the Euclidean sense) depending on α_3; see the first column of figure 5.21.



(a) $\alpha_1 = 0$, $\alpha_3 = 0$ (b) $\alpha_1 = 0.5$, $\alpha_3 = 0$ (c) $\alpha_1 = 1.0$, $\alpha_3 = 0$

(d) $\alpha_1 = 0$, $\alpha_3 = 0.5$ (e) $\alpha_1 = 0.5$, $\alpha_3 = 0.5$ (f) $\alpha_1 = 1.0$, $\alpha_3 = 0.5$

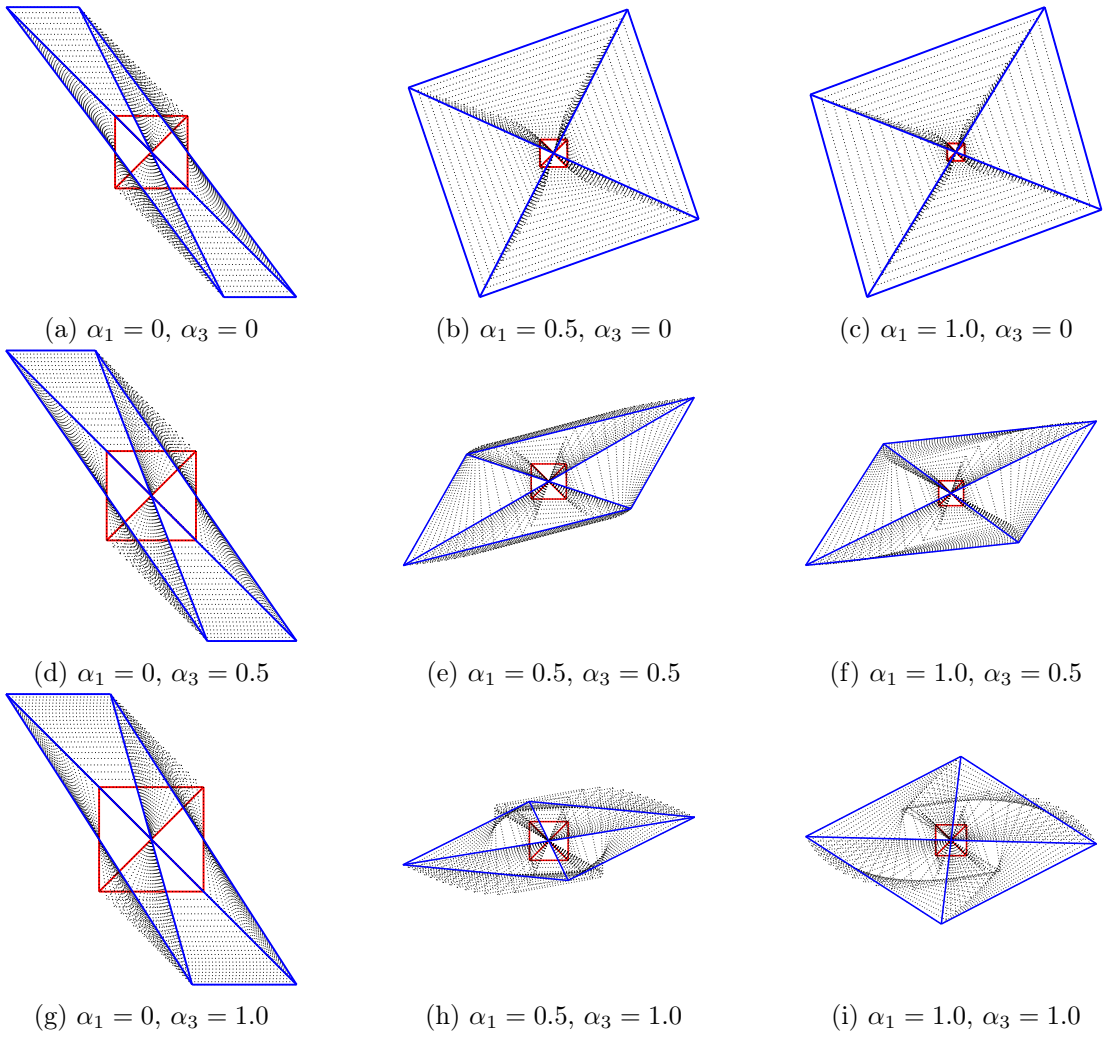
(g) $\alpha_1 = 0$, $\alpha_3 = 1.0$ (h) $\alpha_1 = 0.5$, $\alpha_3 = 1.0$ (i) $\alpha_1 = 1.0$, $\alpha_3 = 1.0$

Figure 5.21. 20 snapshots of geodesics for different values of α_1, α_3, starting from the same initial mesh (shown in red) and produced by the same initial tangent vector, which induces a translation; see subsection 5.5.1. The final mesh is shown in blue.

In figure 5.22 we consider an initial tangent vector which induces a shearing motion. In this case, there is a pronounced difference between the mesh evolution along the Euclidean geodesic ($\alpha_1 = \alpha_3 = 0$, see figure 5.22a) and those with $\alpha_1 > 0$. In the latter case, we can clearly see how the term involving the heights in (5.47) counteracts the impending mesh degeneracy observed along the Euclidean geodesic and helps to maintain a favorable cell aspect ratio.

Figure 5.23 shows the mesh deformation when the initial tangent vector induces a scaling of the mesh, i. e., the tangent vectors at the four corner vertices are pointing inwards. In the Euclidean case ($\alpha_1 = \alpha_3 = 0$), this quickly leads to a nonadmissible (flipped) mesh shown in figure 5.23a (top left). The same is true for the other experiments with $\alpha_1 = 0$ (first column). However, with positive values of α_1 and α_3 we see that the completeness of the metric prevents the mesh from shrinking too much and it remains admissible for all times.

In figure 5.24 we show some geodesics when the initial tangent vector induces a rotation. Here the Euclidean geodesic ($\alpha_1 = \alpha_3 = 0$) does not cause the mesh to become degenerate. Naturally, the Euclidean geodesic resembles a rotation only for small times since all vertices



(a) $\alpha_1 = 0$, $\alpha_3 = 0$ (b) $\alpha_1 = 0.5$, $\alpha_3 = 0$ (c) $\alpha_1 = 1.0$, $\alpha_3 = 0$

(d) $\alpha_1 = 0$, $\alpha_3 = 0.5$ (e) $\alpha_1 = 0.5$, $\alpha_3 = 0.5$ (f) $\alpha_1 = 1.0$, $\alpha_3 = 0.5$

(g) $\alpha_1 = 0$, $\alpha_3 = 1.0$ (h) $\alpha_1 = 0.5$, $\alpha_3 = 1.0$ (i) $\alpha_1 = 1.0$, $\alpha_3 = 1.0$

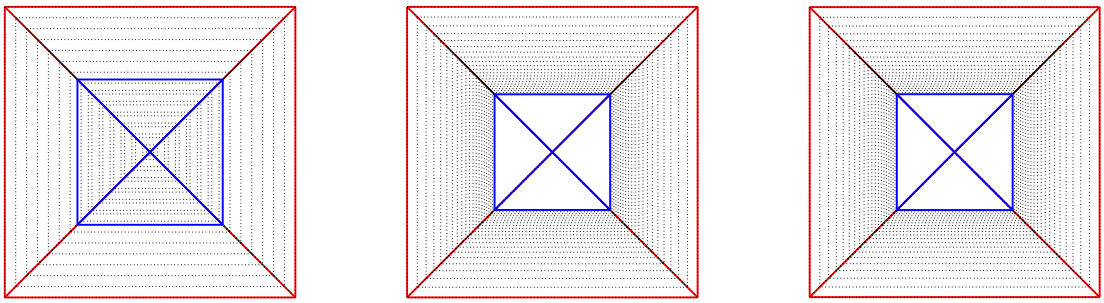
Figure 5.22. 20 snapshots of geodesics for different values of α_1, α_3, starting from the same initial mesh (shown in red) and produced by the same initial tangent vector, which induces shearing; see subsection 5.5.1. The final mesh is shown in blue.

move along straight lines. Interestingly, in the case $\alpha_1 > 0$ and $\alpha_3 = 0$ the geodesics appear to produce exact rotations.

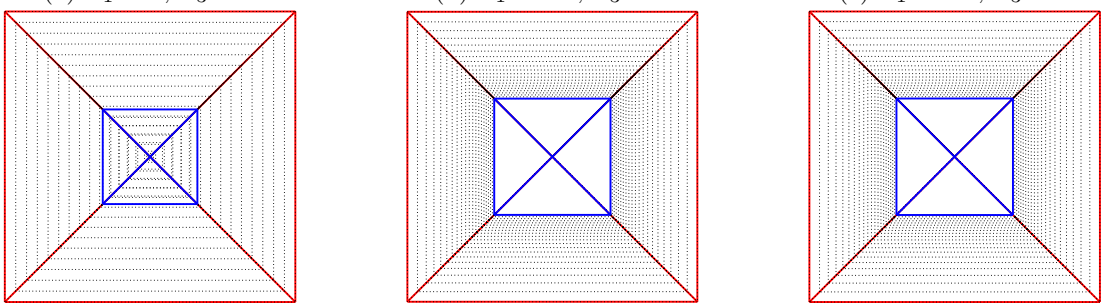
5.5.2 More Complex Initial Tangent Vector

In this example, we revisit the setup depicted in figure 5.2, i. e., we retain a very simple initial mesh but consider the geodesic in the direction of an unfavorable initial tangent vector (figure 5.1). We compare the Euclidean geodesic ($\alpha_1 = \alpha_3 = 0$) with the proposed metric (for values $\alpha_1 = \alpha_3 = 1$). In each case, Q_{ref} is chosen to be the initial mesh. Figure 5.25 shows the respective mesh evolution at 16 snapshots within the interval $[0, 2]$. The scaling of the axes is the same in each snapshot.

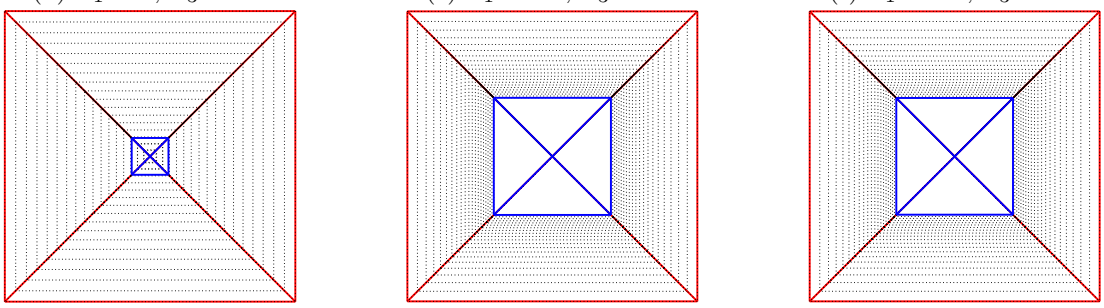
In the Euclidean case, the mesh degenerates very quickly and becomes nonadmissible around $t = 1$. By contrast, the completeness of the proposed metric ensures the mesh to be admissible for all times. Notice that the inward pointing initial tangent vectors at the two



(a) $\alpha_1 = 0$, $\alpha_3 = 0$ (b) $\alpha_1 = 0.5$, $\alpha_3 = 0$ (c) $\alpha_1 = 1.0$, $\alpha_3 = 0$



(d) $\alpha_1 = 0$, $\alpha_3 = 0.5$ (e) $\alpha_1 = 0.5$, $\alpha_3 = 0.5$ (f) $\alpha_1 = 1.0$, $\alpha_3 = 0.5$



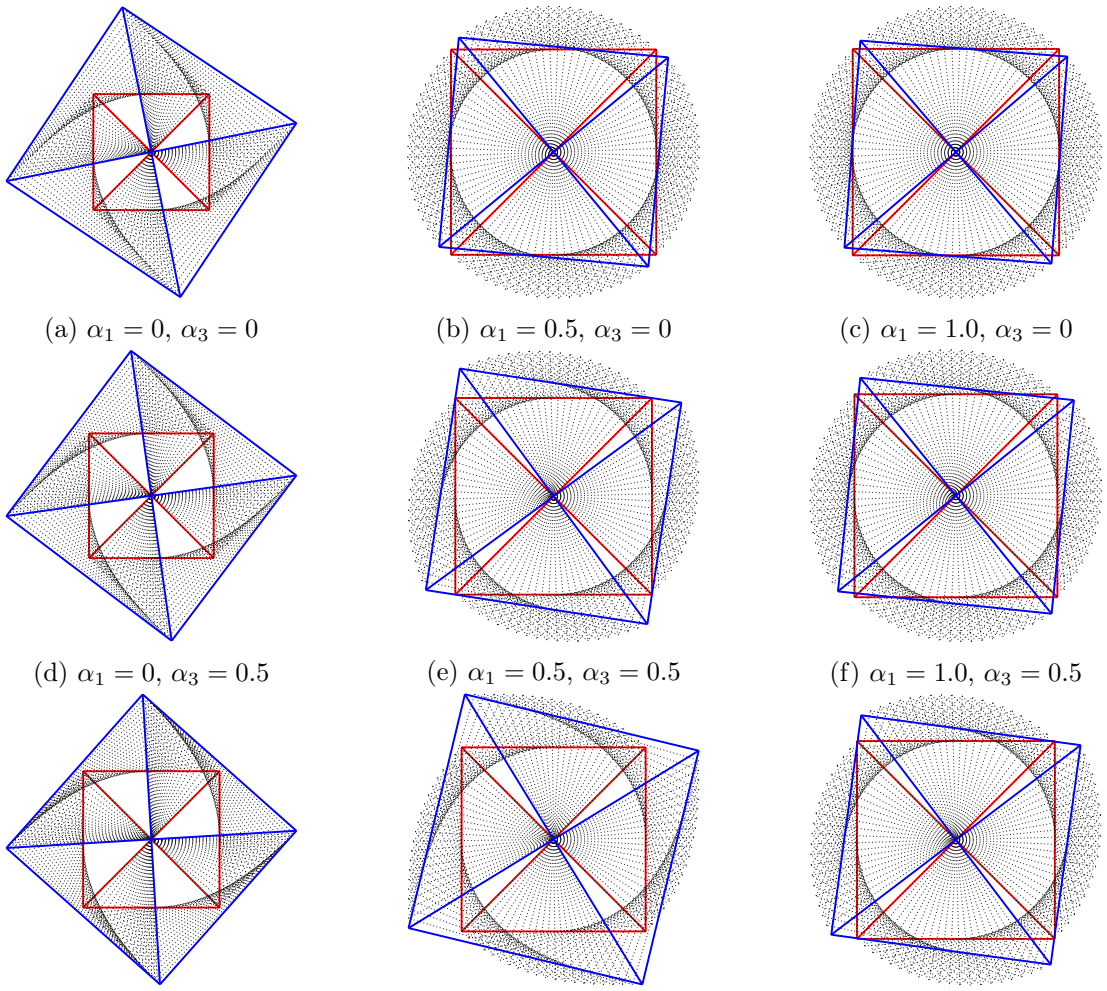
(g) $\alpha_1 = 0$, $\alpha_3 = 1.0$ (h) $\alpha_1 = 0.5$, $\alpha_3 = 1.0$ (i) $\alpha_1 = 1.0$, $\alpha_3 = 1.0$

Figure 5.23. 20 snapshots of geodesics for different values of α_1, α_3, starting from the same initial mesh (shown in red) and produced by the same initial tangent vector, which induces scaling; see subsection 5.5.1. The final mesh is shown in blue.

bottom vertices are repelled and reverse direction around $t = 0.53$ as a consequence of the term involving α_1 in (5.47), which avoids interior self-intersections.

5.5.3 Mesh Quality Experiment

In this experiment, we consider a slightly more complex mesh consisting of 25 vertices and 32 triangles. The initial mesh is a discretized version of the unit circle. The initial tangent vector considered acts only on the boundary of the mesh, i. e., its components pertaining to interior mesh nodes are zero. Such a situation occurs frequently in shape optimization, where the Hadamard structure theorem (Sokołowski, Zolésio, 1992, Thm. 2.27) provides an expression for the shape derivative which is supported only the boundary of the current mesh. One would then usually apply an extension technique to obtain a displacement field (tangent vector) supported in all nodes. A typical example is to achieve this through the solution of an elasticity equation. With our approach, the displacement of all vertices,



(a) $\alpha_1 = 0$, $\alpha_3 = 0$ (b) $\alpha_1 = 0.5$, $\alpha_3 = 0$ (c) $\alpha_1 = 1.0$, $\alpha_3 = 0$

(d) $\alpha_1 = 0$, $\alpha_3 = 0.5$ (e) $\alpha_1 = 0.5$, $\alpha_3 = 0.5$ (f) $\alpha_1 = 1.0$, $\alpha_3 = 0.5$

(g) $\alpha_1 = 0$, $\alpha_3 = 1.0$ (h) $\alpha_1 = 0.5$, $\alpha_3 = 1.0$ (i) $\alpha_1 = 1.0$, $\alpha_3 = 1.0$

Figure 5.24. 20 snapshots of geodesics for different values of α_1, α_3, starting from the same initial mesh (shown in red) and produced by the same initial tangent vector, which induces a rotation; see subsection 5.5.1. The final mesh is shown in blue.

including the interior ones, is achieved automatically by following the geodesic with respect to the metric (5.48).

As a quality measure for the mesh, we consider the mesh aspect ratio

$$AR(\Delta; Q) := \min_{[i_0,i_1,i_2]\in\Delta} \frac{2\, r_Q(i_0, i_1, i_2)}{R_Q(i_0, i_1, i_2)}. \tag{5.49}$$

The aspect ratio takes values between 0 and 1, where the latter is achieved precisely for unilateral triangles.

From figure 5.26, we can observe the mesh aspect ratio degenerates over time for the Euclidean metric, where only the boundary values are displaced. By contrast, the mesh aspect ratio is observed to be bounded away from zero for the metric (5.48) when $\alpha_1, \alpha_3 > 0$.

In figure 5.27 we show three snapshots of three geodesics on the interval $[0, 2]$. We compare the Euclidean geodesic ($\alpha_1 = \alpha_3 = 0$) with the proposed metric for values $\alpha_1 =$

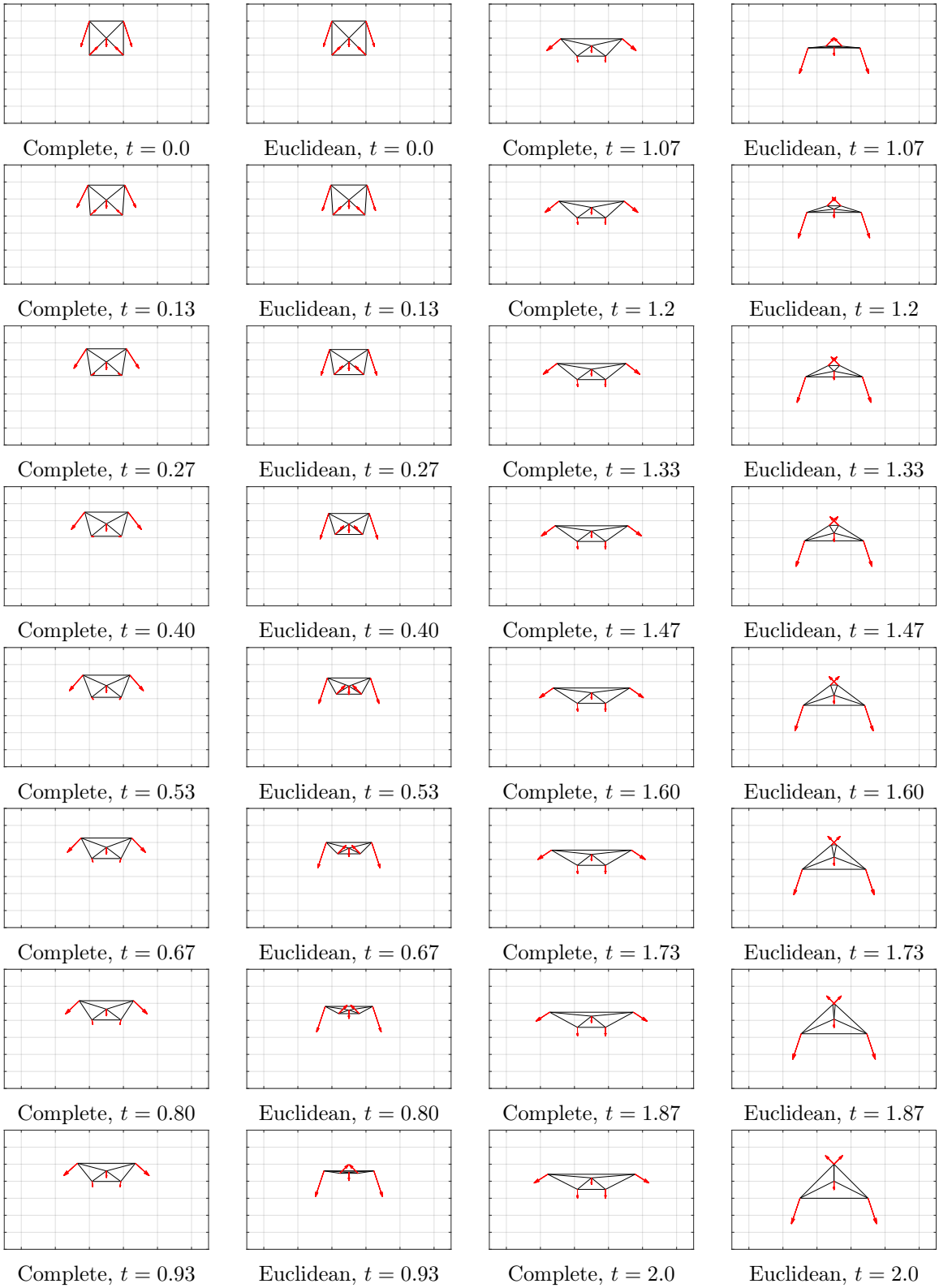


Figure 5.25. Snapshots of the geodesics described in subsection 5.5.2, comparing the complete metric with $\alpha_1 = \alpha_3 = 1$ (first and third columns) and the Euclidean metric with $\alpha_1 = \alpha_3 = 0$ (second and fourth columns).

$\alpha_3 = 0.25$ and $\alpha_1 = \alpha_3 = 0.5$. As in the previous experiments, Q_{ref} is chosen to be the initial mesh.

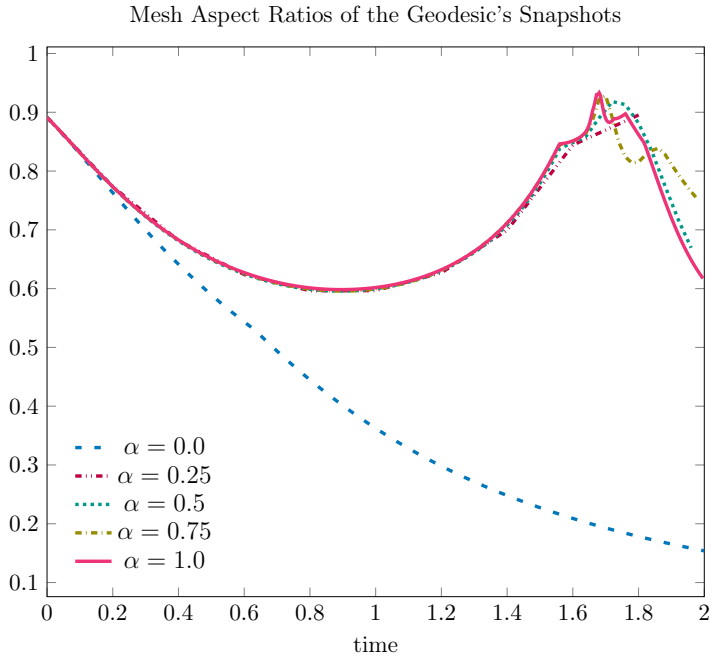


Figure 5.26. Mesh quality plot over time along geodesics for different values of $\alpha = \alpha_1 = \alpha_3$, see subsection 5.5.3.

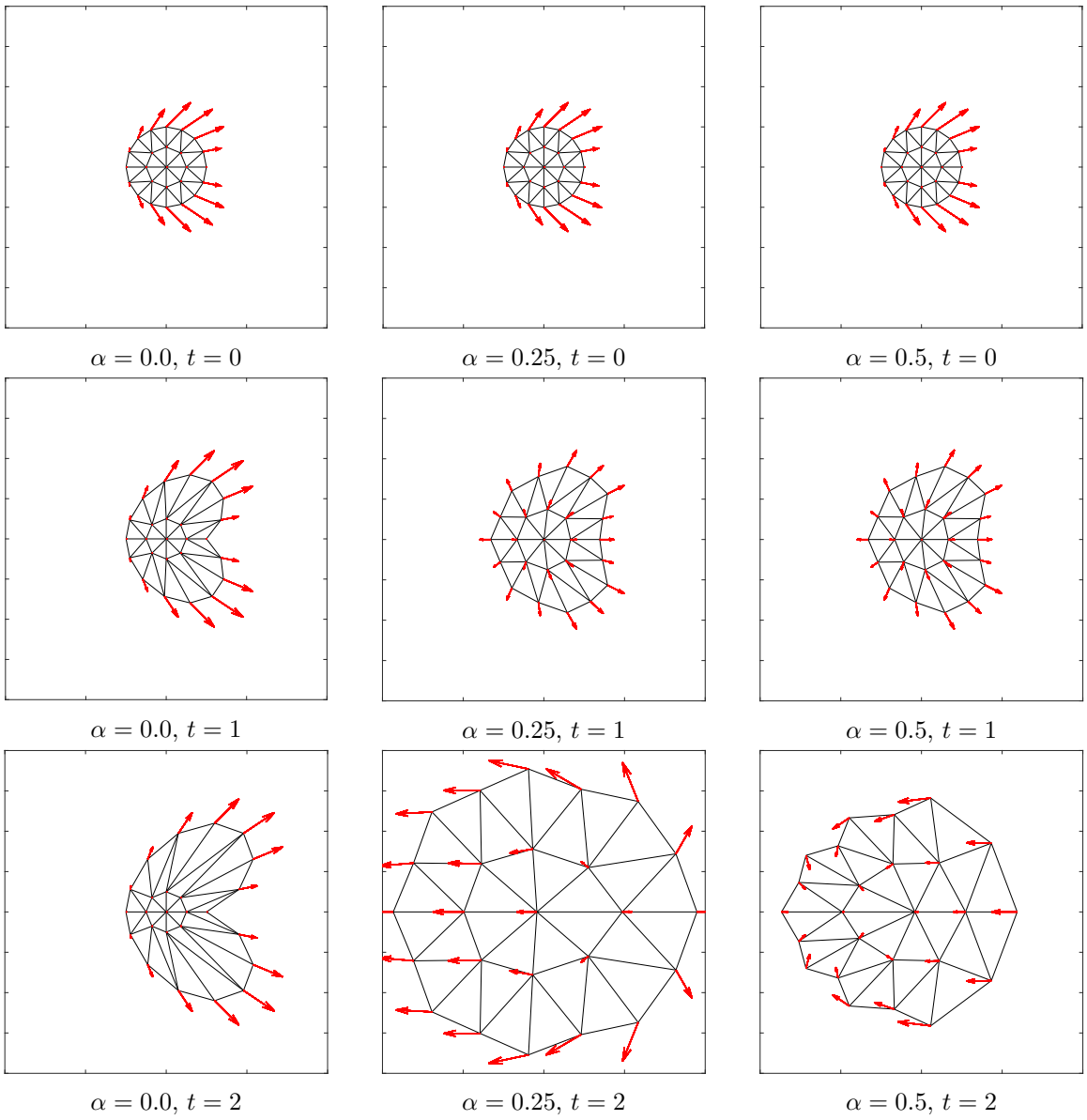


Figure 5.27. Snapshots of the geodesics described in subsection 5.5.3 for different values of $\alpha = \alpha_1 = \alpha_3$. The tangent vectors are also shown. They have been scaled to improve the visualization.

6 Discretize-then-Optimize Approach for PDE-Constrained Shape Optimization

Contents

As already discussed in section 3.4, one major drawback of computational PDE-constrained shape optimization problems, when we discretize the state equation with a finite element method and choose the node positions of the underlying meshes as our optimization variables, is the degeneracy of the mesh quality as the optimization progresses. Usually this degeneracy manifests as some of the mesh cells thinning in the sense that at least one of its heights approaches to zero. In chapter 4, particularly section 4.3, we have described the manifold of planar triangular meshes $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ which represents the set of admissible node positions where we will pose our discretized problem. Moreover, in chapter 5, particularly section 5.3, we have endowed the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ with a complete metric, which basically penalizes the degeneracy which the meshes are usually subjected to. The main aim of this chapter is to apply this newly acquired knowledge to the computational solution of discretized problems.

Let us start by mentioning previous works about the numerical solution of discretized, PDE-constrained shape optimization problems. We found that the literature in this direction is scarce. We start recalling the works from Souli, Zolésio, 1993 and Pironneau, 1984, Sec. 7.2.4, p. 106. In both cases, the authors propose to use the information provided by the shape derivative to update only the boundary nodes. The interior nodes are usually updated via a different function that smoothly distributes the motion of the boundary nodes towards the interior ones. In both works, the Euclidean metric is chosen to transform the shape derivative into a gradient. They also set a limit to the norm of the descent directions to avoid degeneracy of the meshes. Moreover, in the latter publication, the author also mentions there is no reason to believe that this algorithm will keep a good quality of the meshes along the optimization process. To avoid this, the author proposes to set a second limit value of the step length and fix the entries of the descent direction to zero when the movements of the nodes tend to degenerate the cells.

We also mention the work from Delfour, Payre, Zolesio, 1985. Their study is based on optimal triangulations; however, one can migrate these ideas to the context of shape optimization. We particularly refer to section 6.2.1, on page 255, where they present some iterations of the steepest descent method. After iteration 12, three triangles of the mesh had collapsed. As in the previous works, they also propose setting a maximum step length value to avoid mesh degeneracy. Additionally, they aim to find an improved step length such that the oriented surface (signed area in our notation) of the new triangle has the same sign of the

oriented surface of the triangle from the previous iteration. In other words, similarly to our proposal of the manifold of planar triangular meshes $\mathcal{M}_+(\Delta; Q_{\text{ref}})$, the authors also suggest that allowing movements of the nodes which preserve the sign of the oriented surface could be a way to avoid mesh degeneracy. These considerations yield a quadratic inequality, where the variable is the step length. Despite the simplicity of the computation, it has a major disadvantage: errors in the computation of the coefficients from the quadratic inequality can lead to solutions which are useless in practice. Analogously, they propose to zero the entries of the gradient which lead to ill-behaved triangles.

It is worth highlighting that the previously mentioned approaches use the Euclidean scalar product for the transformation of the derivative into the gradient. Furthermore, the node positions of the mesh are treated as elements of a vector space, which they clearly are not. We conjecture that these are the main reasons why the discrete approach for shape optimization problems was not investigated further, and an optimize-then-discretize paradigm became the most commonly used approach.

Throughout, we consider a two-dimensional model problem as in Etling et al., 2020. In continuous form it reads:

$$\text{Minimize} \quad \int_\Omega y \,\mathrm{d}x \quad \text{s.t.} \quad -\Delta y = r \text{ in } \Omega \quad \text{w.r.t.} \quad \Omega \subset \mathbb{R}^2. \tag{6.1}$$

The state y is subject to Dirichlet boundary conditions $y = 0$ on $\partial\Omega$ and the right-hand side function $r\colon \mathbb{R}^2 \to \mathbb{R}$ is given. To discretize it, we represent the unknown domain Ω by a mesh with coordinates $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}}) \subset \mathbb{R}^{2\times N_V}$ and given oriented connectivity complex, as introduced in definitions 4.2.5 and 4.3.10.

We refer to the domain covered by the mesh with node coordinates Q as Ω_Q. We discretize the PDE in (6.1) by the finite element method. To this end, let $S^1(\Omega_Q)$ denote the finite element space of piecewise linear, globally continuous functions, defined over Ω_Q, and let $S^1_0(\Omega_Q)$ denote the subspace of functions with zero Dirichlet boundary conditions, as described in section 2.3. The discrete version of (6.1) then becomes

$$\begin{aligned} \text{Minimize} \quad & \int_{\Omega_Q} y \,\mathrm{d}x \quad \text{w.r.t.} \quad Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}}),\ y \in S^1_0(\Omega_Q) \\ \text{s.t.} \quad & \int_{\Omega_Q} \nabla y \cdot \nabla v \,\mathrm{d}x = \int_{\Omega_Q} r\, v \,\mathrm{d}x \quad \text{for all } v \in S^1_0(\Omega_Q). \end{aligned} \tag{6.2}$$

The main goal of this chapter is to analyze and numerically solve problem (6.2). The existence of solutions of problem (6.1) is obtained when the admissible shapes are assumed to belong the set of quasi-convex sets (cf., Etling et al., 2020, Thm. 2.1). Therefore, the first question we will answer is: does problem (6.2) have a solution?. To this end, in section 6.1 we present our first findings about the possible non-existence of solutions for the problem (6.2), by providing a numerical example. To overcome the possible lack of solutions, we propose to consider a penalized shape optimization problem, whose analysis is presented in section 6.2. The penalization is based on the $\mathcal{C}^3$, proper function f_2 presented in (5.22) (recall we have dropped out the superindex μ in the notation of the $\mathcal{C}^3$-regularizations). Thanks to the properties of f_2 and under usual assumptions about the shape functional, one can obtain the existence of at least one globally optimal solution of the penalized problem, whose result is presented in proposition 6.2.3. Then, we focus on the first-order optimality conditions. We present a detailed computation of the shape derivatives of problem (6.2). We choose the Riemannian steepest descent method with Armijo backtracking line search for the approximation of the solution of problem (6.2), which we describe in section 6.3. We consider four different variants of the steepest descent method depending on the Riemannian metric used to transform the shape derivative into a gradient and the one chosen to navigate along

the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. Finally, in section 6.4, we present three numerical experiments whose aim is to showcase the performance of the steepest descent method when the complete metric $g_{ab}^{\text{invariant}}$ given in (5.34) is chosen for the transformation of derivatives into gradients and/or the navigation of the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. This chapter is based on Herzog, Loayza-Romero, 2021.

6.1 A First Glimpse at the Non-Existence of Solutions

There is a major difference between the continuous and discrete shape optimization problems (6.1) and (6.2). In the former, smooth and bijective reparametrizations of the domain Ω, which preserve the boundary, do not change the solution of the state equation, nor the value of the shape functional. By contrast, the finite element solution of the state equation in the discretized case depends on the positions of *all* nodes, boundary and interior. Moreover, degenerate meshes usually lead to unrealistically small shape functional values, whose infimal value is not attained within $\mathcal{M}_+(\Delta; Q_{\text{ref}})$.

Let us illustrate what happens in the discrete setting, for the simplest possible case. Consider the reference mesh Q_{ref} covering $[-1,1]^2$ shown in figure 6.1a. The nodal positions are recorded in $Q = [q_1, q_2, q_3, q_4, q_5] \in \mathbb{R}^{2\times 5}$. For this experiment, we can even keep the boundary of the shape fixed so that the only remaining unknown is the position of the interior vertex, q_5. It is obvious that $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ holds if and only if $q_5 \in (-1,1)^2$. This leads us to consider the following discrete problem as a particular case of (6.2),

$$\begin{aligned} \text{Minimize} \quad & \int_{\Omega_Q} y \,\mathrm{d}x \quad \text{w.r.t.} \quad q_5 \in (-1,1)^2,\ y \in S_0^1(\Omega_Q) \\ \text{s.t.} \quad & \int_{\Omega_Q} \nabla y \cdot \nabla v \,\mathrm{d}x = \int_{\Omega_Q} r\, v \,\mathrm{d}x \quad \text{for all } v \in S_0^1(\Omega_Q). \end{aligned} \tag{6.3}$$

For this initial experiment, we fix $r \equiv 1$. We emphasize that in this scenario, no quadrature error occurs even for the simplest quadrature formula with one evaluation at each cell center.

Figure 6.1b shows the value of the discrete shape functional as a function of q_5. It can be observed that the shape functional takes values arbitrarily close to zero when q_5 approaches the boundary of Ω_Q. To confirm this, consider for instance $q_5 = (0, 1-\varepsilon)^{\mathrm{T}}$ with a small $\varepsilon > 0$. It can be easily verified that in this case the linear system representing the PDE in (6.2) reads $Ky = b$ with stiffness matrix K and load vector b given by the following expressions:

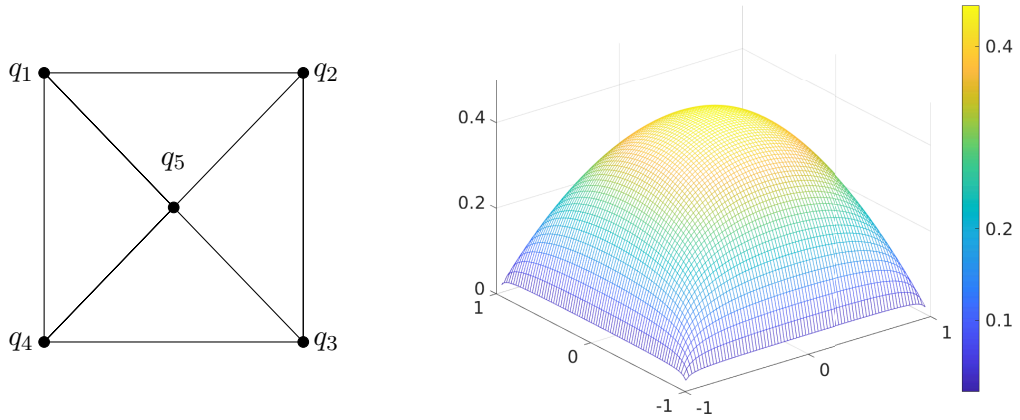
$$K = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 4+1/\varepsilon \end{bmatrix}, \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 4/3 \end{bmatrix}.$$

Consequently, the nodal solution vector y is given by

$$y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \frac{4\varepsilon}{3(4\varepsilon+1)} \end{bmatrix},$$

and satisfies $y \searrow 0 \in \mathbb{R}^5$ as $\varepsilon \searrow 0$. Thus the value of the shape functional approaches zero as well. Similar considerations apply when q_5 is anywhere else near the boundary. Since a location of q_5 exactly on the boundary results in a degenerate mesh with $Q \notin \mathcal{M}_+(\Delta; Q_{\text{ref}})$, we conclude that the simple problem (6.3) does not have a solution in $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. This is

in contrast to the continuous problem. In the continuous setting, due to the fixed boundary, there is no shape to be optimized. The solution to the state equation on $\Omega = (-1, 1)^2$ can be found, e. g., in Elman, Silvester, Wathen, 2014, Ex. 1.1.1, p. 10 and the corresponding value of the shape functional is approximately 0.5622.



(a) Illustration of the reference mesh Q_{ref}.

(b) Shape functional as a function of the nodal position q_5.

Figure 6.1. Reference mesh and shape functional for problem (6.3).

Later we will consider in subsection 6.4.2 more realistic meshes, a different right-hand side function r and, of course, impose no constraints which fix the boundary. However, even this preliminary experiment (6.3) illustrates two fundamental difficulties with discretized shape optimization problems, in which the nodal positions serve as the optimization variables. First, they do not, in general, possess a solution, even if the shape functional is bounded below. Second, poor approximations of the state variable can give rise to unreasonably small shape functional values. Both observations are related to nearly degenerate finite element meshes. Moreover, we want to highlight that this behavior has also been recognized in previous works, for example in Hardesty, Antil, et al., 2020, where they claim «once the boundary shape is resolved it (the optimizer) changes the discretization error by moving the interior nodes in a way that further reduces the objective function, at the expense of mesh quality». Moreover, in Berggren, 2010 the author writes «However, in shape optimization, it does not make much sense to optimize the position of each mesh points independently». We conjecture these claims are only a different way to express the same statement: "In general discrete shape optimization problems have no solution". Despite of these observations, we are not aware of a detailed investigation.

To summarize, it is of paramount importance that formulations and solvers for discretized shape optimization problems maintain control over the mesh quality. Precisely that is the purpose of the penalty function devised in the following section.

6.2 Penalized Discrete Shape Optimization

This section proposes a modification of discrete shape optimization problems over the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ of planar triangular meshes. The modification consists in the addition of a penalty function f, which renders the resulting problem well-posed in the sense that the existence of a globally optimal solution can be proved. To the best of the author's knowledge,

we are not aware of existence results for discretized shape optimization problems (in which the node positions serve as optimization variables) in the literature.

The penalized discrete problem which we consider in this section, with a penalty term f to be specified below added, reads

$$\begin{aligned} \text{Minimize} \quad & \int_{\Omega_Q} y \, \mathrm{d}x + f(Q; Q_{\mathrm{ref}}) \quad \text{w.r.t.} \quad Q \in \mathcal{M}_+(\Delta; Q_{\mathrm{ref}}),\ y \in S_0^1(\Omega_Q) \\ \text{s.t.} \quad & \int_{\Omega_Q} \nabla y \cdot \nabla v \, \mathrm{d}x = \int_{\Omega_Q} r\, v \, \mathrm{d}x \quad \text{for all } v \in S_0^1(\Omega_Q). \end{aligned} \tag{6.4}$$

We recall that $S_0^1(\Omega_Q)$ denotes the finite element space of piecewise linear, globally continuous functions, defined over Ω_Q, with zero Dirichlet boundary conditions.

To motivate our choice of penalization, we present a result which guarantees the existence of solutions to an abstract optimization problem in metric spaces.

Proposition 6.2.1. *Suppose that X is a metric space and $f\colon X \to \mathbb{R}$ a proper function. Moreover, assume that f is bounded from below and lower semi continuous. Then the problem*

$$\textit{Minimize} \quad f(x) \quad \textit{w.r.t.} \quad x \in X \tag{6.5}$$

has at least one globally optimal solution.

Proof. Let us denote by f_0 a lower bound for f. We consider a minimizing sequence $\{x^n\} \subset X$, i. e., $f(x^n) \searrow \inf\{f(x) \,|\, x \in X\}$ holds, which implies that the sequence $\{f(x^n)\} \subset \mathbb{R}$ is bounded. Thus, there exists a constant $K < \infty$ such that $f(x^n) \in [f_0, K]$ holds for all $n \in \mathbb{N}$. Since the interval $[f_0, K]$ is compact in $\mathbb{R}$ and thanks to the properness of f, we know that the set $f^{-1}([f_0, K])$ is compact in X. Since X is a metric space, compactness is equivalent to sequential compactness, which in turn implies that we can extract a convergent subsequence from $\{x^n\} \subset f^{-1}([f_0, K])$, still denoted by $\{x^n\}$. Thanks to the lower semi continuity of f and the uniqueness of the limit for $\{f(x^n)\}$, we obtain the result. □

Indeed, proposition 6.2.1 is a particular case of a classical result in which one assumes f to have at least one nonempty and compact sublevel set. We formulate a simple corollary tailored to problems of the form (6.4):

Corollary 6.2.2. *Let X and f be as in proposition 6.2.1. Moreover, suppose that $j\colon X \to \mathbb{R} \cup \{\infty\}$ is also bounded from below, lower semi continuous and not identically equal to ∞. Then the problem*

$$\textit{Minimize} \quad j(x) + f(x) \quad \textit{w.r.t.} \quad x \in X \tag{6.6}$$

has at least one globally optimal solution.

In what follows, j will play the role of the shape functional such as $\int_{\Omega_Q} y \, \mathrm{d}x$ in (6.4), while f denotes the penalty function. Corollary 6.2.2 suggests to define the latter so that it is proper on $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$. Recall, moreover, that the definition of a complete metric on $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ also relies on a proper function. Therefore, f can and will serve both purposes at the same time. We thus require the penalty function f to be proper, $\mathcal{C}^3$, bounded from below, and invariant with respect to rigid body motions and uniform refinements of the mesh. The function f being proper and bounded from below can be used to show the existence of solutions to optimization problems such as (6.6). The assumption of f being $\mathcal{C}^3$ is required for an augmentation function to define a complete metric as in theorem 5.2.9. The condition of f being invariant under rigid body motions means the following: Suppose that $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ is defined by $T(x) = Rx + b$ with $R \in \mathrm{SO}(2)$ and $b \in \mathbb{R}^2$. Extend R and T to $\mathbb{R}^{2\times N_V}$, operating column by column. Then we ask that $f(Q; Q_{\mathrm{ref}}) = f(TQ; TQ_{\mathrm{ref}})$ holds. Finally, and as already mentioned, the condition of the function f being invariant

under uniform mesh refinements is motivated by applications in PDE-constrained shape optimization. When every edge of the mesh is bisected and thus every triangle split into four congruent ones, the value of the shape functional j will remain nearly the same (up to an improvement in the discretization error), and we wish the same to be true for the penalty function f.

Let us recall that in this chapter we drop out the superscript μ and consider only the $\mathcal{C}^3$ versions of the augmentation functions. Moreover, observe the augmentation function f_1 given in (5.16), which served as the basis of a complete Riemannian metric on $\mathcal{M}_+(\Delta)$ in section 5.2, is already proper, $\mathcal{C}^3$, bounded from below, and invariant under rigid body motions. However, it is not invariant under uniform mesh refinements refinements. On the other hand, the function f_2 given in (5.22), which we recall in what follows:

$$\begin{aligned} f_2(Q;Q_{\mathrm{ref}}) \coloneqq & \sum_{k=1}^{N_T} \frac{1}{N_T} \frac{\beta_1}{\psi_Q(i_0^k, i_1^k, i_2^k)} + \frac{\beta_2}{\sum_{k=1}^{N_T} A_Q\big(i_0^k, i_1^k, i_2^k\big)} \\ & + \sum_{[j_0,j_1]\in E_\partial} \sum_{\substack{i_0\in V_\partial \\ i_0\neq j_0,j_1}} \frac{1}{\#E_\partial \# V_\partial} \frac{\beta_3}{D_Q^\mu(i_0;[j_0,j_1])} + \frac{\beta_4}{2N_V}\|Q-Q_{\mathrm{ref}}\|_F^2 \end{aligned}$$

with

$$\frac{1}{\psi_Q(i_0,i_1,i_2)} \coloneqq \frac{\big(E_Q^0(i_0,i_1,i_2)\big)^2 + \big(E_Q^1(i_0,i_1,i_2)\big)^2 + \big(E_Q^2(i_0,i_1,i_2)\big)^2}{4\sqrt{3}\, A_Q(i_0,i_1,i_2)}$$

satisfies all the required properties. The function E_Q^ℓ denotes the length of the ℓ-th edge, A_Q refers to the area of the triangle and D_Q^μ is a $\mathcal{C}^3$-regularization of the 1-norm based distance of a vertex to an edge.

A proof of the properness was presented in theorem 5.3.3. Since the terms D_Q^μ, A_Q and $\|\cdot\|_F$ are always non-negative, f_2 is bounded below by zero. The term associated with β_2 penalizes small total areas of the entire mesh. Even in the continuous case, the inclusion of such a term into the shape functional makes sense in order to avoid domains shrinking to a point becoming optimal. In remark 5.3.6, we have also commented about the invariance properties of the function f_2.

The properness of f_2 provided by theorem 5.3.3 guarantees the existence of solutions to the penalized discrete shape optimization model problem (6.4). The proof of this result is presented in proposition 6.2.3 under the customary assumption of a hold-all domain. We define the latter by requiring that all nodal positions belong to a certain box, i. e.,

$$D \coloneqq \{Q=[q_1,\dots,q_{N_V}] \in \mathcal{M}_+(\Delta;Q_{\mathrm{ref}}) \,|\, q_i \in [\underline{a},\overline{a}]\times[\underline{b},\overline{b}] \text{ for all } i=1,\dots,N_V\} \tag{6.7}$$

for some constants $\underline{a}<\overline{a}$ and $\underline{b}<\overline{b}$. Notice that this implies that the mesh Ω_Q itself lies inside $[\underline{a},\overline{a}]\times[\underline{b},\overline{b}]$.

Proposition 6.2.3. *Let f_2 be as in (5.22) or (5.33) with $\beta_1,\beta_2,\beta_3,\beta_4>0$. Suppose, moreover, that Q_{ref} belongs to the hold-all D as in (6.7). Denote by $I_D(Q)$ the characteristic function of D. Finally, suppose that r belong to $L^\infty([\underline{a},\overline{a}]\times[\underline{b},\overline{b}])$. Then the problem*

$$\begin{aligned} \textit{Minimize} \quad & \int_{\Omega_Q} y \,\mathrm{d}x + I_D(Q) + f_2(Q;Q_{\mathrm{ref}}) \quad \textit{w.r.t.} \quad Q\in\mathcal{M}_+(\Delta;Q_{\mathrm{ref}}),\ y\in S_0^1(\Omega_Q) \\ \textit{s.t.} \quad & \int_{\Omega_Q} \nabla y\cdot\nabla v \,\mathrm{d}x = \int_{\Omega_Q} r\, v \,\mathrm{d}x \quad \textit{for all } v\in S_0^1(\Omega_Q) \end{aligned} \tag{6.8}$$

has at least one globally optimal solution in $\mathcal{M}_+(\Delta;Q_{\mathrm{ref}})$.

Proof. By virtue of corollary 6.2.2 and theorem 5.3.3 it is enough to show that the function $\int_{\Omega_Q} y \,\mathrm{d}x + I_D(Q)$ is bounded from below, lower semi continuous and not identically equal to ∞. First we note that I_D is lower semi continuous since D is closed in $\mathbb{R}^{2\times N_V}$ and thus closed in $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$. On the other hand, the continuity of $\int_{\Omega_Q} y \,\mathrm{d}x$ follows from the continuity of the mass matrix and the inverse of the stiffness matrix associated with the weak formulation of the partial differential equation, as a function of the vertex coordinates $Q \in \mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$. Notice, moreover, that j is everywhere finite on $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ and I_D is not identically equal to ∞ since $Q_{\mathrm{ref}} \in D$.

Thanks to the definition of the characteristic function, it remains to be proved that $\int_{\Omega_Q} y \,\mathrm{d}x$ is bounded from below on D. Using $L^2(\Omega_Q) \subset L^1(\Omega_Q)$ and Poincaré's inequality, one can obtain the following estimate:

$$\int_{\Omega_Q} y \,\mathrm{d}x \geq -\|y\|_{L^1(\Omega_Q)} \geq -|\Omega_Q|^{1/2} \|y\|_{L^2(\Omega_Q)} \geq -|\Omega_Q|^{1/2} \operatorname{diam}(\Omega_Q) \|\nabla y\|_{L^2(\Omega_Q)},$$

where $|\Omega_Q|$ stands for the volume of Ω_Q and $\operatorname{diam}(\Omega_Q)$ is the diameter of Ω_Q. From the weak formulation of the state equation and under similar arguments as before, it follows that

$$\|\nabla y\|_{L^2(\Omega_Q)} \leq \|r\|_{L^\infty(B)} \, |\Omega_Q|^{1/2} \operatorname{diam}(\Omega_Q),$$

where we abbreviate $B \coloneqq [\underline{a}, \overline{a}] \times [\underline{b}, \overline{b}]$. Altogether this implies that

$$\int_{\Omega_Q} y \,\mathrm{d}x \geq -\|r\|_{L^\infty(B)} \, |\Omega_Q| \operatorname{diam}(\Omega_Q)^2.$$

Moreover, it holds that $Q \in D$ implies $\Omega_Q \subset B$, thus

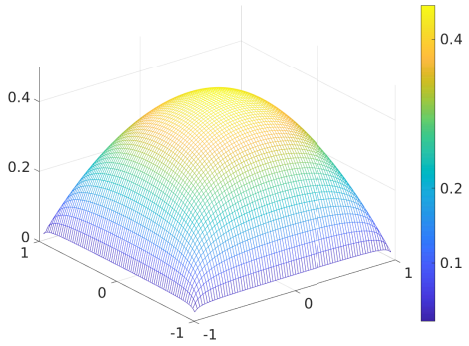
$$\int_{\Omega_Q} y \,\mathrm{d}x \geq -\|r\|_{L^\infty(B)} \, |B| \operatorname{diam}(B)^2,$$

which concludes the proof. □

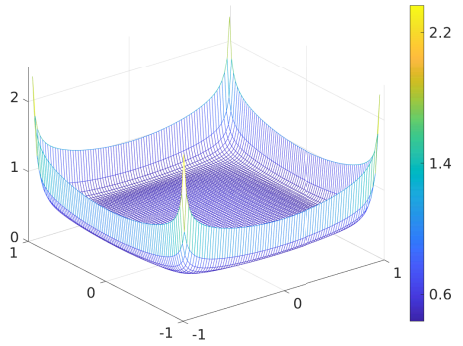
Remark 6.2.4. *Depending on the specific form of the shape functional j, it may be possible to obtain an existence result even with one or several of the coefficients β_j in (5.22) equal to zero. For instance, suppose that the $j\colon \mathcal{M}_+(\Delta; Q_{\mathrm{ref}}) \to \mathbb{R}$ is such that there exists $A_0 > 0$ and $\varepsilon > 0$ for which $A_Q < A_0$ implies $j(Q) \geq j^* + \varepsilon$, where j^* is the infimum of j on $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$. Then, the second term in (5.22) can be omitted, i. e., β_2 can be chosen equal to zero.*

Moreover, if the shape functional $j(Q)$ is bounded below on $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$, such as a quadratic tracking-type or compliance-type objective, the existence of solutions follows from corollary 6.2.2 and there is no need to impose a hold-all domain constraint.

We now revisit example (6.3), which served as a counterexample to the existence of solution for discrete shape optimization problems in section 6.1. With the penalty f_2 added, the existence of a solution now follows from proposition 6.2.3. The definition of a hold-all is actually not required since the boundary is fixed. For the same reason, the boundary self-intersection term in f_2 is not necessary, i. e., we use the cut-off function described in remark 5.3.4. To confirm the existence of a solution for this simple example, figure 6.2b shows a comparison of the shape functionals with and without penalization, the latter with parameters $\beta_1 = 0.1$, $\beta_2 = 0.01$, $\beta_3 = 0$ and $\beta_4 = 0.05$. As in figure 6.1a, the right-hand side is chosen as $r \equiv 1$ in (6.3).



(a) Shape functional without penalization as a function of the nodal position q_5.



(b) Shape functional with penalization ($\beta_1 = 0.1$, $\beta_2 = 0.01$, $\beta_3 = 0$ and $\beta_4 = 0.05$) as a function of the nodal position q_5.

Figure 6.2. Transforming problem (6.3) (left) into one which has a solution (right) by adding the penalty function f_2.

6.2.1 Optimality Conditions

Having established the existence of solutions of the penalized problem (6.8), we derive its first-order optimality conditions. The addition of the indicator function I_D can be considered only of a formal nature. In practice, one can choose a large enough hold-all domain, so no constraint needs to be considered. This is precisely the approach we follow in what remains of this chapter.

We use the Lagrangian approach described in subsection 2.1.3. Similarly to theorem 2.1.9, we know that on a manifold a stationary point $Q^* \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ of $j + f_2 \colon \mathcal{M}_+(\Delta; Q_{\text{ref}}) \to \mathbb{R}$ is characterized by vanishing directional derivatives, i. e.,

$$\mathrm{d}_Q[j + f_2]\,[Q^*][V] = 0 \quad \text{for all } V \in \mathcal{T}_{Q^*}\mathcal{M}_+(\Delta; Q_{\text{ref}}), \tag{6.9}$$

where $\mathcal{T}_{Q^*}\mathcal{M}_+(\Delta; Q_{\text{ref}})$ denotes the tangent space to $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ at Q^* (cf., Boumal, 2020, Prop. 4.4, p. 61). Recall that $\mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$ agrees with $\mathbb{R}^{2\times N_V}$.

Using any Riemannian metric $(\cdot\,, \cdot)_{Q^*}$ on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$, we can define the gradient via

$$(\operatorname{grad}[j + f_2]\,(Q^*)\,, V)_{Q^*} = \mathrm{d}_Q[j + f_2]\,[Q^*][V] \quad \text{for all } V \in \mathcal{T}_{Q^*}\mathcal{M}_+(\Delta; Q_{\text{ref}}) \tag{6.10}$$

and, equivalently to (6.9), write

$$\operatorname{grad}[j + f_2]\,(Q^*) = 0. \tag{6.11}$$

This leads to the following formulation of the first-order necessary optimality conditions.

Proposition 6.2.5. *Let Q be a locally optimal solution to (6.4) with associated state y. Then, there exists a unique adjoint state $p \in S_0^1(\Omega)$ such that the following system of equations is*

satisfied:

$$\textit{(state equation)} \quad \int_{\Omega_Q} \nabla y \cdot \nabla e_a \,\mathrm{d}x - \int_{\Omega_Q} r\, e_a \,\mathrm{d}x = 0 \qquad \textit{for all } a = 1, \dots, N_V, \tag{6.12a}$$

$$\textit{(adjoint equation)} \quad \int_{\Omega_Q} \nabla p \cdot \nabla e_b \,\mathrm{d}x + \int_{\Omega_Q} e_b \,\mathrm{d}x = 0 \qquad \textit{for all } b = 1, \dots, N_V, \tag{6.12b}$$

$$\begin{aligned}\textit{(design equation)} \quad & \int_{\Omega_Q} y \operatorname{div} V_i \,\mathrm{d}x + \int_{\Omega_Q} \nabla y \cdot \left[\left(\operatorname{div} V_i - DV_i - DV_i^{\mathrm{T}}\right) \nabla p \,\mathrm{d}x\right] \\ & - \int_{\Omega_Q} \operatorname{div}(rV_i)\, p \,\mathrm{d}x + \frac{\partial f_2(Q; Q_{\mathrm{ref}})}{\partial (\operatorname{vec} Q)_i} = 0 \quad \textit{for all } i = 1, \dots, 2\,N_V.\end{aligned} \tag{6.12c}$$

Here $\{e_a\}_{a=1}^{N_V}$ *is the standard nodal finite element basis of* $S_0^1(\Omega_Q)$*. The vector fields* V_i *are defined as follows:*

$$\begin{cases} V_i = (e_{(i+1)/2}, 0)^{\mathrm{T}} & \textit{if } i \textit{ is odd,} \\ V_i = (0, e_{i/2})^{\mathrm{T}} & \textit{if } i \textit{ is even.} \end{cases} \tag{6.13}$$

Proof. Let us start by fixing some notation. Recall that Ω_Q is the resulting triangular mesh with connectivity Δ and node coordinates Q. In the same way, we denote by K_Q the 2-faces of $\Sigma_Q(\Delta)$ given by $\operatorname{conv}_Q(i_0^k, i_1^k, i_2^k)$.

Assuming that $y \in S_0^1(\Omega_Q)$ implies $y = \sum_{b=1}^{N_V} \vec{y}_b e_b$, where $\vec{y} = [\vec{y}_1, \dots, \vec{y}_{N_V}]^{\mathrm{T}}$ is a vector of real numbers. The Lagrangian of the problem is then given by the following expression:

$$\begin{aligned} L(Q, \vec{y}, \vec{p}) = & \sum_{K_Q \in \Omega_Q} \sum_{b=1}^{N_V} \vec{y}_b \int_{K_Q} e_b \,\mathrm{d}x + f_2(Q; Q_{\mathrm{ref}}) + \sum_{K_Q \in \Omega_Q} \sum_{a=1}^{N_V} \sum_{b=1}^{N_V} \vec{y}_b \vec{p}_a \int_{K_Q} \nabla e_a \cdot \nabla e_b \,\mathrm{d}x \\ & - \sum_{K_Q \in \Omega_Q} \sum_{a=1}^{N_V} \vec{p}_a \int_{K_Q} r(\xi_{K_Q})\, e_a \,\mathrm{d}x, \end{aligned}$$

where $\vec{p} \in \mathbb{R}^{N_V}$, and ξ_{K_Q} denotes the center of the triangle K_Q. Notice that the basis functions e_a also depend on Q, even though we do not state it explicitly.

Following the Lagrangian approach, we know the adjoint equation is given by

$$0 = \mathrm{d}_y L(Q, \vec{y}, \vec{p}) = \sum_{K_Q \in \Omega_Q} \sum_{a=1}^{N_V} \vec{p}_a \int_{K_Q} \nabla e_a \cdot \nabla e_b \,\mathrm{d}x + \sum_{K_Q \in \Omega_Q} \int_{K_Q} e_b \,\mathrm{d}x. \tag{6.14}$$

for all $b = 1, \dots, N_V$. Therefore, $p = \sum_{a=1}^{N_V} \vec{p}_a e_a$ belongs to $S_0^1(\Omega_Q)$ and satisfies (6.12b).

On the other hand, the derivatives with respect to the node positions are computed using the Eulerian semi-derivative, i. e.,

$$(\mathrm{d}_Q L(\ Q, \vec{y}, \vec{p}\ ))_i = \frac{\partial L(\ Q, \vec{y}, \vec{p}\ )}{\partial (\operatorname{vec} Q)_i} = \left.\frac{\partial L(\ {}^tQ, {}^t\vec{y}, {}^t\vec{p}\ )}{\partial t}\right|_{t=0}, \qquad \forall i = 1, \dots, 2\,N_V,$$

where ${}^tQ = {}^tT^i(\Omega_Q)$ is the collection of perturbed node positions, and we denote by $\Omega_{{}^tQ}$ its associated perturbed domain. Moreover, ${}^t\vec{y}$, ${}^t\vec{p}$ are the solutions of the state and adjoint equations at the domain $\Omega_{{}^tQ}$, respectively. As suggested in Berggren, 2010, Eq. (9), we

choose the transformation ${}^tT^i$ given by the following expression:

$$ {}^tT^i(x) = x + t \begin{bmatrix} e_{(i+1)/2}(x) \\ 0 \end{bmatrix} \quad \text{if } i \text{ is odd}, \quad {}^tT^i(x) = x + t \begin{bmatrix} 0 \\ e_i(x) \end{bmatrix} \quad \text{if } i \text{ is even}. \tag{6.15} $$

We recall e_i is the continuous piecewise linear finite element global basis function at the node i. Choosing this transformation has two direct implications. First, the planar edges of the mesh will remain planar under this transformation. Second, the material derivative of the basis functions equals zero (cf., Berggren, 2010, Ex. 3, p. 34). This property is sometimes also referred to as the global basis functions being convected, see e. g., Souli, Zolésio, 1993, Lem. 3.2, p. 191, and Cor. 3.1, p. 192. Finally, thanks to the definition of V_i given in (6.13), the transformations satisfy ${}^tT^i(x) = (\mathrm{id} + tV_i)(x)$.

Note that the computation of $\partial\, {}^t\vec{y}/\partial t$, $\partial\, {}^t\vec{p}/\partial t$ can be understood as the analog of the material derivative in the discrete setting. Using the fact that the global basis functions are convected we obtain the following systems of equations for $\partial\, {}^t\vec{y}/\partial t$ and $\partial\, {}^t\vec{p}/\partial t$. The derivative of ${}^t\vec{y}$ w.r.t. t, denoted by $\dot{\vec{y}}_b = \left.\frac{\partial\, {}^t\vec{y}_b}{\partial t}\right|_{t=0}$ solves the following equation:

$$ \begin{aligned} \sum_{K_Q \in \Omega_Q} \sum_{b=1}^{N_V} \dot{\vec{y}}_b \int_{K_Q} \nabla e_b \cdot \nabla e_a \,\mathrm{d}x = &- \sum_{K_Q \in \Omega_Q} \sum_{b=1}^{N_V} \vec{y}_b \int_{K_Q} \nabla e_b \cdot [A'(0) \nabla e_a] \,\mathrm{d}x \\ &+ \sum_{K_Q \in \Omega_Q} \int_{K_Q} \mathrm{div}(rV_i)\, e_a \,\mathrm{d}x, \end{aligned} \tag{6.16} $$

for all $a = 1, \ldots, N_V$, and with $A'(0) = \mathrm{div}(V_i) - DV_i - DV_i^{\mathrm{T}}$.

In the same way, $\dot{\vec{p}}_a = \left.\frac{\partial\, {}^t\vec{p}_a}{\partial t}\right|_{t=0}$ solves the following system:

$$ \begin{aligned} \sum_{K_Q \in \Omega_Q} \sum_{a=1}^{N_V} \dot{\vec{p}}_a \int_{K_Q} \nabla e_a \cdot \nabla e_b \,\mathrm{d}x = &- \sum_{K_Q \in \Omega_Q} \sum_{a=1}^{N_V} \vec{p}_a \int_{K_Q} \nabla e_a \cdot [A'(0) \nabla e_b] \,\mathrm{d}x \\ &+ \sum_{K_Q \in \Omega_Q} \int_{K_Q} \mathrm{div}(V_i)\, e_b \,\mathrm{d}x, \end{aligned} \tag{6.17} $$

for all $b = 1, \ldots, N_V$. Equations (6.16) and (6.17) define uniquely the quantities $\dot{\vec{y}}$ and $\dot{\vec{p}}$. Therefore, the derivative of $L({}^tQ, {}^t\vec{u}, {}^t\vec{p})$ w.r.t. t, can be obtained by using the substitution rule of integration given in (2.29), deriving with respect to t and owing the definitions of $\dot{\vec{y}}$ and $\dot{\vec{p}}$ given in equations (6.16) and (6.17). Thus,

$$ \begin{aligned} 0 = \frac{\partial L}{\partial (\mathrm{vec}\, Q)_i} = &\sum_{K_Q \in \Omega_Q} \sum_{b=1}^{N_V} \vec{y}_b \int_{K_Q} \mathrm{div}(V_i)\, e_b \,\mathrm{d}x \\ &+ \sum_{K_Q \in \Omega_Q} \sum_{a=1}^{N_V} \sum_{b=1}^{N_V} \vec{y}_b \vec{p}_a \int_{K_Q} \nabla e_a \cdot [A'(0) \nabla e_b] \,\mathrm{d}x \\ &- \sum_{K_Q \in \Omega_Q} \sum_{a=1}^{N_V} \vec{p}_a \int_{K_Q} \mathrm{div}(rV_i)\, e_a \,\mathrm{d}x + \frac{\partial f_2(Q; Q_{\mathrm{ref}})}{\partial (\mathrm{vec}\, Q)_i}. \end{aligned} \tag{6.18} $$

Using the notation $\nabla y = \sum_{b=1}^{N_V} \vec{y}_b \nabla e_b$ and $\nabla p = \sum_{a=1}^{N_V} \vec{p}_a \nabla e_a$ then the optimality system for problem (6.4) given in equations (6.12a) to (6.12c) is obtained. □

6.3 Steepest Descent Method on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$

In this section, we briefly describe a general steepest descent method for the solution of the model problem (6.4) on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. The description of the method is kept generic since we wish to conduct numerical experiments for various choices of the Riemannian metric and the retraction later on in section 6.4. Clearly, higher-order optimization methods such as quasi-Newton or Newton methods are known to be advantageous with respect to their local convergence properties. However, a quasi-Newton method would require an implementation of the parallel transport or, more generally, a vector transport associated with the chosen retraction. By contrast, a Newton method would require the evaluation of the second-order covariant derivative of the penalized shape functional. Both of these topics are outside the scope of this thesis.

As in the vector space case, the Riemmanian steepest descent method is an iterative method that generates a sequence of improved estimates from a given initial guess. The negative of the Riemannian gradient (6.10) provides the direction in which the value of the shape functional will be improved. To generate a new iteration of the algorithm (which also belongs to the manifold) for the given initial velocity provided by the negative Riemannian gradient, one usually needs a generalization of the notion of straight lines. To this end, the definitions of the exponential map and retractions are required.

The exponential map $\exp_Q \colon \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}}) \to \mathcal{M}_+(\Delta; Q_{\text{ref}})$ at the point Q which belongs to $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is defined as

$$V \mapsto \exp_Q V \coloneqq \gamma_{Q,V}(1), \tag{6.19}$$

where $\gamma_{Q,V}(t)$ denotes the geodesic, starting at Q with initial velocity V, evaluated at time t.

Loosely speaking a retraction is a function that for a given pair $(Q, V) \in \mathcal{T}\mathcal{M}_+(\Delta; Q_{\text{ref}})$ (the tangent bundle of $\mathcal{M}_+(\Delta; Q_{\text{ref}})$) picks a particular curve starting at Q, with initial velocity $V \in \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$, which remains on the manifold. Additionally, it is assumed that the choice of the curve depends smoothly on (Q, V). For now, a retraction is to be understood as a function defined for each Q on $\mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\text{ref}})$ such that $V \mapsto \operatorname{retr}_Q(V) \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$. See definition A.3.1 for a formal definition.

Having introduced the notions of Riemannian gradient, exponential map, and retractions, we are now ready to describe the Riemannian steepest descent method given in algorithm 2.

As already mentioned, there exist various possible choices of the Riemannian metrics and retractions, which were described in subsections 5.1.1 and 5.1.2 and sections 5.2 and 5.3. To keep this chapter as self-contained as possible, we briefly recall them. The first and most obvious choice is the Euclidean Riemannian metric, given by $(V\,,\,\widetilde{V})_Q^{\text{Euc}} = (\operatorname{vec} V) \cdot (\operatorname{vec} \widetilde{V})$, where vec is the vectorization operation, which stacks V column by column. The geodesics associated to the Euclidean metric are straight lines, which in this case, coincide with the perturbation of identity, i. e.,

$$\operatorname{retr}_Q^{\text{euc}}(tV) = Q_i + t\, V_i, \tag{6.20}$$

for all $i = 1, \ldots, 2N_V$, and $t \in \mathbb{R}$. The second option is the linear elasticity Riemannian metric given by the following expression.

$$(V\,,\,\widetilde{V})_Q^{\text{elas}} \coloneqq (\operatorname{vec} V) \cdot \mathbb{K}\,(\operatorname{vec} \widetilde{V}) + \delta\,(\operatorname{vec} V) \cdot \mathbb{M}\,(\operatorname{vec} \widetilde{V}).$$

where the matrix $\mathbb{K}$ is the finite element stiffness matrix for piecewise linear elements over the mesh defined by Q associated to the linear elasticity operator given in (5.4), and $\mathbb{M}$ is the mass matrix. The final choice, and the one in which we put particular emphasis is the

Algorithm 2: General formulation of the steepest descent method on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ for (6.4).

Data: reference mesh $Q_{\text{ref}} \in \mathcal{M}_+(\Delta) \subset \mathbb{R}^{2\times N_V}$ with oriented connectivity complex Δ
Data: maximum number of iterations $N_{\max}$
Data: Riemannian metric $(\cdot\,,\,\cdot)_Q$ on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$
Data: retraction $\operatorname{retr}_Q(\cdot)$ on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$
Result: approximate stationary point of the problem (6.4) on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$

while *stopping criterion is not satisfied and* $n < N_{\max}$ **do**
  set $Q_0 \coloneqq Q_{\text{ref}}$ and $n \coloneqq 0$
  compute the state y by solving (6.12a)
  compute the adjoint state p by solving (6.12b)
  evaluate the derivative $\mathrm{d}_Q[j + f_2](Q^n) \in \mathcal{T}^*_{Q^n}\mathcal{M}_+(\Delta; Q_{\text{ref}})$ via the left-hand side of (6.12c)
  find the negative gradient $d^n \in \mathcal{T}_{Q^n}\mathcal{M}_+(\Delta; Q_{\text{ref}})$ by solving the linear system

$$(d^n\,,\,V)_{Q^n} = -\,\mathrm{d}_Q[j + f_2](Q^n)[V] \quad \text{for all } V \in \mathcal{T}_{Q^n}\mathcal{M}_+(\Delta; Q_{\text{ref}})$$

  find a step size s_n via Armijo backtracking, satisfying

$$(j + f_2)(\operatorname{retr}_{Q^n}(s_n\, d^n)) \leq (j + f_2)(Q^n) + \sigma\, s_n\, \mathrm{d}_Q[j + f_2](Q^n)[d^n]$$

  update $Q^{n+1} \coloneqq \operatorname{retr}_{Q^n}(s_n\, d^n)$
  set $n \coloneqq n + 1$
end
return $Q^{n+1} \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$, *an approximate stationary point of* $j + f_2$

invariant under mesh refinements complete metric given in (5.34).

$$g_{ab}^{\text{invariant}} = \delta_a^b + \frac{\partial f_2}{\partial(\operatorname{vec} Q)^a}\frac{\partial f_2}{\partial(\operatorname{vec} Q)^b}, \quad a, b = 1, \ldots, 2\,N_V,$$

where f_2 is given in (5.22).

As already highlighted in subsection 5.4.2 this complete metric has the following properties. First, the representation of the metric is merely a rank-1 perturbation of the identity matrix; which implies the solution of the linear system to obtain the respective gradient of any function from its derivative is very efficient. This holds in particular for the penalized shape function. Second, we can, in principle, follow the geodesic with respect to this metric in negative gradient direction in the Armijo line search procedure. In other words, we can use the exponential map as the retraction. Due to the completeness of the metric, no artificial restriction of the step sizes is then required to avoid degenerate meshes, i. e., to remain on $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. This will be numerically verified in subsection 6.4.2.

Despite the simplicity of the metric (5.34), the geodesic equation must be solved numerically, as described in section 5.4. In practice, as confirmed by our experiments in section 6.4, this step in algorithm 2 is prohibitively expensive. However, even when combined with the Euclidean retraction, the new metric (5.34) performs very favorably in practice, at a lower numerical cost than the elasticity metric.

6.4 Numerical Investigations

This section aims to compare the performance of different combinations of Riemannian metrics and retractions within the steepest descent method, given in algorithm 2, for the

solution of a discretized, PDE-constrained shape optimization problem. We stick to the model problem (6.4) with right-hand side $r(x_1, x_2) = 2.5\,(x_1 + 0.4 - x_2^2)^2 + x_1^2 + x_2^2 - 1$, as previously used in Etling et al., 2020. Bartels, Wachsmuth, 2020 suggest to use this model based on the simple interpretation of the expected solution. Recall that our goal is to minimize $\int_{\Omega_Q} y \,\mathrm{d}x$ and notice that the sublevel set $\{x \in \mathbb{R}^2 \,|\, r(x) \leq 0\}$ is connected. Due to the maximum principle, we can therefore expect to find an optimal shape close to this sublevel set, at least in the continuous setting where a maximum principle is available. In the discrete setting, the maximum principle hinges upon the condition of nonobtuse triangles, which is not guaranteed a priori. Indeed, we did find obtuse triangles in most our experiments to occur. Figure 6.3 shows a contour plot of r for comparison with the obtained optimal shapes.

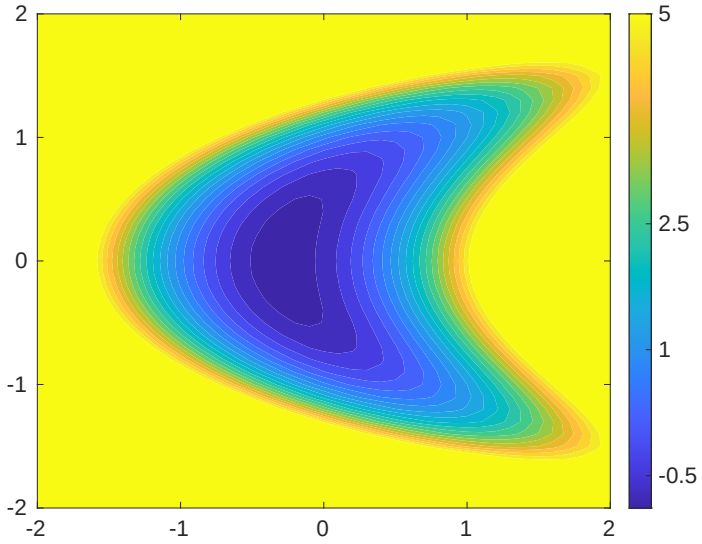


Figure 6.3. Contour plot of r.

The variants we compare are termed Euclidean-Euclidean, Elasticity-Euclidean, Complete-Euclidean and Complete-Complete. The first component of the name refers to the metric used for the evaluation of the shape gradient; see (6.10). The three choices indicate the Euclidean metric, the elasticity metric (5.3) and the new complete metric (5.34). Their precise parameters are specified further below. The second component of the name refers to the choice of the retraction, which is either Euclidean (6.20) or the exponential map (6.19), evaluated via numerical integration using algorithm 1.

In subsection 6.4.1 we describe the implementation details used throughout the numerical experiments. Then, three experiments are conducted to explore various points. In the first experiment we consider problem (6.4) without a penalty term in subsection 6.4.2. We confirm that, as expected, this problem then does not possess a solution. Consequently, this leads any gradient descent method, regardless of the metric employed, to ultimately produce degenerate meshes in the pursuit of smaller and smaller shape functional values. However, the variants Elasticity-Euclidean, and Complete-Complete still produce "good" iterates along the way, albeit at different iteration counts, while Euclidean-Euclidean breaks down early.

Our second experiment in subsection 6.4.3 targets the penalized problem, for which the existence of a solution can be proved. It turns out that now, as expected, the gradient descent method finds this solution regardless of the metric chosen, yet at different iteration numbers.

Computationally, we observe that the new metric outperforms Euclidean-Euclidean and also Elasticity-Euclidean for the problem under consideration, when combined with the Euclidean retraction (Complete-Euclidean).

In our third experiment in subsection 6.4.4, we therefore revisit the first strategy and compare the two most promising candidates, Elasticity-Euclidean and Complete-Euclidean, using finer meshes than before. Once again, it turns out that the use of the new metric may maintain better-quality meshes and requires less time per iteration compared to the elasticity metric.

The results thus far indicate that the typically ill-posed problem of minimizing a discrete shape functional may be tackled either by early stopping or by the addition of a penalty term. The penalty approach may be criticized since it requires the user to make a somewhat arbitrary choice of parameters β_1, β_2, β_3, β_4.

We found numerically that the hold-all domain assumption required for the proof of proposition 6.2.3 did not require to be enforced.

6.4.1 Implementation Details

Our implementation is achieved in MATLAB, using the `initmesh` function of the PDE toolbox for the generation of all initial meshes and the code provided by Koko, 2016b; a to assemble the elasticity stiffness and mass matrices required for the elasticity metric (5.3). All experiments were performed on a computer with an Intel Core i7-7500 CPU with 2.7 GHz and 16GiB RAM.

Initialization of the Armijo Backtracking Procedure As already described in algorithm 2, we use Armijo's condition

$$(j + f_2)(\mathrm{retr}_{Q^n}(s_n\, d^n)) \leq (j + f_2)(Q^n) + \sigma\, s_n\, \mathrm{d}_Q[j + f_2](Q^n)[d^n], \tag{6.21}$$

in order to guarantee sufficient decrease of the (penalized) shape functional through a backtracking procedure. It is well-known that the steepest descent method is not scale invariant and therefore relies on a judicious choice of the initial line search step size. We use the technique presented in Nocedal, Wright, 2006, p. 59, i. e., the candidate for the initial step size in iteration n is given by

$$\overline{s}_n = s_{n-1} \frac{\mathrm{d}_Q[j + f_2](Q^{n-1})[d^{n-1}]}{\mathrm{d}_Q[j + f_2](Q^n)[d^n]}.$$

This candidate step size gets overwritten in the initial iteration or when $\overline{s}_n$ becomes too small. We use the rule

$$s_n^{\text{initial}} = \begin{cases} \frac{1}{\|d^n\|_{Q^n}} & \text{if } n = 0 \text{ or } \overline{s}_n \,\|d^n\|_{Q^n} < 10^{-4}, \\ \overline{s}_n & \text{otherwise} \end{cases} \tag{6.22}$$

for this purpose. Should a trial step size fail to satisfy the Armijo condition (6.21), we repeatedly multiply it by a factor $\tau \in (0, 1)$ specified further below.

We recall that some of the variants of the algorithm involve the Euclidean retractions (6.20). In this case, mesh nodes move independently of each other and thus extra care needs to be taken regarding the trial step sizes in order to avoid degenerate meshes. We proceed as follows. When the Euclidean retraction is used, we deliberately fail the Armijo condition (6.21) for the trial step size s as long as the distance a node would travel is relatively large compared to the heights of any of its incident triangles. More precisely, we fail the Armijo condition as long as

$$s\, \|d_{i_\ell^k}\|_Q^{\text{euc}} \geq 0.5\, h_Q^\ell(i_0^k, i_1^k, i_2^k) \quad \text{for any } k = 1, \ldots, N_T \text{ and any } \ell = 0, 1, 2 \tag{6.23}$$

holds. Here $\|d_{i_\ell^k}\|_Q^{\mathrm{euc}}$ denotes the Euclidean norm of the subvector of the negative gradient direction d pertaining to the ℓ-th vertex of the k-th triangle, and h_Q^ℓ is the corresponding height, see (5.5). (For the purpose of readability, we temporarily dropped the iteration index n here.)

Armijo Backtracking with the Exponential Map In the experiment in subsection 6.4.2, we use Complete-Complete as one of the variants of algorithm 2. As opposed to all other variants using the Euclidean retraction, the geodesic equation with respect to the metric (5.34) must be integrated numerically, which is expected to be expensive. In order to avoid repeated evaluations of the geodesic in case the Armijo's condition (6.21) happens to fail for the initial trial step size, we make use of the re-scaling lemma; see e. g., Lee, 2011, Lem. 5.18, p. 127. This lemma states that for every initial data $Q \in \mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ and $d \in \mathcal{T}_Q\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$, trial step size and backtracking parameter $\tau > 0$, we have $\gamma_{Q,\tau\, s\, d}(1) = \gamma_{Q,s\, d}(\tau)$. When integrating the initial trial geodesic with velocity $s^{\mathrm{initial}}d$ until $t = 1$, our implementation of the numerical integrator thus stores the values at $t \in \{\tau, \tau^2, \ldots\}$. This can be conveniently achieved by setting $\tau = 0.5$ and using a number of time steps divisible by a sufficiently large power of 2.

Parameter Choices We keep the following parameters fixed for all experiments. For the Armijo line search, we use the acceptance and backtracking parameters $\sigma = 10^{-4}$ and $\tau = 0.5$. The linear elasticity metric given in (5.3) uses Lamé constants given by

$$\mu = \frac{E}{2(1+\nu)}, \quad \lambda = \frac{E\,\nu}{(1+\nu)(1-2\nu)}, \quad \delta = 0.2\,E, \tag{6.24}$$

with Young's modulus $E = 1$ and Poisson ratio $\nu = 0.4$.

As a measure of the quality of the generated meshes, we monitor the function

$$\Lambda(Q) = \sum_{k=1}^{N_T} \frac{1}{N_T} \frac{1}{\psi_Q(i_0^k, i_1^k, i_2^k)}, \tag{6.25}$$

which is part of the penalty function's definition (5.22), where $1/\psi$ is given by (5.23). We remind the reader that $\Lambda(Q) \geq 1$ holds, and 1 constitutes the best value while bad quality meshes correspond to large values of Λ.

We also recall that the penalty function f_2 serves two purposes: it renders the penalized problem well-posed if added to the shape functional, and it forms the basis for the complete metric (5.34). For flexibility, we allow two different sets of parameters β_j, $j = 1, \ldots, 4$ for both occurrences. They are denoted as $\beta_j^{\mathrm{penalty}}$ and $\beta_j^{\mathrm{metric}}$, respectively. For the problem under consideration, we do not run the risk of exterior self-intersections so we set $\beta_3^{\mathrm{penalty}} = \beta_3^{\mathrm{metric}} = 0$ for all experiments. This can be justified using a thresholding function as in remark 5.3.4. The remaining parameters are specified in each of the following sections as needed.

Derivative-Gradient Transformation The evaluation of the gradient (6.10) requires the solution of a linear system whenever the metric is not the Euclidean one. In case of the linear elasticity metric (5.3), we assemble the stiffness and mass matrices using the code provided by Koko, 2016b; a. The subsequent solve of the linear system was achieved using the default sparse direct solver of MATLAB. For the moderate size of the experiments conducted, a more sophisticated strategy such as a geometric multigrid method does not pay off.

In case of the complete metric (5.34), we exploit the fact that the associated matrix is is a rank-1 perturbation of the identity matrix and use the computation aspects described in subsection 5.4.2, i. e., the linear system (6.10) is solved using two iterations of the conjugate gradient method without preconditioning. Our implementation is matrix-free. The most

expensive part of this process is the evaluation of the first-order derivatives of the penalty function f_2.

Definition of Unsuccessful Experiments As a precautionary measure, we keep track of several indicators during the iteration of the gradient descent method algorithm 2. In particular, we verify that each search direction d^n is indeed a descent direction, i. e., $\mathrm{d}_{Q^n}[j + f_2](Q^n)[d^n] < 0$ holds. Moreover, we make sure that the signed areas (4.2) of all triangles remain positive for all iterates, which is a requirement for them to belong to the manifold $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$. As expected, these indicators were never found to be violated.

It can happen, however, that a close-to-degenerate mesh enforces very small trial step sizes due to (6.23) when the Euclidean retraction is used. Indeed, we declare a gradient descent run unsuccessful and stop as soon as a trial step size becomes smaller than 10^{-6}.

Stopping Criteria Choosing a stopping criterion is a delicate task. This is especially true in case of the unpenalized problem, which may not possess solutions, and early stopping (before the norm of the gradient becomes too small) becomes essential. Since the attempt to approximate the infimum results in degenerate meshes, using any criterion involving the value of the shape functional alone will also not be suitable. As a compromise, we therefore settle on a fixed number of iterations for the experiments in subsections 6.4.2 and 6.4.4, which concern the unpenalized problem.

For the penalized problem in subsection 6.4.3, which does have a solution, we can use a more classic approach. Since we compare different metrics, which entail different ways to measure the norm of the gradient, the gradient norm does not allow a fair comparison. We therefore resort to measuring the absolute change of the values of the penalized shape functional over a span of 5 past iterations, and use it as an stopping criterion. This results in stopping as soon as

$$\max_{m=1,\dots,5} \left\{ (j + f_2)(Q^{n-m}) - (j + f_2)(Q^n) \right\} < \mathrm{tol}. \tag{6.26}$$

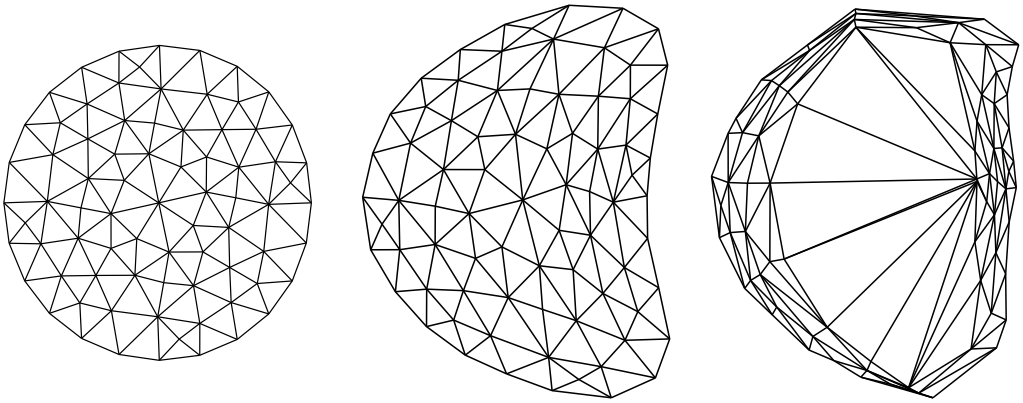
This is motivated by a condition proposed in Laurain, 2018, Sec. 6.15, p. 1324.

6.4.2 Experiment 1: Lack of Solutions for the Unpenalized Problem

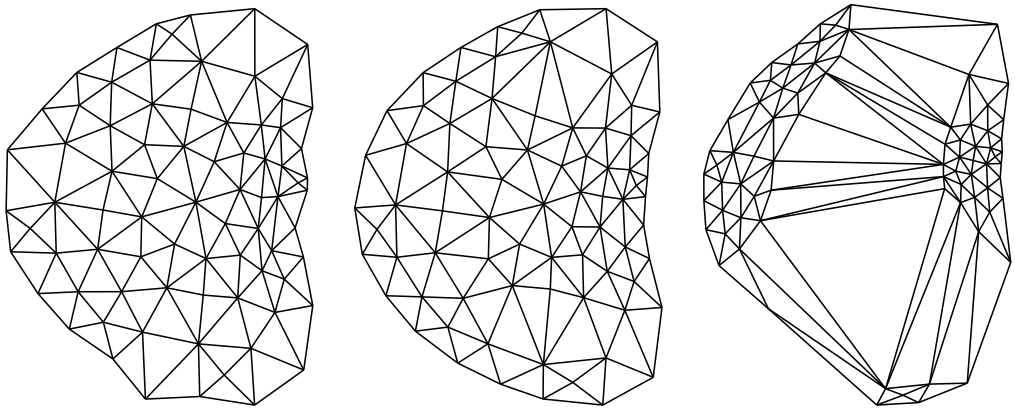
As was argued in section 6.1, discretized shape optimization problems in which the node positions serve as optimization variables can not be expected to possess a solution. Here we confirm this observation for our model problem (6.4) without a penalty, i. e., we set $\beta_j^{\mathrm{penalty}} = 0$ for all $j = 1, 2, 3, 4$.

Consequently, this leads any gradient descent method, regardless of the metric employed, to ultimately produce degenerate meshes in the pursuit of smaller and smaller shape functional values. We also trace back the specific nature of the degeneracy observed to an exploitation of the quadrature formula for the problem at hand.

We compare the variants Euclidean-Euclidean, Elasticity-Euclidean, and Complete-Complete. For the latter, we use the parameters $\beta_1^{\mathrm{metric}} = 10$, $\beta_2^{\mathrm{metric}} = 1$, $\beta_3^{\mathrm{metric}} = 0$ and $\beta_4^{\mathrm{metric}} = 0.077$. The initial mesh for this first experiment is a coarse triangulation of the unit disc containing $N_V = 77$ nodes and $N_T = 128$ triangles. The results are shown in figure 6.4 and table 6.1. The Euclidean-Euclidean variant breaks down in iteration 60 with too small a trial step size and a disastrous value of the mesh quality measure Λ and it is thus evaluated as an unsuccessful experiment. By contrast, the Elasticity-Euclidean and Complete-Complete variants produce meshes of comparable quality and similarly small values of the shape functional at iteration counts 150 and 15, respectively. As expected, both enter a phase of producing increasingly degenerate meshes afterwards before being stopped at iteration 1000. However, we observe that the deterioration of the mesh quality is more pronounced for the Elasticity-Euclidean variant.



Iterates from left to right: 15 (too early), 150 (good), 1000 (too late) for variant Elasticity-Euclidean.



Iterates from left to right: 5 (too early), 15 (good), 1000 (too late) for variant Complete-Complete.

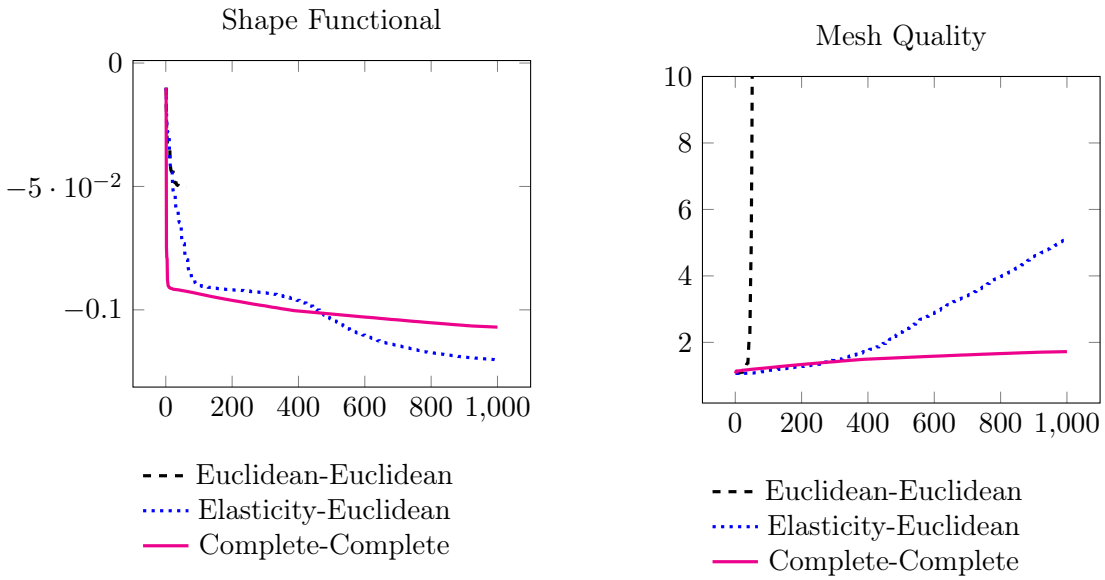
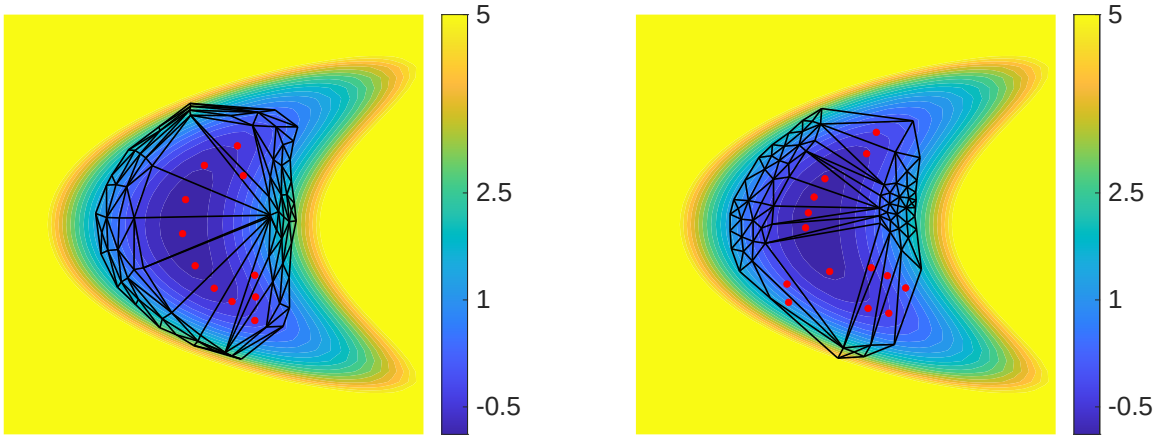


Figure 6.4. Results for the experiment described in subsection 6.4.2.

variant	iter (n)	$j(Q^n)$	$\Lambda(Q^n)$
Euclidean-Euclidean	60	-0.0502	259.9215
Elasticity-Euclidean	1000	-0.1157	3.6000
Complete-Complete	1000	-0.1031	1.5149

Table 6.1. Summary of the results obtained for the experiment described in subsection 6.4.2.

As announced earlier, it is illustrative to study the meshes for the Elasticity-Euclidean and Complete-Complete variants at the final iteration 1000. As shown in figure 6.5, large triangles are produced where the values of the PDE's right-hand side function r are smallest. This is due to the discrete shape functional involving a quadrature formula for the evaluation of the element load vector, which evaluates the right-hand side only in the triangle centers, some of which are marked by red dots. Given the opportunity, it thus can be concluded that the optimizer exploits the quadrature error.



(a) Variant Elasticity-Euclidean (b) Variant Complete-Complete

Figure 6.5. Location of the centers of the larger triangles at iterate 1000, superimposed on a contour plot of the right-hand side r.

A first conclusion at this point is that a gradient method, applied to an unpenalized problem without a solution, might be successful to produce a reasonably good approximation to the solution of the continuous shape optimization problem, provided that it is stopped sufficiently early. As already noted, the Complete-Complete variant reaches this convenient stopping point at a much earlier iteration number. However, the picture changes when comparing the respective run-times.

Table 6.2 shows the timings for the first 5 gradient iterations of each of the variants. The column `state` summarizes the time devoted to solving the state equation at least once per iteration, depending on the number of Armijo backtracking steps. The column `dObj` represents the time invested in assembling the derivative of the shape derivative. Likewise, the column `backt` presents the time required to check whether the line search trial step sizes satisfy both (6.23) (in case of the Euclidean retraction) and Armijo condition (6.21). The column `grad` shows the time needed in the transformation of the derivative to the gradient, i. e., for the solution of the linear system (6.10). Finally, the column `retr` shows the time to evaluate the retraction. This is not relevant for the Euclidean retraction, but

only in case the geodesic equation associated with the metric (5.34) is solved numerically. The latter is achieved using the implementation of the Störmer–Verlet scheme detailed in subsection 5.4.2. We used 1024 time steps for this purpose to ensure convergence of the fixed-point solver for the implicit sub-step.

As the timings clearly show, the numerical integration of the geodesic equation associated with the metric (5.34) is prohibitively expensive in the Complete-Complete variant. Therefore, we replace the Complete-Complete variant by Complete-Euclidean for further experiments, i. e., we combine the metric (5.34) with the Euclidean retraction.

Variant	total	state	dObj	backt	grad	retr
Euclidean-Euclidean	0.489 s	0.159 s	0.069 s	0.113 s	–	–
Elasticity-Euclidean	0.284 s	0.110 s	0.027 s	0.043 s	0.047 s	–
Complete-Complete	709.68 s	0.179 s	0.032 s	0.076 s	0.030 s	709.296 s

Table 6.2. Execution times for 5 iterations for the variants used in subsection 6.4.2.

6.4.3 Experiment 2: Solving the Penalized Problem

Our second experiment targets the penalized problem, for which the existence of a solution was proved in proposition 6.2.3. Due to the excessive time associated with the numerical integration of the geodesic equation associated with the metric (5.34), we only consider the Euclidean retraction (6.20) from now on. We thus compare the variants Euclidean-Euclidean, Elasticity-Euclidean and Complete-Euclidean. We solve the penalized problem with three different sets of parameters given in table 6.3. The initial mesh is again a coarse triangulation of the unit disc containing $N_V = 146$ nodes and $N_T = 258$ triangles. The parameters for the metric (5.34) β_j^{metric} are the following: $\beta_1^{\text{metric}} = 10$, $\beta_2^{\text{metric}} = 1$, $\beta_3^{\text{metric}} = 0$ and $\beta_4^{\text{metric}} = 1.46$.

Parameter set	β_1^{penalty}	β_2^{penalty}	β_3^{penalty}	$\beta_4^{\text{penalty}}/N_V$
1	1	0.5	0.0	0.1
2	0.1	0.01	0.0	0.001
3	0.015	0.005	0.0	0.0005

Table 6.3. Description of the parameter set for the experiment in subsection 6.4.3.

Since we know that the problem has a solution, we can use the stopping criterion in (6.26) with a tolerance of tol $= 10^{-6}$. The number of iterations and the final values of the shape functional and the penalized shape functional are shown in table 6.4. Figure 6.6 shows the final iterates obtained for each variant, which are very similar to each other.

The first fact to highlight is that variant Euclidean-Euclidean performs surprisingly well on the penalized problem, even for moderately small values of the penalty parameters β_j^{penalty} (parameter sets 1 and 2). However, it does not quite converge within 1000 iterations for parameter set 3. Variants Elasticity-Euclidean and Complete-Euclidean perform equally well, but the latter is faster; see table 6.5. Both variants are also comparable to each other and better compared to Euclidean-Euclidean with respect to the values of the shape functional and the mesh quality, as shown in figure 6.7.

We also mention that the evaluation of the derivative of the penalty function (column `dPen`), which might be a concern, does not require a major computational effort, at least not for the meshes of this size.

In conclusion, we find that the presence of the penalty terms helps preserve the mesh quality for all variants. The variant Complete-Euclidean performs fastest at a numerical cost very close to that of Euclidean-Euclidean. This is partly due to the small cost of solving for the gradient, see (6.10). Admittedly, the differences are small for the coarse mesh under consideration. Therefore, we conduct a series of experiments in the following subsection 6.4.4 with finer meshes.

Parameter set	Variant	iter (n)	$j(Q^n)$	$j(Q^n) + f_2(Q^n)$	$\Lambda(Q^n)$
1	Euclidean-Euclidean	56	-0.056	1.158	1.042
	Elasticity-Euclidean	87	-0.056	1.158	1.042
	Complete-Euclidean	59	-0.056	1.158	1.042
2	Euclidean-Euclidean	363	-0.091	0.019	1.050
	Elasticity-Euclidean	261	-0.091	0.019	1.050
	Complete-Euclidean	281	-0.091	0.019	1.049
3	Euclidean-Euclidean	1000	-0.0919	-0.0729	1.1165
	Elasticity-Euclidean	276	-0.0921	-0.0733	1.0895
	Complete-Euclidean	289	-0.0923	-0.0734	1.0945

Table 6.4. Summary of the results obtained for the experiment described in subsection 6.4.3.

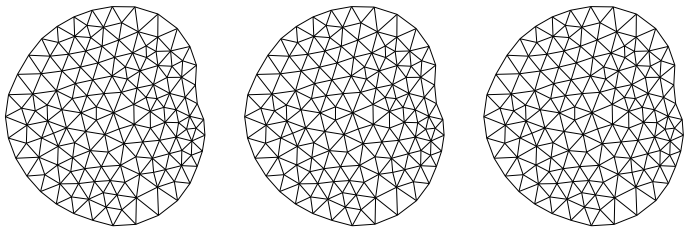
Variant	Total	state	dObj	dPen	backt	grad
Euclidean-Euclidean	0.488 s	0.150 s	0.045 s	0.031 s	0.118 s	–
Elasticity-Euclidean	0.436 s	0.181 s	0.036 s	0.023 s	0.038 s	0.057 s
Complete-Euclidean	0.327 s	0.131 s	0.029 s	0.022 s	0.034 s	0.026 s

Table 6.5. Execution times for 5 iterations for the variants used in subsection 6.4.3.

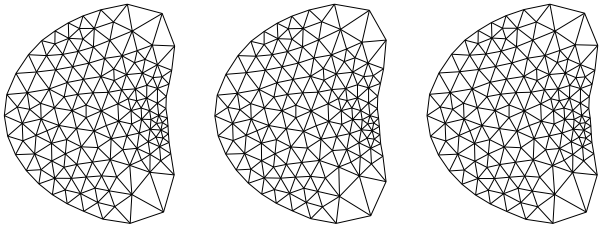
6.4.4 Experiment 3: Unpenalized Problem with Finer Meshes

The penalty approach may be criticized since it requires the user to make a somewhat arbitrary choice of the penalty parameters β_j^{penalty}, $j = 1, \dots, 4$. Therefore we revisit here the unpenalized problem, aware of the fact that the discretized problem does not possess a solution any gradient method could converge to. In contrast to the results of subsection 6.4.2, the meshes are now finer, and we only compare the two most promising gradient descent variants, Elasticity-Euclidean and Complete-Euclidean. We consider four mesh levels. The first one contains $N_V = 541$ nodes and $N_T = 1016$ triangles. The second one has $N_V = 775$ nodes and $N_T = 1468$ elements. The third possesses $N_V = 2191$ nodes and $N_T = 4252$ triangles. Finally, mesh level four has $N_V = 13455$ nodes and $N_T = 26588$ triangles.

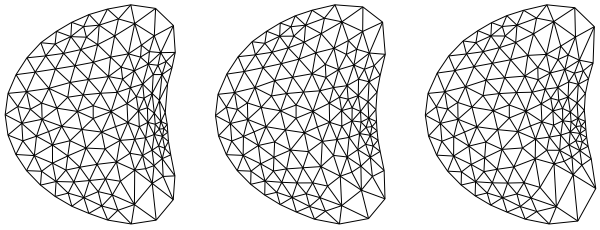
We allow the algorithm to run 500 iterations and are mainly interested in comparing the values of the shape functional and the mesh quality. The results can be seen in figures 6.8



(a) Parameter set 1



(b) Parameter set 2

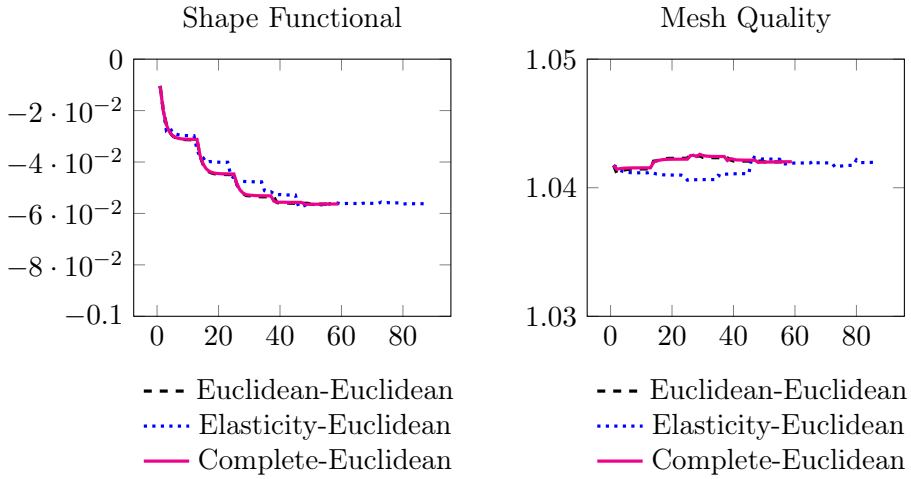


(c) Parameter set 3

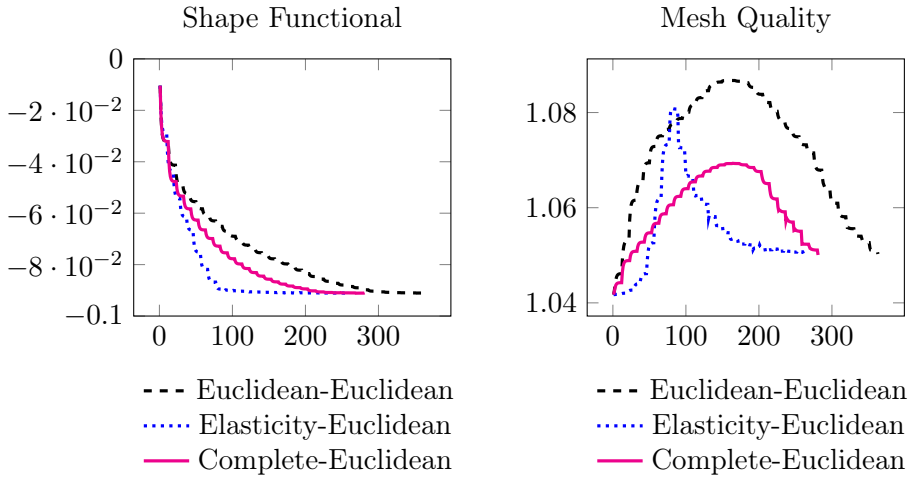
Figure 6.6. Final iterates obtained for the penalized problem with variants Euclidean-Euclidean (left), Elasticity-Euclidean (middle) and Complete-Euclidean (right) as described in subsection 6.4.3.

and 6.9. We infer that both variants, Elasticity-Euclidean and Complete-Euclidean, achieve a similar decrease of the shape functional. The variant Elasticity-Euclidean needs fewer iterations to reach the plateau, but the Complete-Euclidean maintains a better mesh quality measure and has less numerical cost per iteration. The latter is reflected in table 6.6. Here we separately display the time required to "assemble" the matrices representing the Riemannian metric in column `assemG`. More precisely, as in all experiments before, we only actually form this matrix in case of Elasticity-Euclidean, and employ a sparse direct solver to obtain the solution of the gradient equation (6.10). In case of Complete-Euclidean, we continue to work with matrix-vector products and the conjugate gradient solver. In this case, the column `assemG` is dominated by the time to evaluate the first-order derivative of the penalty function. We also observe that the time required to solve the gradient equation (6.10) remains essentially constant in case of Complete-Euclidean while the time for the direct solver in case of Elasticity-Euclidean grows with the problem size.

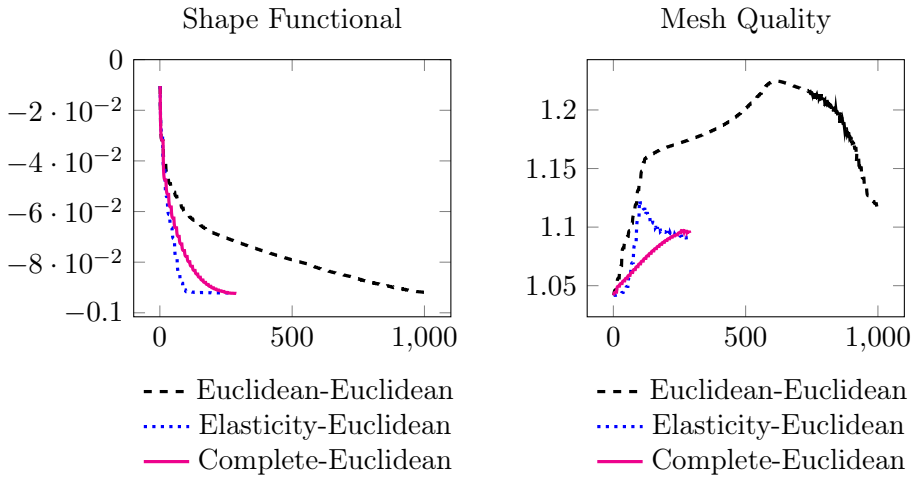
An inspection of the meshes at iteration 500 in figure 6.8 shows triangles closer to equilateral when using Complete-Euclidean and more elongated in case of Elasticity-Euclidean, as reflected by mesh quality plot in figure 6.9. Moreover, the triangles are smaller and the vertices more dense in regions which have deformed most compared to the initial circle mesh. We can consider this behavior as a natural redistribution of the nodes promoted by the use of the complete metric.



(a) Parameter set 1

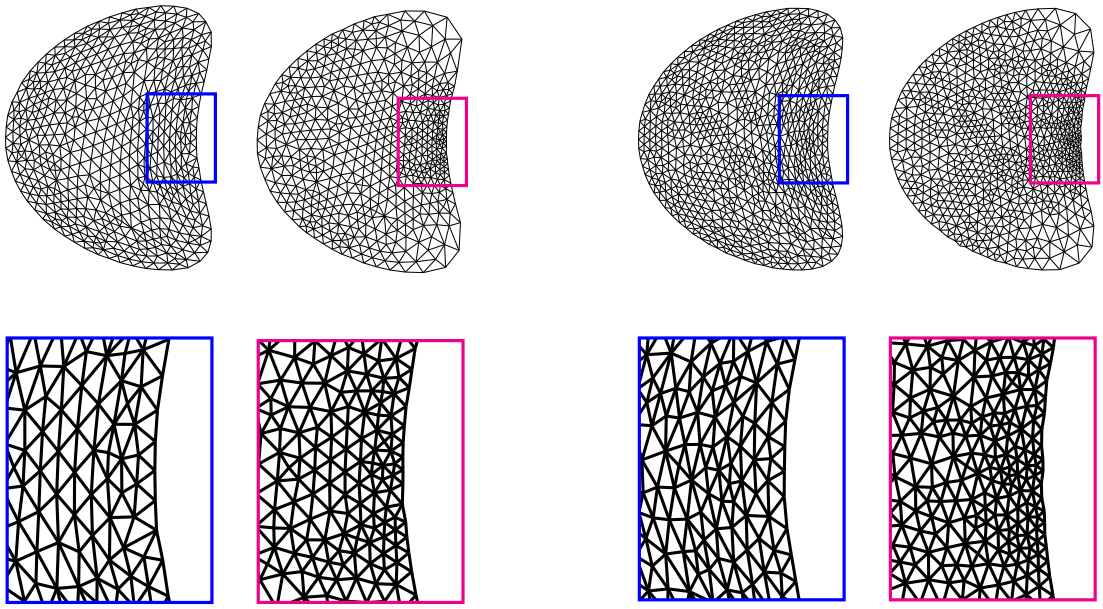


(b) Parameter set 2



(c) Parameter set 3

Figure 6.7. Shape functional and mesh quality for the penalized problem described in subsection 6.4.3.



Mesh Level 1 Mesh Level 2

Figure 6.8. 500th iterate in case of Elasticity-Euclidean (blue) and Complete-Euclidean (magenta).

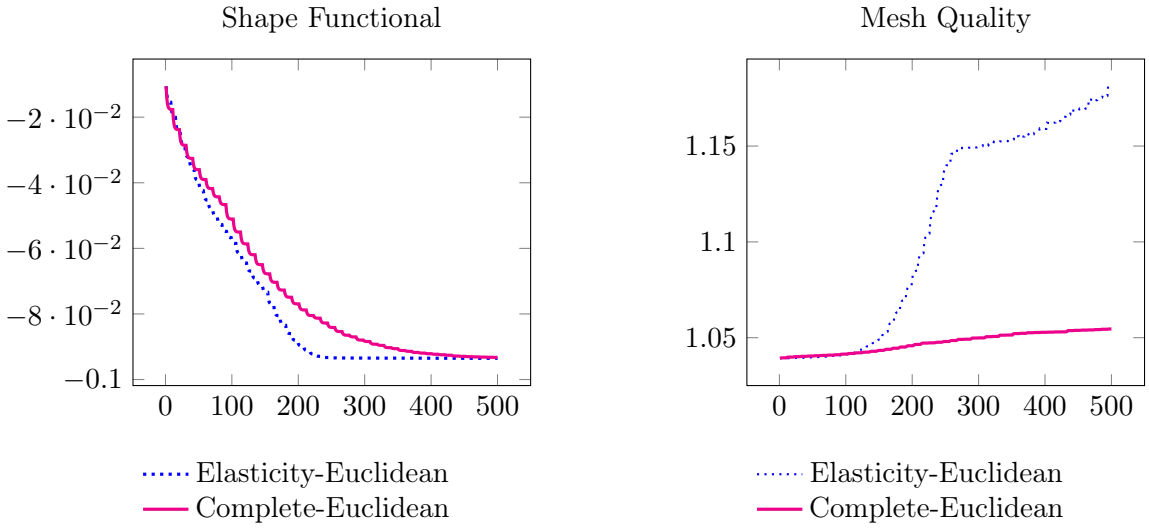


Figure 6.9. Shape functional and mesh quality for the unpenalized problem at mesh level 2 described in subsection 6.4.4.

Mesh level	Variant	Total	dObj	backt	assemG	grad
1	Elasticity-Euclidean	0.557 s	0.064 s	0.128 s	0.049 s	0.021 s
	Complete-Euclidean	0.378 s	0.041 s	0.044 s	0.034 s	0.022 s
2	Elasticity-Euclidean	0.400 s	0.042 s	0.061 s	0.044 s	0.053 s
	Complete-Euclidean	0.352 s	0.041 s	0.052 s	0.028 s	0.013 s
3	Elasticity-Euclidean	0.656 s	0.090 s	0.115 s	0.093 s	0.062 s
	Complete-Euclidean	0.637 s	0.090 s	0.104 s	0.059 s	0.016 s
4	Elasticity-Euclidean	2.530 s	0.409 s	0.574 s	0.475 s	0.345 s
	Complete-Euclidean	1.964 s	0.383 s	0.547 s	0.281 s	0.021 s

Table 6.6. Execution times for 5 iterations for the variants used in subsection 6.4.4.

7 Conclusions and Outlook

In this thesis, we analyzed two-dimensional PDE-constrained shape optimization problems under the discretize-then-optimize paradigm. In what follows, we summarize the obtained results and briefly overview possible future research directions.

We introduced the background knowledge about the analysis of general PDE-constrained problems in chapter 2. Section 2.1 collected the main results about the existence of solutions of linear, elliptic partial differential equations, the existence of solutions for optimal control problems, and their optimality conditions. In section 2.2 we discussed the main discretization concepts, with special focus in the differences between the *optimize-then-discretize* and *discretize-then-optimize* approaches. Section 2.3 was dedicated to describe the generalities of the finite element method.

We devoted chapter 3 to the description of the optimize-then-discretize approach for the solution of PDE-constrained shape optimization problems. Since this is the most commonly used approach, one can also consider this chapter as the state-of-the-art of computational shape optimization. Working under the optimize-then-discretize approach means we have to choose a continuous shape representation. Therefore, we presented an overview of the most common continuous shape representations in section 3.1. It is well-known that shape optimization problems often do not have a solution. To elaborate on this, we presented in section 3.2 two examples which do not possess a solution. In the first example topological changes to the mesh where allowed, while in the second one they were not. To derive the first-order optimality conditions of PDE-constrained shape optimization problems, we briefly described the basic notions of shape calculus in section 3.3, emphasizing the different formulations of the shape derivative. Section 3.4 described the main features of the discretization of shape optimization problems. To highlight the difference between the continuous and discretized problem, we recalled the example provided in Glowinski, He, 1998, where the steepest descent algorithm does not converge when using the optimize-then-discretize approach. However, it does converge and with good convergence properties under the discretize-then-optimize paradigm. The finite element method is the most used approach to discretize the state equation, and this implies that the underlying mesh is used to represent the discrete shape. In this context, it is common to experience degeneracy on the mesh quality as the optimization progresses. Various techniques had been developed to circumvent this obstacle, and we have collected some of them in section 3.5.

Chapters 4 and 5 built the foundations to analyze and numerically solve PDE-constrained shape optimization problems under the discretize-then-optimize paradigm. In Schulz, 2014, it was suggested that formulating shape optimization problems on Riemannian manifolds offers great advantages. For this reason, we focused on the description of discrete shapes which belong to Riemannian manifolds. In this sense, chapter 4 was devoted to the presentation of discrete shape manifolds. In section 4.1 we provided a list of possible shape representations whose collection constitutes a Riemannian manifold. In order to formally define the manifold of planar triangular meshes, we used the language of simplicial complexes; consequently, section 4.2 collected the fundamentals of simplicial complexes both from the abstract and the geometric perspective. Moreover, we gathered a list of inequalities involving the geometric measurements of triangles used in this thesis. Section 4.3 can be considered as the first main contribution of this thesis, where we formally constructed the manifold of planar triangular meshes as the set of all oriented meshes, which can be obtained

through continuous deformations of a reference oriented mesh. We have proved this set is indeed an open submanifold of $\mathbb{R}^{2\times N_V}$, and endowed it with the standard smooth structure. This allowed us to conclude, that the manifold of planar triangular meshes $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is a smooth manifold, which is connected by definition. Additionally, we characterized its tangent space.

Chapter 5 presented our second main contribution: the construction of complete metrics for the manifold of planar triangular meshes. As preparation, we present two Riemannian metrics for the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ in section 5.1. First, we introduced the Euclidean metric, defined as the restriction of the metric from $\mathbb{R}^{2\times N_V}$ to $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. Second, inspired by Schulz, Siebenborn, Welker, 2016, we used the Lamé system of the linear elasticity to define a metric in $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. We provided in theorem 5.1.1 a proof that it is, indeed, a Riemannian metric for $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. Using Gordon, 1973, Thm. 1, we proposed two complete metrics for $\mathcal{M}_+(\Delta; Q_{\text{ref}})$. The first one was described in section 5.2 and depends on a function f_1 defined in (5.6). This function penalizes any direction pending to self-intersection on a mesh. One of the most important results of this chapter is the proof of properness of the function f_1 given in theorem 5.2.6. We also proved that this function enjoys invariance properties for rigid body motions. However, the values of this function get arbitrary large for fine meshes, even when there is no cell close to degeneracy. For this reason, we proposed a second complete metric in section 5.3, based on a proper function f_2 given in (5.22), which inherits all the properties of the first one, and additionally is invariant under uniform mesh refinements. As a drawback, we obtained geodesic equations that can only be integrated numerically, and we used their Hamiltonian formulation and the Störmer–Verlet scheme for this purpose, which we described in section 5.4. To improve our understanding of the manifold of planar triangular meshes, we presented three numerical experiments in section 5.5. We first investigated how meshes deform under elementary transformations (translations, scaling, shearing, and rotations). The second experiment allowed us to verify the completeness of the proposed metrics. The last experiment showed that the meshes along the geodesics associated with the complete metric keep their aspect ratios around acceptable values.

Chapter 6 addressed our primary concern: the study of discretized PDE-constrained shape optimization problems posed on the manifold of planar triangular meshes. To this end, we focused on a simple problem, taken from Etling et al., 2020, which in its continuous version has at least one globally optimal solution, and it is shape differentiable. After discretization, the node positions are the optimization variables, and in section 6.1, we provided numerical evidence that, in general, this kind of problems possesses no solution which belongs to the manifold of planar triangular meshes, even when the shape functional is bounded below. The example exhibited that the optimizer exploits the poor approximation state variable, as a result of the poor quality of the mesh, to further decrease the shape functional values. On the other hand, the lack of existence of solutions belonging to the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is somewhat reminiscent of a class of ill-posed inverse problems, which can be dealt by, e. g., the addition of an appropriate penalty function to the shape functional. Inspired by these facts, we proposed a novel penalty function in section 6.2, whose main purpose is to control the mesh quality during the optimization process. Since the complete metrics proposed in chapter 5 are based on functions that penalize all the possible directions pending to self-intersections, the augmentation functions f_1^μ and f_2^μ were our first candidates. However, the second function f_2^μ, given in (5.22), was a better fit since, similarly to the shape functional, it is also invariant under uniform mesh refinements. Moreover, we devised a result of existence of solutions for an abstract optimization problem in metric spaces (cf., proposition 6.2.1), which allowed us to exploit the already

known properness of the function f_2^μ. Thus, obtaining the existence of at least one globally optimal solution for a generic penalized, discretized PDE-constrained shape optimization problem, in corollary 6.2.2. Up to the author's knowledge this is the fist result of this kind. Furthermore, the existence of solutions for our example problem was presented in proposition 6.2.3, assuming the existence of a hold-all domain. We also derived the first-order optimality conditions of the problem. The Riemannian steepest descent method was chosen to numerically solve the discretized problems and was described in section 6.3. Thanks to the versatility of this method, we conceived four different variants depending on the chosen Riemannian metrics. We termed the variants Euclidean-Euclidean, Elasticity-Euclidean, Complete-Euclidean, and Complete-Complete. The first component of the name referred to the metric used to evaluate the shape gradient, while the second component referred to the choice of retraction. Under these considerations, we conducted three numerical experiments in section 6.4. First, we considered the unpenalized problem with the variants Euclidean-Euclidean, Elasticity-Euclidean, and Complete-Complete. Keeping in mind that this problem may not have a solution, we studied the behavior of the different variants for long runs. We fixed 1000 as the maximum number of iterations and observed the following. The Euclidean-Euclidean failed at iteration 60 since the only acceptable step length at this iteration was lower than 10^{-6}. The values of the shape functional for variants Elasticity-Euclidean and Complete-Complete never reached a plateau. Moreover, the quality of the mesh at iteration 1000 was undesirable, supporting our claim that this problem may not possess a solution that belongs to the manifold of planar triangular meshes. However, it is worth highlighting that variant Complete-Complete reached an acceptable value of the shape functional with an acceptable mesh quality in only 15 iterations, compared with the 150 iterations required by the Elasticity-Euclidean variant. Although we put some effort into an efficient implementation, the solution of the geodesic equation remains computationally involved, and is without any doubt, the most expensive step in a shape optimization loop. As a compromise, in the last two experiments we studied how the variant Complete-Euclidean behaves compared to the variants Euclidean-Euclidean and Elasticity-Euclidean. Solving the penalized problem presented significant advantages like using standard stopping criteria, since now, the existence of solutions of the problem is guaranteed by virtue of proposition 6.2.3. Moreover, the numerical experiments revealed that even using the variant Euclidean-Euclidean can lead to acceptable solutions. However, in most cases, the variant Complete-Euclidean generated meshes with better quality than the other variants. Finally, one could argue about the rather arbitrary choice of the penalization parameters. Therefore, in the last experiment, we revisited the unpenalized problem and studied the behavior of the most promising variants, namely the variants Elasticity-Euclidean and Complete-Euclidean for finer meshes. We let the algorithm iterate 500 times for each variant. A close up to the final iterate of the variant Complete-Euclidean revealed the natural redistribution of the nodes –as equilateral as possible– promoted by using the complete metric.

Summarizing, in this thesis we made the following contributions:

- Using the language of simplicial complexes, we characterized the set of all node positions which generate admissible meshes with connectivity Δ, denoted by $\mathcal{M}_0(\Delta)$, and proved that it is a smooth submanifold of $\mathbb{R}^{2\times N_V}$.
- Using the orientation of simplicial complexes, we defined the set of admissible oriented meshes with connectivity Δ denoted by $\mathcal{M}_+(\Delta)$ and proved that it is also a smooth submanifold of $\mathbb{R}^{2\times N_V}$.
- We proved in theorem 4.3.11 that the set $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ defined in definition 4.3.10, and termed the manifold of planar triangular meshes is a connected, smooth, submanifold of $\mathbb{R}^{2\times N_V}$.

- By considering the discretization of the Lamé system of linear elasticity with piecewise linear elements over the mesh defined by $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$, we proved in theorem 5.1.1 it coincides with the matrix representation of a Riemannian metric for $\mathcal{M}_+(\Delta; Q_{\text{ref}})$.
- We proposed two functions f_1 given in (5.6) and f_2 given in (5.22) which penalize all the possible directions pending to self-intersection of a mesh. The proofs of properness were given in theorems 5.2.6 and 5.3.3, respectively.
- We constructed a family of $\mathcal{C}^3$-regularizations f_1^μ and f_2^μ for f_1 and f_2, respectively, which are proper and also approximate arbitrary good the original ones when $\mu \to +\infty$. See appendix B for more details.
- Using Gordon, 1973, Thm. 1, we constructed two complete metrics for the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ in theorem 5.2.9 and proposition 5.3.5 associated with the functions f_1^μ and f_2^μ.
- Using the Störmer–Verlet scheme, we numerically solved the geodesic equations associated with the complete metrics.
- We designed numerical experiments in which we verify the completeness of the metrics compared with the Euclidean metric.
- We provided numerical evidence that discretized, PDE-constrained shape optimization problems, in general, do not possess a solution that belongs to the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$, even if the shape functional is bounded below.
- We proposed a penalized version of the discretized, PDE-constrained shape optimization problem, for which we guarantee the existence of solutions in corollary 6.2.2.
- The proposed penalization coincides with the proper function f_2^μ and therefore served two purposes. First, render well-posed problems in the sense that they have at least one globally optimal solution. Secondly, it can be used to construct a complete metric as suggested in proposition 5.3.5.
- Four variants of the Riemannian steepest descent method were considered.
- Solving the penalized problem showed that the obtained results are acceptable even using the variant Euclidean-Euclidean. However, the quality of the obtained meshes is better by using of the proposed complete metric.
- For the unpenalized problem early stopping is mandatory. Moreover, following the geodesics associated with the complete metric can undoubtedly accelerate the optimization process in terms of the number of iterations, but not necessarily with respect to the times of execution. In the absence of the complete metric to update the mesh, the results are still favorable since the complete metrics naturally redistribute the nodes as equilateral as possible.

Finally, we want to mention possible future research lines.

Topological properties of $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ and a limit manifold: In this thesis, we have studied the manifold of planar triangular meshes from a rather applied perspective. However, there are still open questions regarding the topological properties of this set. It will be interesting to understand if the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ is bounded or if one can characterize its boundary. Furthermore, while considering uniform refinements of the mesh, how do the associated connectivity complexes relate?. Is there a relationship between their corresponding manifolds of planar triangular meshes? All these considerations could be used to build an infinite-dimensional manifold as the limit of the manifold $\mathcal{M}_+(\Delta; Q_{\text{ref}})$ when $N_V \to +\infty$ and therefore find continuous interpretation to our augmentation functions and corresponding

penalty term. One could use this continuous version of the penalization to render problems with at least one optimal solution also in the continuous case.

Linear elasticity Riemannian metric: Having proved the bilinear form associated with the linear elasticity is indeed a Riemannian metric opens the natural question of studying the associated geodesic equation. Could we also use the Störmer–Verlet scheme to approximate their solutions? Moreover, by virtue of Gordon, 1973, Thm. 2, one could, in principle, try to construct a proper function f such that the matrix representation of the metric with respect to the vec chart $(g_{ij} - f_i f_j)$ is positive definite. If so, the metric will also be complete, and then one could search for advantages on the approximation of the geodesics associated with this metric by exploiting previously proposed techniques for the solution of the linear elasticity equation.

Higher-order Riemannian line search methods: A natural extension of the work presented in this thesis is the formulation of higher-order Riemannian line search methods, like L-BFGS or Newton methods. These algorithms will require the efficient approximation of the parallel transport and covariant derivatives associated with the proposed complete metrics.

Mesh morphing, registration, interpolation: It was proved in Alexa, Cohen-Or, Levin, 2000; Alexa, 2002; Baghaie, Yu, D'souza, 2014 that working with triangular meshes is an advantage while performing image registration or image morphing. Using the notions presented in this thesis, these problems can be solved by means of the logarithmic map associated with the proposed complete metrics on the manifold of planar triangular meshes. Therefore, it will be interesting to propose iterative algorithms to approximate the logarithmic map, which do not involve the computation of the entire geodesic to update the iterates as the shooting method suggests.

Proposal of mesh-preserving retractions: The solution of the geodesic equation is, without any doubt, the most expensive step in a shape optimization loop; despite our efforts of an efficient implementation. For this reason, replacing the exponential map with retractions, i. e., its first-order approximations, becomes of paramount importance. The design of quality preserving retractions could significantly reduce the execution times of the algorithms while keeping the advantage of performing large deformations on the meshes without compromising their aspect ratios. In this way, one could consider more realistic applications.

Manifold of tetrahedral meshes: The theory developed in this thesis is currently limited to planar shapes. Unfortunately, there is not an easy way to generalize it to higher dimensions. For this reason, it is interesting the study of complete Riemannian metrics on higher dimensional shape spaces. The first natural extension is the proposal of a complete metric, using Gordon, 1973, Thm. 1, for the space of discrete shells introduced in Heeren et al., 2012. Furthermore, it will be interesting to pursue the proposal of the manifold of tetrahedral meshes and endow it with a corresponding complete Riemannian metric based on a mesh preserving proper function. The function proposed in this thesis will not be enough but can be a good start. The heights of the 2-faces and the distance from boundary faces can be combined with dihedral angles, and the heights of the tetrahedra to obtain a proper function for this manifold.

A Appendix: Fundamentals on Differential Geometry

The main aim of this chapter is to collect the most important notions of differential geometry used along this thesis. The chapter is structured as follows, we will start with the notion of topological manifolds, and endow them with a smooth structure in appendix A.1 In appendix A.2, we revisit the notions of Riemannian metric, and define the Riemannian manifolds together with the notion of geodesics. Then, we introduce concept of geodesically completeness which is key in this thesis, and present their main implications. We end this chapter by describing the Riemannian steepest descent method on manifolds in appendix A.3. The contents of this chapter are based on the following bibliography Lee, 2018; do Carmo, 1992; Herzog, 2018; Boumal, 2020.

A.1 Smooth Manifolds

We start by recalling the notion of topological manifolds. Let us consider $\mathcal{M}$ a topological space, we say $\mathcal{M}$ is a d-dimensional **topological manifold** if $\mathcal{M}$ is Hausdorff, second-countable and is locally homeomorphic to an open subset of $\mathbb{R}^d$. In other words, every point $q \in \mathcal{M}$ has an open neighborhood U which is homeomorphic to an open subset of $\mathbb{R}^d$. Figure A.1 shows an illustration of a set being locally homeomorphic to an open subset of $\mathbb{R}^d$.

Given a d-manifold $\mathcal{M}$, a pair (U, ϕ) is said to be a **chart (coordinate)** if $U \subset \mathcal{M}$ is open and $\phi\colon U \to \phi(U)$ is a homeomorphism onto the open subset $\phi(U) \subset \mathbb{R}^d$. In other words, $\phi\colon \mathcal{M} \supset U \to \phi(U) \subset \mathbb{R}^d$ is continuous with continuous inverse. A **chart about a point** $q \in \mathcal{M}$ is a chart (U, ϕ) such that $q \in U$. A collection of charts $\{(U_\alpha, \phi_\alpha)\}_{\alpha \in A}$ is called an **atlas** for $\mathcal{M}$ if the chart domain covers $\mathcal{M}$, i. e., $\mathcal{M} = \bigcup_{\alpha \in A} U_\alpha$.

Since we have defined the manifold of planar triangular meshes, as a open submanifold of $\mathbb{R}^{2\times N_V}$. We present the result which guarantees that an open subset of a manifold is also a manifold, whose proof can be found in Lee, 2011, Prop. 2.53, p. 39.

Theorem A.1.1. *Suppose that $\mathcal{M}$ is a topological manifold of dimension d and $\mathcal{N} \subset \mathcal{M}$ is open.*

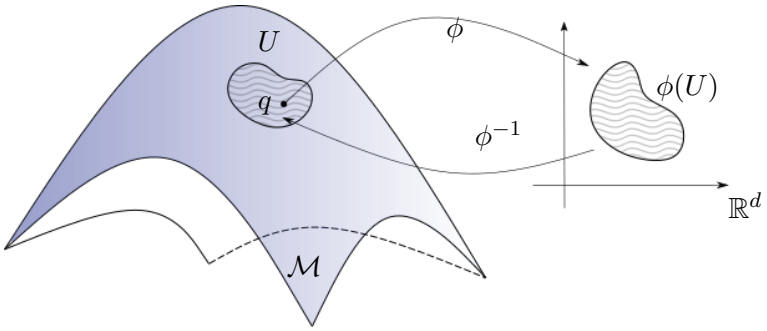


Figure A.1. Illustration of $\mathcal{M}$ being locally homeomorphic to on open subset of $\mathbb{R}^2$.

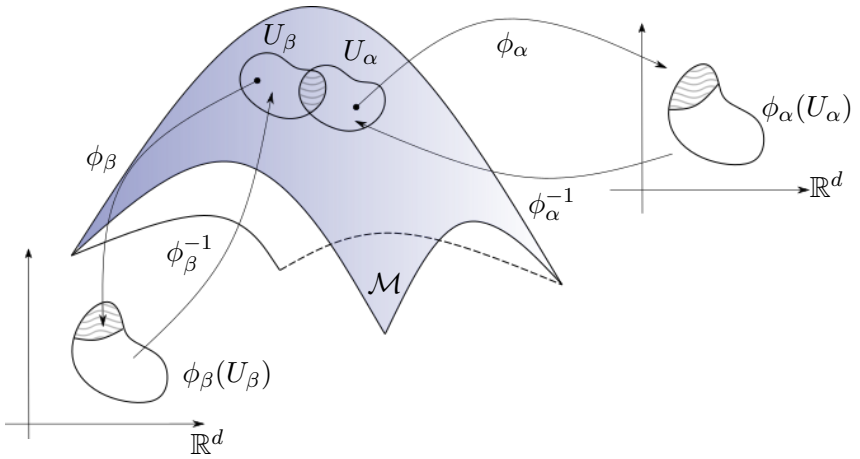


Figure A.2. Illustration of a transition map.

(a) *Let us endow $\mathcal{N}$ with the subspace topology, (open sets in $\mathcal{N}$ are defined as the intersection of $\mathcal{N}$ with the open subsets of $\mathcal{M}$). Then, $\mathcal{N}$ is a topological manifold of dimension d and we refer to it as an open submanifold of $\mathcal{M}$.*
(b) *If $\{(U_\alpha, \phi_\alpha)\}_{\alpha\in A}$ is an atlas of $\mathcal{M}$, then $\{(\mathcal{N}\cap U_\alpha, \phi_\alpha)\}_{\alpha\in A}$ is an atlas for $\mathcal{N}$.*

Since continuity is a merely topological property, one can also define continuous functions between topological manifolds. Suppose that $\mathcal{M}, \mathcal{N}$ are topological manifolds of dimension $d_\mathcal{M}$ and $d_\mathcal{N}$, respectively. Moreover, let us denote as $\{(U_\alpha, \phi_\alpha)\}_{\alpha\in A}$ and $\{(V_\beta, \psi_\beta)\}_{\beta\in B}$, atlases of $\mathcal{M}$ and $\mathcal{N}$, respectively. A function $F\colon \mathcal{M}\to\mathcal{N}$ is said to be **chart continuous** if the function

$$\psi_\beta \circ F \circ \phi_\alpha^{-1}\colon \mathbb{R}^{d_\mathcal{M}} \supset \phi_\alpha(U_\alpha \cap F^{-1}(V_\beta)) \to \psi_\beta(V_\beta) \subset \mathbb{R}^{d_\mathcal{N}} \tag{A.1}$$

is continuous, for all charts (U_α, ϕ_α) and (V_β, ψ_β).

Now, we endow a topological manifold, with extra structure, namely, a smooth structure. To this end, we study transition maps. Let $\mathcal{M}$ be a d-dimensional topological manifold. For any two charts (U_α, ϕ_α) and (U_β, ϕ_β) of $\mathcal{M}$ the map given by

$$\phi_\alpha \circ \phi_\beta^{-1}\colon \mathbb{R}^d \supset \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta) \subset \mathbb{R}^d \tag{A.2}$$

is called a **transition map** (see figure A.2 for an illustration). Moreover, two charts (U_α, ϕ_α) and (U_β, ϕ_β) of $\mathcal{M}$ are said to be **smoothly compatible** if both of the transition maps, i. e., $\phi_\alpha \circ \phi_\beta^{-1}$ and $\phi_\beta \circ \phi_\alpha^{-1}$ are smooth on $\phi_\beta(U_\alpha \cap U_\beta)$, and $\phi_\alpha(U_\alpha \cap U_\beta)$, respectively. Finally, an atlas $\{(U_\alpha, \phi_\alpha)\}_{\alpha\in A}$ for $\mathcal{M}$ is said to by **smooth** if all its transition maps are smooth.

In general, there are many possible choices of atlases which represent the same smooth structure, in the sense that they all determine the same collection of smooth functions on the manifold $\mathcal{M}$. One possibility is to consider equivalence classes of smooth atlases, and work with representatives of this equivalence classes. A second option, is to consider maximal atlases. A smooth atlas $\mathcal{A}$ of $\mathcal{M}$ is said to be **maximal** if it is not contained in any strictly larger smooth atlas. A smooth structure on $\mathcal{M}$ is a maximal smooth atlas. Altogether, allows us to define a **smooth manifold** as a topological manifold which possess a smooth structure.

The next remark defines the standard smooth structure of $\mathbb{R}^d$.

Remark A.1.2. *The space $\mathbb{R}^d$ is a smooth d-dimensional manifold with the smooth structure determined by the atlas consisting on the single chart $(\mathbb{R}^d, \mathrm{id})$. This is usually called the* standard smooth structure.

The next remark justifies our claim on remark 4.3.13.

Remark A.1.3. *If $\mathcal{M}$ is a d-dimensional smooth manifold with $\{(U_\alpha, \phi_\alpha)\}_{\alpha\in A}$ a smooth atlas, and $\mathcal{N} \subset \mathcal{M}$ is an open subset. Then, $\mathcal{N}$ has a natural smooth structure consisting of all smooth charts $\{(U_\alpha \cap \mathcal{N}, \phi_\alpha)\}_{\alpha\in A}$. By virtue of theorem A.1.1, and the previously defined natural smooth structure, it holds that every open subset of a d-dimensional smooth manifold is a d-dimensional smooth manifold in a natural way.*

Now, we study the notion of smooth functions between smooth manifolds. Let $\mathcal{M}$ and $\mathcal{N}$ be smooth manifolds of dimension $d_\mathcal{M}$ and $d_\mathcal{N}$, respectively. Let us denote by $\mathcal{A} = \{(U_\alpha, \phi_\alpha)\}_{\alpha\in A}$ and $\mathcal{B} = \{(V_\beta, \psi_\beta)\}_{\beta\in B}$ the smooth atlases for $\mathcal{M}$ and $\mathcal{N}$, respectively. A function $\mathcal{F}\colon \mathcal{M} \to \mathcal{N}$ is said to be smooth if the mapping

$$\psi_\beta \circ \mathcal{F} \circ \phi_\alpha^{-1}\colon \mathbb{R}^{d_\mathcal{M}} \supset \psi_\beta(U_\alpha \cap \mathcal{F}^{-1}(V_\beta)) \to \psi_\beta(V_\beta) \subset \mathbb{R}^{d_\mathcal{N}}$$

is smooth for all ϕ_α and ψ_β.

Another key concept in the theory of smooth manifolds, and which is of extreme importance for optimization, is the notion of tangent vectors. Their main purpose is to formalize the notion of linear approximations near a point. There are various equivalent definitions, in what follows we mention two of them. For every point $q \in \mathcal{M}$, a **tangent vector** (in the algebraic sense) **at a point** $q \in \mathcal{M}$ is a linear map $v\colon C^\infty(\mathcal{M}) \to \mathbb{R}$ which is a derivation at q, i. e.,

$$v(fg) = f(q)vg + g(q)vf \qquad \text{for all } \ f, g \in \mathcal{C}^\infty(\mathcal{M}).$$

This relation is also known as the product or Leibniz rule.

Secondly, a **tangent vector** (in the geometric sense) **at a point** $q \in \mathcal{M}$ is an equivalence class of $\mathcal{C}^1$-curves defined on an open interval around zero, which satisfies the following relation: If $\{(U_\alpha, \phi_\alpha)\}_{\alpha\in A}$ is the atlas of $\mathcal{M}$, then, two differentiable curves c_1 , $c_2\colon \mathbb{R} \to \mathcal{M}$ such that $c_1(0) = c_2(0) = q$ are said to be equivalent if the following relation holds

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_\alpha(c_1(t))\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\phi_\alpha(c_2(t))\bigg|_{t=0},$$

for all charts (U_α, ϕ_α) about q.

Regardless of the chosen definition, the collection of all tangent vectors at a point $q \in \mathcal{M}$ is denoted by $\mathcal{T}_q\mathcal{M}$, and called the **tangent space at** q. Assuming appropriate definitions of the addition and multiplication with a scalar of curves, it is possible to prove that $\mathcal{T}_q\mathcal{M}$ is a vector space, and its dimension coincides with the dimension of the manifold. See, e. g., do Carmo, 1992, Ch. 0, p. 8. This proof is based on the coordinate functions, which we describe in what follows.

Let us consider the d-dimensional smooth manifold $(\mathcal{M}, \mathcal{A})$ such that $\mathcal{A} = \{(U_\alpha, \phi_\alpha)\}_{\alpha\in A}$. The a-th **coordinate function** x^a induced by the chart (U_α, ϕ_α) is defined by

$$x^a\colon \mathcal{M} \supset U_\alpha \ni q \mapsto x^a(q) \coloneqq [\phi_\alpha(q)]^a \in \mathbb{R}.$$

In other words, x^a assigns to a point $q \in \mathcal{M}$ the a-th coordinate of its image under ϕ_α. There exists d (dimension of the manifold) coordinate functions, one for each component of $\phi(q)$. The smoothness of these functions is immediately obtained by the chain rule and the fact that each chart ϕ_α is also a smooth function.

Using the notion of coordinate functions; now, we consider the so-called coordinate vectors. Let ϕ be a fixed but arbitrary chart around $q \in \mathcal{M}$, and $(x^1, \ldots, x^d)$ be the coordinate functions of ϕ. The **coordinate vectors** denote by $\partial/\partial x^1|_q, \ldots, \partial/\partial x^d|_q$ are defined through the following expression

$$\frac{\partial}{\partial x^a} f = \frac{\partial}{\partial x^a}\bigg|_{\phi(q)} (f \circ \phi^{-1}).$$

It can be proved that these vectors form a basis for the space $\mathcal{T}_q\mathcal{M}$, and therefore, once the chart has been fixed, every tangent vector $v \in \mathcal{T}_q\mathcal{M}$ can be written uniquely as follows:

$$v = \sum_{a=1}^{d} v(x^a) \left.\frac{\partial}{\partial x^a}\right|_q .$$

Notice that $v(x^a) \in \mathbb{R}$ for all $a = 1, \dots, d$ and we refer to them as the components of v w.r.t. the basis $\{\partial/\partial x^a|_q\}_{a=1}^d$.

Now, we present the result which characterizes the tangent space of an open submanifold. We refer the reader to Lee, 2012, Prop. 3.8 and Prop. 3.9 for details on the precise function which allows us to identify the tangent space of an open submanifold with the tangent space of the manifold.

Proposition A.1.4. *Let $\mathcal{M}$ be a d-dimensional smooth manifold and let $\mathcal{N}$ be an open smooth submanifold in the sense of remark A.1.3. Then, for any $q \in \mathcal{N}$, we have $\mathcal{T}_q\mathcal{N} \cong \mathcal{T}_q\mathcal{M}$.*

The differential of a smooth function at a certain point $q \in \mathcal{M}$, can also be called the push-forward. For the smooth function $j\colon \mathcal{M} \to \mathbb{R}$, and $q \in \mathcal{M}$, we consider the linear map

$$\begin{aligned} \mathrm{d}j_q\colon \mathcal{T}_q\mathcal{M} &\to \mathcal{T}_{j(q)}\mathbb{R} \cong \mathbb{R} \\ v &\mapsto \mathrm{d}j_q v \coloneqq v(\cdot \circ j) \end{aligned} \tag{A.3}$$

and we call $\mathrm{d}j_q$ the **differential** of j at q.

Alongside the tangent space, we introduce of the contangent space to $\mathcal{M}$ at q as the dual $(\mathcal{T}_q\mathcal{M})'$ of the tangent space $\mathcal{T}_q\mathcal{M}$ and it is denoted as $\mathcal{T}_q^*\mathcal{M}$. The elements of $\mathcal{T}_q^*\mathcal{M}$ are called cotangent vectors to $\mathcal{M}$ at q.

A.2 Riemannian Manifolds

As already mentioned, the main purpose of a Riemannian metric is to extend geometric notions such as vectors lengths, angles between vectors, among others, to smooth manifolds.

Definition A.2.1. *Let $\mathcal{M}$ be a d-dimensional smooth manifold. A Riemannian metric on $\mathcal{M}$ is a correspondence which associates to each point $q \in \mathcal{M}$ an inner product (symmetric, bilinear and positive-definite form) $(\cdot\,,\,\cdot)_q\colon \mathcal{T}_q\mathcal{M} \times \mathcal{T}_q\mathcal{M} \to \mathbb{R}$ which varies smoothly from point to point in the following sense. If (U, ϕ) is a chart about q with coordinate functions $x^1, \dots, x^d$ and coordinate vectors $\partial/\partial x^1|_q, \dots, \partial/\partial x^d|_q$, then the functions*

$$g_{ab}(x^1, \dots, x^d) = \left(\left.\frac{\partial}{\partial x^a}\right|_q , \left.\frac{\partial}{\partial x^b}\right|_q \right)_q \tag{A.4}$$

are differentiable on U. Moreover, each function g_{ab} is called the local representation of the Riemannian metric in the coordinate system $(x^1, \dots, x^d)$.

Definition A.2.1 allows us to introduce the notion of Riemannian manifold. Suppose that $\mathcal{M}$ is a smooth manifold. Then, $\mathcal{M}$ together with a Riemannian metric is called a **Riemannian manifold**. On a Riemannian manifold one can compute, for example, the **norm or length** of a tangent vector $v \in \mathcal{T}_q\mathcal{M}$ as follows:

$$\|v\|_q = \sqrt{(v\,,\,v)_q}. \tag{A.5}$$

In what follows we detail the local representation of the Euclidean metric on $\mathbb{R}^d$.

Remark A.2.2. *The Euclidean metric can be represented in standard coordinates as follows. Let us consider $v, \widetilde{v} \in \mathcal{T}_q\mathcal{M}$ which can be written as $v = \sum_a v(x^a)\partial/\partial x^a|_q$ and*

$\widetilde{v} = \sum_b \widetilde{v}(x^b)\partial/\partial x^b|_q$, *then*

$$(v\,,\,\widetilde{v})_q^{\mathrm{Euc}} = \sum_a v(x^a)\,\widetilde{v}(x^a) = \sum_{a,b} \delta_a^b v(x^a)\,\widetilde{v}(x^b),$$

therefore $\tilde{g}_{ab} = \delta_a^b$.

One of the most important applications of the Riemannian metric in the context of optimization is the ability of converting the derivative $(\mathrm{d}j_q) \in \mathcal{T}_q^*\mathcal{M}$ of a function $j\colon \mathcal{M} \to \mathbb{R}$ into a tangent vector $v \in \mathcal{T}_q\mathcal{M}$.

Definition A.2.3. *Let $j\colon \mathcal{M} \to \mathbb{R}$ be a smooth function on a Riemannian manifold $\mathcal{M}$, and $q \in \mathcal{M}$. The Riemannian gradient of j with respect to the metric g at the point q is the tangent vector* $\operatorname{grad} j(q) \in \mathcal{T}_q\mathcal{M}$ *which satisfies*

$$(\operatorname{grad} j(q)\,,\,v)_q = (\mathrm{d}j_q)(v), \tag{A.6}$$

for all $v \in \mathcal{T}_q\mathcal{M}$.

It is easy to see that, as in the Euclidean case, $\operatorname{grad} j$ is the steepest-ascent direction of j at q (see e. g., Absil, Mahony, Sepulchre, 2008, Ch. 3, p. 46).

Now, we focus on the notion of geodesics on a manifold. Let (U, ϕ) be a chart of $\mathcal{M}$ with coordinate functions x^a, $a = 1, \ldots, d$ and let $\gamma^a \coloneqq x^a \circ \gamma$ denote the coordinates of a curve γ. Then, γ is a **geodesic** if and only if its coordinate curves solve the following system of second-order nonlinear ordinary differential equations

$$\frac{\mathrm{d}^2\gamma^c}{\mathrm{d}t^2} + \sum_{a,b=1}^{d} \Gamma_{ab}^c \frac{\mathrm{d}\gamma^a}{\mathrm{d}t}\frac{\mathrm{d}\gamma^b}{\mathrm{d}t} = 0, \quad c = 1, \ldots, n,$$

where Γ_{ab}^c are evaluated at $\gamma(t)$. We refer to Γ_{ab}^c as the **Christoffel symbols**, and they are defined by the following expression:

$$\Gamma_{ab}^c = \frac{1}{2}\sum_{e=1}^{d} g^{ce}\left(\frac{\partial g_{ea}}{\partial x^b} + \frac{\partial g_{eb}}{\partial x^a} - \frac{\partial g_{ab}}{\partial x^e}\right). \tag{A.7}$$

The components of the matrix representation of the metric with respect to ϕ are denoted by g_{ab} and g^{ab} are the components of its corresponding inverse.

For a given initial point $q \in \mathcal{M}$ and an initial tangent vector $v \in \mathcal{T}_q\mathcal{M}$, one can prove that there exists an open interval $I \subset \mathbb{R}$ containing zero, such that the geodesic equation has a solution on I and it is unique; see e. g., Lee, 2018, Thm. 4.27, p. 103. The unique solution of (5.35) with initial conditions $\gamma(0) = q \in \mathcal{M}$ and $\dot{\gamma}(0) = v \in \mathcal{T}_q\mathcal{M}$ is denoted by $\gamma_{q,v}(t)$.

In this context, we also introduce the notion of maximal geodesic. A geodesic $\gamma\colon I \to \mathcal{M}$ is said to be maximal if does not exist a geodesic $\tilde{\gamma}\colon \tilde{I} \to \mathcal{M}$ defined on an interval $\tilde{I}$ properly containing I and satisfying $\tilde{\gamma}|_I = \gamma$. In other words, it cannot be extended to a geodesic on a larger interval. It can also be proved that for every Riemannian manifold and for each $q \in \mathcal{M}$, and $v \in \mathcal{T}_q\mathcal{M}$ there exists a unique maximal geodesic (cf., Lee, 2018, Cor. 4.28, p. 105).

In example A.2.4, we show that indeed the geodesics associated to the Euclidean metric on $\mathbb{R}^d$ are straight lines.

Example A.2.4. *Let us consider $\mathbb{R}^d$ endowed with the standard smooth structure, as given in remark A.1.2. Moreover, let us endow $\mathbb{R}^d$ with the Euclidean metric, whose matrix representation in this chart is given by* $\mathrm{id}_{d\times d}$ *(the identity matrix of dimension d) as suggested by remark A.2.2. Let us denote by $q = \gamma(0)$, and $v = \dot{\gamma}(0)$. In this case, the Christoffel symbols given in (A.7), are $\Gamma_{ab}^c = 0$ for all $a, b, c = 1, \ldots, d$. Then, it follows, the geodesic*

satisfies $\mathrm{d}^2\gamma^c/\,\mathrm{d}t^2 = 0$ *for all* $c = 1, \ldots, d$. *Solving this second order, linear, homogeneous and decoupled system of ordinary differential equations, gives as* $\gamma_{q,v}(t) = q + tv$, *for* $t \in \mathbb{R}$.

Another important property worth to be highlighted is that geodesics with proportional initial tangent vectors are related. This property was exploited in the optimization algorithm described in algorithm 2 to reduce the computation costs of the Armijo backtracking. We refer the reader to subsection 6.4.1 for more details. This result is commonly known as the *rescaling lemma*, and its proof can be found in Lee, 2018, Lem. 5.18, p. 127.

Lemma A.2.5. *For every* $q \in \mathcal{M}$, $v \in \mathcal{T}_q\mathcal{M}$ *and* $c, t \in \mathbb{R}$

$$\gamma_{q,cv}(t) = \gamma_{q,v}(ct)$$

whenever either side is defined.

Up until now, we have been able to generalize the straight lines for a given point $q \in \mathcal{M}$ and a give $v \in \mathcal{T}_q\mathcal{M}$. To improve our understanding about geodesics, we are also interested in studying their behavior while varying the initial tangent vector. To this end we define, the **exponential map**, which for $q \in \mathcal{M}$ is given by the following expression:

$$v \mapsto \exp_q(v) \coloneqq \gamma_{q,v}(1).$$

One could also define the exponential map on the tangent bundle and give it a more general view. However, we do not purse that definition here. For more details, we refer the reader to Lee, 2018, Ch. 5, p. 126.

We end this section by introducing the notion of geodesically completeness, which is key in this thesis.

Definition A.2.6. *A Riemannian manifold* $\mathcal{M}$ *is said to be geodesically complete if every maximal geodesic is defined for all* $t \in \mathbb{R}$.

Even though the notion of geodesically completeness is associated directly to the manifold, it entirely depends on the choice of metric. In other words, we can transform a geodesically incomplete manifold into a geodesically complete one, only by endowing it with a different Riemannian metric. A simple example of this is the positive real line.

Example A.2.7. *Let us consider the smooth manifold* $\mathcal{M} \coloneqq \{x \in \mathbb{R} \,|\, x > 0\}$. *If we endow it with the Euclidean metric, it is easy to find a point* $q \in \mathcal{M}$ *and* $v \in \mathcal{T}_q\mathcal{M}$ *such that the value of the exponential map at a finite time escapes* $\mathcal{M}$. *However, if we endow it with the metric whose matrix representation with respect to the identity chart is given by the following expression:*

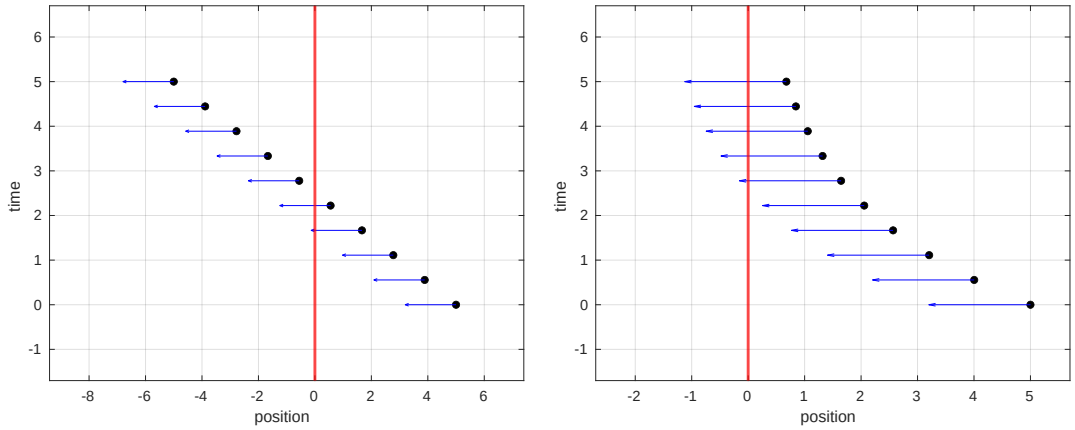
$$g_{11} = \frac{1}{q^2}.$$

Then, the resulting Riemannian manifold is geodesically complete. Figure A.3 we show ten snapshots of the geodesics associated with the Euclidean and complete metrics. For a proof of the completeness of this metric we refer the reader to e. g., Moakher, Zéraï, 2011, Thm. 3

Verifying if a metric is not complete, is relatively easy: for a given point $q \in \mathcal{M}$ and given tangent vector $v \in \mathcal{T}_q\mathcal{M}$, one needs to find a $\bar{t} \in \mathbb{R}$ for which the value of the geodesic $\gamma_{q,v}(\bar{t})$ at time $\bar{t}$ does not belong to $\mathcal{M}$. Conversely, constructing complete metrics is, in general, not an easy task. However, a recipe to construct complete Riemannian metrics on connected smooth manifolds such as $\mathcal{M}_+(\Delta; Q_{\mathrm{ref}})$ was presented in Gordon, 1973, whose main ingredient is a proper function.

Definition A.2.8. *A function* $f : \mathcal{M} \to \mathbb{R}$ *is said to be proper if the preimages* $f^{-1}(K)$ *of compact sets* $K \subset \mathbb{R}$ *are compact in* $\mathcal{M}$.

This allows us to present the theorem, in which we base the construction of complete metrics.



(a) Euclidean metric. (b) Complete metric.

Figure A.3. 10 snapshots of the geodesics for the positive real line with initial point $q = 5$ and initial tangent vector $v = -2$ on the interval $[0, 5]$.

Theorem A.2.9 (Gordon, 1973, Thm. 1). *Suppose that $\mathcal{M}$ is a connected manifold of class $\mathcal{C}^3$, endowed with a (not necessarily complete) Riemannian metric $\widetilde{g}$ with component functions $\widetilde{g}_{ab}$. If $f\colon \mathcal{M} \to \mathbb{R}$ is any proper function of class $\mathcal{C}^3$, then the Riemannian metric g defined by*

$$g_{ab} = \widetilde{g}_{ab} + \frac{\partial f}{\partial x^a}\frac{\partial f}{\partial x^b} \tag{A.8}$$

is geodesically complete.

Moreover, Gordon, 1973, Thm. 2 shows that this construction is the only way to obtain complete Riemannian metrics on connected smooth manifolds.

We end this section by introducing the important implications of a manifold being geodesically complete. Particularly, we are interested in the result which guarantees that a geodesically complete manifold is also complete in the sense of metric space, i. e., every Cauchy sequence converges. This result is known as the Hopf–Rinow theorem, whose proof can be found in Lee, 2011, Thm. 6.19, p. 169, and we state in what follows.

Theorem A.2.10 (Hopf-Rinow). *A connected Riemannian manifold is complete in the sense of a metric space if and only if it is geodesically complete.*

Corollaries A.2.11 to A.2.13 are a direct consequence of theorem A.2.10.

Corollary A.2.11. *If $\mathcal{M}$ is a connected Riemannian manifold and there exists a point $q \in \mathcal{M}$ such that the restricted exponential map $\exp_q$ is defined on all of $\mathcal{T}_q\mathcal{M}$, then $\mathcal{M}$ is complete.*

Corollary A.2.12. *If $\mathcal{M}$ is a complete, connected Riemannian manifold, then any two points in $\mathcal{M}$ can be joined by a minimizing geodesic segment.*

Corollary A.2.13. *If $\mathcal{M}$ is a compact Riemannian manifold, then every maximal geodesic in $\mathcal{M}$ is defined for all time.*

A.3 Riemannian Steepest Descent Method

This section aims to describe the simplest iterative optimization algorithm on manifolds, i. e., the Riemannian steepest descent method. The definition of local and global minima, and first-order optimality conditions presented in chapter 2 can be naturally generalized to manifolds. We refer the reader to Absil, Mahony, Sepulchre, 2008; Boumal, 2020 for more information on this and further optimization algorithms on manifolds.

We start by recalling the line search optimization methods on Euclidean spaces provided in Nocedal, Wright, 2006. These methods generate a sequence of improved estimates until they terminate, from an initial guess of the unknown. The guarantee that the algorithm had terminated in a solution of the problem is provided by the first-order optimality conditions. In order to update the estimates, these algorithms usually use the values of the objective function together with its derivatives of first and/or second-order. How long we move along the chosen direction is computed by approximating a one dimensional minimization problem where the *step length* is the unknown.

As mentioned, we focused on the simplest of the line search methods, namely the steepest descent method. In the Euclidean context, by considering the problem

$$\text{Minimize} \quad j(x) \quad \text{w.r.t. } x \in \mathbb{R}^n,$$

the standard steepest descent iteration is given by the following expression:

$$x_{n+1} = x_n - s_n \nabla j(x_n).$$

It can be interpreted as follows: the next iteration of the method is generated by following a straight line which starts at x_n, whose slope equals to $-\nabla j(x_n)$, and with step length s_n.

Let us now generalize these concepts to Riemannian manifolds. We consider the following generic formulation of an unconstrained problem on a Riemannian manifold $\mathcal{M}$.

$$\text{Minimize } j(q), \quad \text{w.r.t. } q \in \mathcal{M}.$$

By recalling that geodesics are the generalization of straight lines on manifolds, the standard iteration of the Riemannian steepest descent method is given by

$$q_{n+1} = \exp_{q_n}(-s_n \operatorname{grad} j(q_n)).$$

As mentioned, the steepest descent method is completely determined by choice of step length. Ideally one would like to solve the following problem:

$$\min_{s \in \mathbb{R}} j\big(\exp_{q_n}(-s \operatorname{grad} j(q_n))\big)$$

in each iteration. However, depending on the objective function, solving this problem can be as time consuming as the original one. Therefore, many techniques had been developed to approximate its solution, without requiring too much computational effort.

We work with the backtracking line search associated to the Armijo condition (in its Riemannian version) given by the following expression:

$$j\big(\exp_{q_n}(-s \operatorname{grad} j(q_n))\big) \leq j(q_n)) + \sigma s(\operatorname{grad} j(q_n)\,,\, \operatorname{grad} j(q_n))_q,$$

where $\sigma \in (0,1)$ usually takes the value of 10^{-4}.

The backtracking line search is then described along the lines of algorithm 3.

Algorithm 3: Armijo backtracking line search

Data: Set the parameters $\sigma \in (0,1)$, $\tau \in (0,1)$
Data: $q \in \mathcal{M}$, $\bar{s} > 0$
Result: step length with ensures the sufficient decrease condition is satisfied
set $s \leftarrow \bar{s}$
while $j\big(\exp_{q_n}(-s \operatorname{grad} j(q_n))\big) > j(q_n) + \sigma s(\operatorname{grad} j(q_n)\,,\, \operatorname{grad} j(q_n))_q$ **do**
  Set $s \leftarrow \tau s$;
end
return s

Altogether, allows us to present the steps which define the Riemannian steepest descent method on algorithm 4.

Algorithm 4: Riemannian steepest descent method

Data: $q_0 \in \mathcal{M}$, $\bar{s} > 0$
1 **for** $n = 0, 1, 2, \ldots$ **do**
2 Compute $\mathrm{d}j_{q_n}$;
3 Compute $\operatorname{grad} j(q_n)$ by solving (A.6);
4 Choose s_n with algorithm 3;
5 Update $q_{n+1} = \exp_{q_n}(-s_n \operatorname{grad} j(q_n))$;
6 Set $n \leftarrow n + 1$;
7 **end**
8 **return** q

Finally, we would like to remark that the evaluation of the exponential map in line 5 of algorithm 4 may result too computationally expensive (it involves the solution of the geodesic equation). To reduce the computational its computational costs, one could consider a retraction. Informally, it is a function which associates for each point $q \in \mathcal{M}$ and each initial velocity $v \in \mathcal{T}_q\mathcal{M}$, a curve completely contained in the manifold. In what follows we present its formal definition.

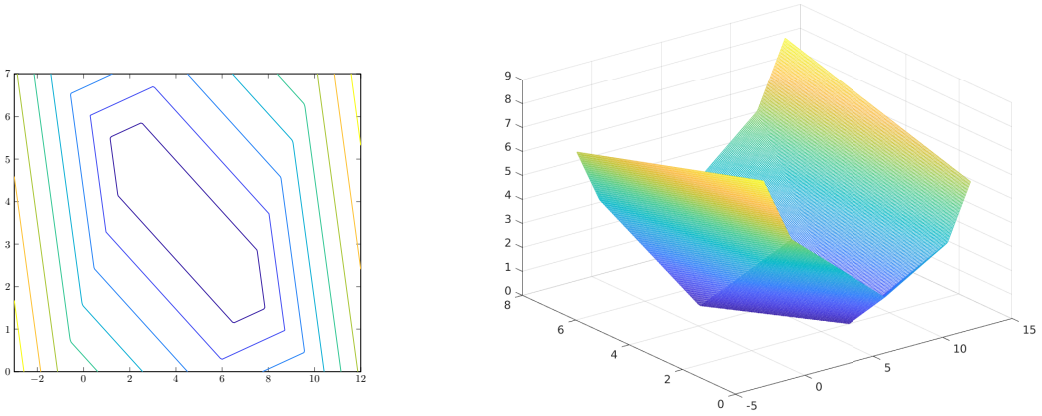
Definition A.3.1. *A retraction on a Riemannian manifold $\mathcal{M}$ is a smooth mapping $\mathcal{R}$ from the tangent bundle $\mathcal{TM}$ onto $\mathcal{M}$ with the following properties. Let $\mathcal{R}_q$ denote the restriction of $\mathcal{R}$ to $\mathcal{T}_q\mathcal{M}$ such that*

(a) $\mathcal{R}_q(0_q) = q$, where 0_q denotes the zero element of $\mathcal{T}_q\mathcal{M}$.
(b) $D\mathcal{R}_q(0_q) = \mathrm{id}_{\mathcal{T}_q\mathcal{M}}$, where $\mathrm{id}_{\mathcal{T}_q\mathcal{M}}$ denotes the identity mapping on $\mathcal{T}_q\mathcal{M}$.

We assume the canonical identification $\mathcal{T}_{0_q}\mathcal{T}_q\mathcal{M} \simeq \mathcal{T}_q\mathcal{M}$, and $D\mathcal{R}_q$ is the differential of $\mathcal{R}_q$.

B Appendix: An example regularization for $D_Q(i_0;[j_0,j_1])$

This chapter aims to present an example of a family of functions which can be used to approximate f_1^μ and f_2^μ as suggested in (5.16), and (5.22), respectively. To this end we use a regularized version of the 1- norm based distance from a vertex to an edge, $D_Q(i_0;[j_0,j_1])$ described in (4.13). In figure B.1 we depict $D_Q(i_0;[j_0,j_1])$ as a function of the vertex $\{q_{i_0}\}$, together with its contour plots, where we can clearly observe the lack of differentiability of this function. In appendix B.1 we present the construction of the $\mathcal{C}^3$-regularizations. Having guaranteed the regularity of the functions f_1^μ and f_2^μ, we will compute their first-order derivatives in appendix B.2.



(a) contour plot.

(b) $D_Q(i_0;[j_0,j_1])$

Figure B.1. Illustration of $D_Q(i_0;[j_0,j_1])$ when $q_{j_0}=[2,5]^{\mathrm{T}}$ and $q_{j_1}=[7,2]^{\mathrm{T}}$ as a function of the vertex $q_{i_0}\in[-3,12]\times[0,7]$.

We start by noticing that the 1- norm based distance of a vertex $\{q_{i_0}\}$ to an edge $\mathrm{conv}_Q(j_0,j_1)$ can be written as

$$D_Q(i_0;[j_0,j_1]) = g\big(\tilde{q}^1_{i_0};[\tilde{q}^1_{j_0},\tilde{q}^1_{j_1}]\big) + \big|\tilde{q}^2_{i_0}-\tilde{q}^2_{j_0}\big|. \tag{B.1}$$

Here $\tilde{q}^\ell$ stands for the first ($\ell=1$) or second ($\ell=2$) component of a vector q rotated about the origin so that its first coordinate aligns with the edge $\mathrm{conv}_Q(j_0,j_1)$, as shown in figure B.2.

For convenience, the convention here is that $\tilde{q}^1_{j_0}<\tilde{q}^1_{j_1}$ holds. Furthermore, the function g is the distance of a point to an interval in $\mathbb{R}$, i. e.,

$$g(x;[y,z]) = \begin{cases} |y-x| & \text{if } y\geq x,\\ 0 & \text{if } y\leq x\leq z,\\ |x-z| & \text{otherwise.}\end{cases} \tag{B.2}$$

which is depicted in figure B.3.

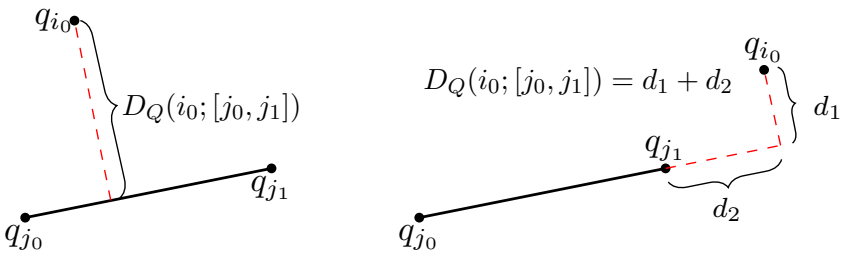


Figure B.2. Illustration of the distance (4.13) of a vertex to an edge in an edge oriented coordinate system. The two cases shown are when the projection of the vertex onto the infinite line generated by the edge belongs to the edge (left), and when it does not (right).

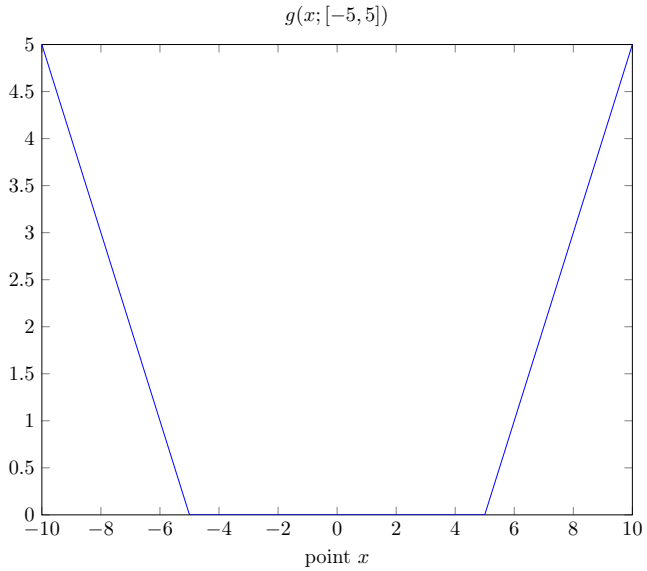


Figure B.3. Example of the distance of a point $x \in \mathbb{R}$ to the interval $[-5,5]$.

B.1 Construction of $\mathcal{C}^3$–Regularizations

We construct a regularizing function $D_Q^\mu(i_0;[j_0,j_1])$ based on $\mathcal{C}^3$-regularizations of the function g given in (B.2) and the absolute value.

Definition B.1.1. *Suppose that $\mu \geq 1$. We define the* ***regularized*** *1-**norm based distance from a vertex to an edge** as follows:*

$$D_Q^\mu(i_0;[j_0,j_1]) = g^\mu\big(\tilde{q}^1_{i_0};[\tilde{q}^1_{j_0},\tilde{q}^1_{j_1}]\big) + h^\mu\big(\tilde{q}^2_{i_0} - \tilde{q}^2_{j_1}\big), \tag{B.3}$$

where for $x,y,z \in \mathbb{R}$,

$$g^\mu(x;[y,z]) = \begin{cases} |y-x| - \frac{1}{4\mu} & \text{if } \frac{1}{2\mu} \leq y-x, \\ 32\mu^5(y-x)^6 - 48\mu^4(y-x)^5 + 20\mu^3(y-x)^4 & \text{if } 0 \leq y-x \leq \frac{1}{2\mu}, \\ 0 & \text{if } y \leq x \leq z, \\ 32\mu^5(x-z)^6 + 48\mu^4(x-z)^5 + 20\mu^3(x-z)^4 & \text{if } 0 \leq x-z \leq \frac{1}{2\mu}, \\ |x-z| - \frac{1}{4\mu} & \text{if } x-z \geq \frac{1}{2\mu}. \end{cases} \tag{B.4}$$

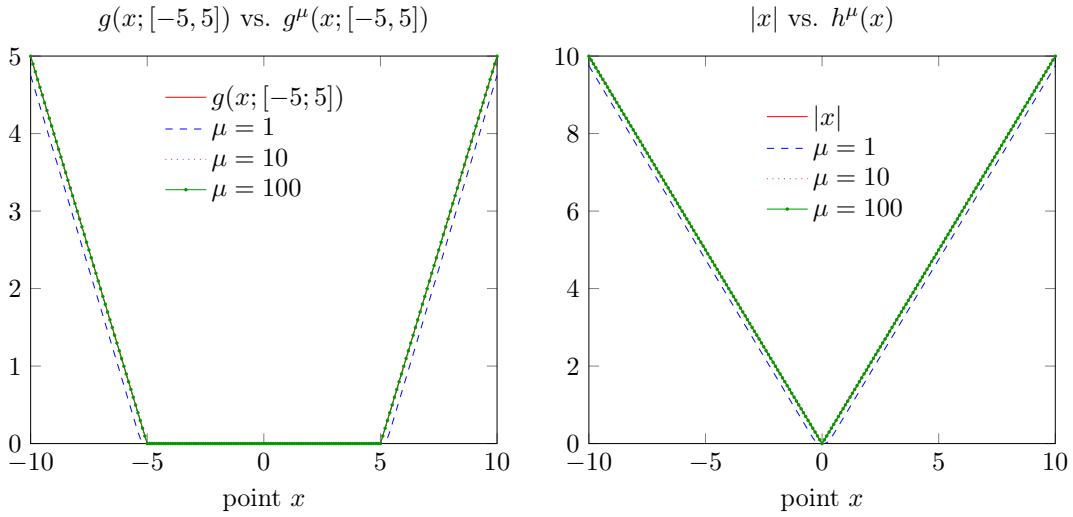


Figure B.4. Illustration of the functions involved in the computation of the vertex-edge 1-norm distances and its corresponding $\mathcal{C}^3$-regularizations.

and

$$h^\mu(x) = \begin{cases} |x| - \frac{1}{4\mu} & \text{if } |x| \geq \frac{1}{2\mu}, \\ 40\mu^5|x|^6 - 64\mu^4|x|^5 + 32\mu^3|x|^4 - 4\mu^2|x|^3 + \frac{\mu}{2}|x|^2 & \text{otherwise.} \end{cases} \tag{B.5}$$

Figure B.4 shows the regularization functions g^μ and h^μ for different values of the parameter μ and the original functions $g, |\cdot|$.

We now focus on proving that the proposed family of functions D_Q^μ satisfy the assumptions of theorem 5.2.9. Specifically, we verify $0 \leq D_Q^\mu \leq D_Q$ in proposition B.1.2 and argue that $Q \mapsto D_Q^\mu$ is of class $\mathcal{C}^3$ in proposition B.1.3. Moreover, proposition B.1.4 shows that in addition, the regularization is consistent, i. e., $f^\mu(Q) \to f(Q)$ holds when $\mu \to \infty$, for all $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$.

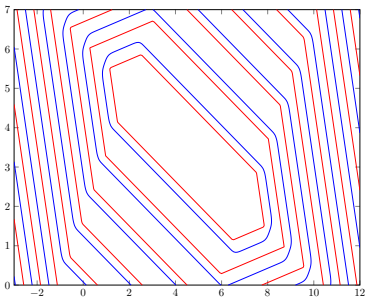
Proposition B.1.2. *For any $Q \in \mathcal{M}_+(\Delta; Q_{\text{ref}})$ and all $\mu \geq 1$, we have $0 \leq D_Q^\mu(i_0; [j_0, j_1]) \leq D_Q(i_0; [j_0, j_1])$.*

Proof. We need to establish $0 \leq h^\mu(x) \leq |x|$ and $0 \leq g^\mu(x; [y, z]) \leq g(x; [y, z])$ for all $x, y, z \in \mathbb{R}$ such that $y < z$.
We start by proving $h^\mu(x) \leq |x|$. When $|x| \geq \frac{1}{2\mu}$, then the first case in (B.5) applies and $h^\mu(x) \leq |x|$ is immediate. When $|x| < \frac{1}{2\mu}$, we estimate

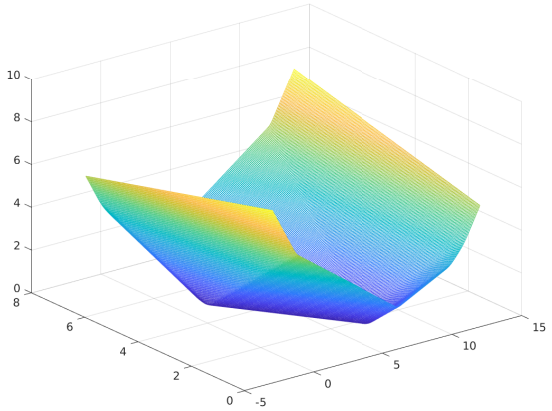
$$\begin{aligned} h^\mu(x) - |x| &= 40\mu^5|x|^6 - 64\mu^4|x|^5 + 32\mu^3|x|^4 - 4\mu^2|x|^3 + \frac{\mu}{2}|x|^2 - |x|, \\ &\leq \frac{1}{4\mu}(5\mu|x| - 8) + \frac{1}{2\mu}(8\mu|x| - 1) + \frac{1}{2\mu}\left(\frac{\mu|x|}{2} - 1\right), \\ &< -\frac{67}{80\mu^2} \leq 0, \end{aligned}$$

where we have used the fact that $|x| < 1/(2\mu)$. The claim $h^\mu \geq 0$ is immediately obtained from the definition when $|x| \geq 1/(2\mu)$. Conversely, for $(1 - 2\mu|x|) > 0$, we notice that

$$h^\mu(x) = 40\mu^5|x|^6 + 32\mu^3|x|^4(1 - 2\mu|x|) + \frac{\mu}{2}|x|^2(1 - 2\mu|x|) \geq 0.$$



(a) Contour plots of the distance D_Q (red) and regularized distance D_Q^μ (blue) with $\mu = 1$.



(b) $D_Q^\mu(i_0;[j_0,j_1])$ with $\mu = 1$.

Figure B.5. Illustration of $D_Q(i_0;[j_0,j_1])$ vs. $D_Q^\mu(i_0;[j_0,j_1])$ when $q_{j_0} = [2,5]^{\mathrm{T}}$ and $q_{j_1} = [7,2]^{\mathrm{T}}$ as function of the vertex q_{i_0}.

Next we prove $0 \leq g^\mu(x;[y,z]) \leq g(x;[y,z])$. We focus on the second case in (B.4), i. e., $y - x \leq \frac{1}{2\mu}$, since the other cases are simpler. Using the definitions, we have

$$\begin{aligned} g^\mu(x;[y,z]) - g(x;[y,z]) &= 32\mu^5(y-x)^6 - 48\mu^4(y-x)^5 + 20\mu^3(y-x)^4 - (y-x) \\ &= 16\mu^4(y-x)^5(2\mu(y-x)-3) + (y-x)\big(20\mu^3(y-x)^3 - 1\big) \\ &\leq -\frac{1}{4\gamma} < 0. \end{aligned}$$

The claim now follows immediately from the definition (B.3). In the same way, $g^\mu(x;[y,z]) \geq 0$ is immediately obtained from its definition when $1/(2\mu) \leq (y-x)$, $y \leq x \leq z$, $1 - 2\mu(x - z) \geq 0$, and $(x - z) \geq 1/(2\mu)$. Finally, for $1 - 2\mu(y - x) \geq 0$ the relation holds since

$$g^\mu(x;[y,z]) = 4\mu^3(y-x)^4\big(8\mu^2(y-x)^2 - 12\mu(y-x) + 5\big) \geq 0. \qquad \square$$

To illustrate the fact that $D_Q^\mu(i_0;[j_0,j_1])$ is an underestimate of $D_Q(i_0;[j_0,j_1])$, we revisit the example depicted in figure B.1, and plot the regularized distance, and also their corresponding contour plots.

Proposition B.1.3. *The function $\mathcal{M}_+(\Delta) \ni Q \mapsto D_Q^\mu(i_0;[j_0,j_1])$ is of class $\mathcal{C}^3$.*

Proof. The rotation $\mathcal{M}_+(\Delta) \ni Q \mapsto [\tilde{q}_{i_0}, \tilde{q}_{i_1}, \tilde{q}_{i_2}]$ is of class $\mathcal{C}^\infty$. The functions $x \mapsto h^\mu(x)$ and $x \mapsto g^\mu(x;[y,z])$ are of class $\mathcal{C}^3$ on $\mathbb{R}$ by construction. This can be verified in a straightforward way. In addition, $y \mapsto g^\mu(x;[y,z])$ and $z \mapsto g^\mu(x;[y,z])$ are of class $\mathcal{C}^3$. Since $D_Q^\mu(i_0;[j_0,j_1])$ consists of the composition of these functions with the rotation, it is of class $\mathcal{C}^3$ as well. $\square$

Proposition B.1.4. *For any $Q \in \mathcal{M}_+(\Delta;Q_{\mathrm{ref}})$, $D_Q^\mu(i_0;[j_0,j_1]) \to D_Q(i_0;[j_0,j_1])$ as $\mu \to \infty$. Consequently, $f_1^\mu(Q) \to f_1(Q)$, $f_2^\mu(Q) \to f_2(Q)$ hold as well.*

Proof. We start by proving $h^\mu(x) \to |x|$. When $x \neq 0$, then $h^\mu(x) = |x| - \frac{1}{4\mu}$ for μ sufficiently large and thus $h^\mu(x) \to |x|$. When $x = 0$, then $h^\mu(x) = 0$ for all μ.

Concerning g^μ, we distinguish the following cases. When $x < y < z$ holds, then we are in the first case in (B.4) for sufficiently large μ and thus for $\mu \to \infty$ we get $g^\mu(x;[y,z]) \to |y-x|$. When $x = y$, then the second case in (B.4) applies and $g^\mu(x;[y,z]) = 0$ for all μ. When $y < x < z$, then the third case is relevant for sufficiently large μ and thus $g^\mu(x;[y,z]) \to 0$ as $\mu \to \infty$. The remaining cases are similar.
The claim now follows immediately from the definition (B.3). □

B.2 Derivatives of the Regularized Augmentation Functions

This section is devoted to the presentation of the derivatives of the functions $f_1^\mu(Q;Q_{\mathrm{ref}})$ and $f_2^\mu(Q;Q_{\mathrm{ref}})$. We start by recalling the definition of $f_1^\mu(Q;Q_{\mathrm{ref}})$, given in (5.16).

$$f_1^\mu(Q;Q_{\mathrm{ref}}) \coloneqq \sum_{k=1}^{N_T}\sum_{\ell=0}^{2}\frac{\alpha_1}{h_Q^\ell\big(i_0^k,i_1^k,i_2^k\big)} + \sum_{[j_0,j_1]\in E_\partial}\sum_{\substack{i_0\in V_\partial\\ i_0\neq j_0,j_1}}\frac{\alpha_2}{D_Q^\mu(i_0;[j_0,j_1])} + \frac{\alpha_3}{2}\|Q-Q_{\mathrm{ref}}\|_F^2.$$

We also recall the definition of the heights $h_Q^\ell(i_0,i_1,i_2)$

$$h_Q^\ell(i_0,i_1,i_2) = \frac{2\,A_Q(i_0,i_1,i_2)}{E_Q^\ell(i_0,i_1,i_2)}, \quad \ell = 0,1,2,$$

and the signed area $A_Q(i_0,i_1,i_2)$

$$A_Q(i_0,i_1,i_2) \coloneqq \frac{1}{2}\det\left[q_{i_1}-q_{i_0},\ q_{i_2}-q_{i_1}\right].$$

Now, we fix the notation we use in the remainder of this chapter. Let $Q \in \mathcal{M}_+(\Delta;Q_{\mathrm{ref}})$, and recall that $Q = [q_1,\dots,q_{N_V}]$, where $q_i \in \mathbb{R}^2$. Moreover, q_i^ℓ stands for the first ($\ell = 1$) or second ($\ell = 2$) component of a vector q_i. We consider a fixed but arbitrary 2-face $[i_0,i_1,i_2] \in \Delta$ and its associated triangle $\mathrm{conv}_Q(i_0,i_1,i_2)$. We start by presenting the derivative of the α_3-term, i. e., the one involving the Frobenius norm. The derivative of the α_3-term with respect to $(\mathrm{vec}\,Q)_i$ is given by the following expression:

$$\frac{\partial\big(\|Q-Q_{\mathrm{ref}}\|_F^2\big)}{\partial(\mathrm{vec}\,Q)_i} = (\mathrm{vec}\,Q)_i - (\mathrm{vec}\,Q_{\mathrm{ref}})_i \tag{B.6}$$

Since the sum of the reciprocal of the heights can be computed in terms of the area and the perimeter of each triangle, we compute their first-order derivative as intermediate results. We use the notation $\mathrm{Per}_Q(i_0,i_1,i_2) \coloneqq E_Q^0(i_0,i_1,i_2) + E_Q^1(i_0,i_1,i_2) + E_Q^2(i_0,i_1,i_2)$. The first-order derivative of the area of each triangle $[i_0,i_1,i_2]$ is given by the following expression:

$$\frac{\partial A_Q(i_0,i_1,i_2)}{\partial(\mathrm{vec}\,Q)_i} = \begin{cases} q_{i_1}^2 - q_{i_2}^2 & \text{if } i = 2\,i_0 - 1,\\ q_{i_2}^1 - q_{i_1}^1 & \text{if } i = 2\,i_0,\\ q_{i_2}^2 - q_{i_0}^2 & \text{if } i = 2\,i_1 - 1,\\ q_{i_0}^1 - q_{i_2}^1 & \text{if } i = 2\,i_1,\\ q_{i_0}^2 - q_{i_1}^2 & \text{if } i = 2\,i_2 - 1,\\ q_{i_1}^1 - q_{i_0}^1 & \text{if } i = 2\,i_2,\\ 0 & \text{otherwise.}\end{cases} \tag{B.7}$$

In the same way, the derivative of the perimeter with respect to $(\operatorname{vec} Q)_i$ is mathematically specified by:

$$\frac{\partial \mathrm{Per}_Q(i_0,i_1,i_2)}{\partial(\operatorname{vec} Q)_i} = \begin{cases} -\frac{q^1_{i_1}-q^1_{i_0}}{\|q_{i_1}-q_{i_0}\|_2} + \frac{q^1_{i_0}-q^1_{i_2}}{\|q_{i_0}-q_{i_2}\|_2} & \text{if } i = 2\,i_0 - 1, \\ -\frac{q^2_{i_1}-q^2_{i_0}}{\|q_{i_1}-q_{i_0}\|_2} + \frac{q^2_{i_0}-q^2_{i_2}}{\|q_{i_0}-q_{i_2}\|_2} & \text{if } i = 2\,i_0, \\ \frac{q^1_{i_1}-q^1_{i_0}}{\|q_{i_1}-q_{i_0}\|_2} - \frac{q^1_{i_2}-q^1_{i_1}}{\|q_{i_2}-q_{i_1}\|_2} & \text{if } i = 2\,i_1 - 1, \\ \frac{q^2_{i_1}-q^2_{i_0}}{\|q_{i_1}-q_{i_0}\|_2} - \frac{q^2_{i_2}-q^2_{i_1}}{\|q_{i_2}-q_{i_1}\|_2} & \text{if } i = 2\,i_1, \\ \frac{q^1_{i_2}-q^1_{i_1}}{\|q_{i_2}-q_{i_1}\|_2} - \frac{q^1_{i_0}-q^1_{i_2}}{\|q_{i_0}-q_{i_2}\|_2} & \text{if } i = 2\,i_2 - 1, \\ \frac{q^2_{i_2}-q^2_{i_1}}{\|q_{i_2}-q_{i_1}\|_2} - \frac{q^2_{i_0}-q^2_{i_2}}{\|q_{i_0}-q_{i_2}\|_2} & \text{if } i = 2\,i_2, \\ 0, & \text{otherwise.} \end{cases} \tag{B.8}$$

Thus, the derivative of the α_1-term in the definition of f_1^μ with respect to $(\operatorname{vec} Q)_i$ satisfies:

$$\begin{aligned} \frac{\partial}{\partial(\operatorname{vec} Q)_i}\left(\sum_{\ell=0}^{2} \frac{1}{h^\ell_Q(i_0,i_1,i_2)}\right) = {} & \frac{1}{2A_Q(i_0,i_1,i_2)} \frac{\partial \mathrm{Per}_Q(i_0,i_1,i_2)}{\partial(\operatorname{vec} Q)_i} \\ & - \frac{\mathrm{Per}_Q(i_0,i_1,i_2)}{2A_Q(i_0,i_1,i_2)^2} \frac{\partial A_Q(i_0,i_1,i_2)}{\partial(\operatorname{vec} Q)_i}. \end{aligned} \tag{B.9}$$

Now, we present the derivatives of the regularized 1-norm based distance from a vertex to an edge $D^\mu_Q(i_0;[j_0,j_1])$. First of all, we recall its definition given in terms of the functions g^μ and h^μ.

$$D^\mu_Q(i_0;[j_0,j_1]) = g^\mu\big(\tilde{q}^1_{i_0};[\tilde{q}^1_{j_0},\tilde{q}^1_{j_1}]\big) + h^\mu\big(\tilde{q}^2_{i_0} - \tilde{q}^2_{j_1}\big),$$

where g^μ is given in (B.4) and h^μ is given in (B.5). Moreover, $\tilde{q}$ stands for the rotation of q about the origin so that its first coordinate aligns with the edge $\mathrm{conv}_Q(j_0,j_1)$. The derivatives of the functions g^μ and h^μ are given by the following expressions:

$$\mathrm{d}_x g^\mu \coloneqq \frac{\partial g^\mu}{\partial x} = \begin{cases} -1 & \text{if } \frac{1}{2\mu} \le y - x, \\ -192\mu^5(y-x)^5 + 240\mu^4(y-x)^4 - 80\mu^3(y-x)^3 & \text{if } 0 \le y - x \le \frac{1}{2\mu}, \\ 0 & \text{if } y \le x \le z, \\ 192\mu^5(x-z)^5 + 240\mu^4(x-z)^4 + 80\mu^3(x-z)^3 & \text{if } 0 \le x - z \le \frac{1}{2\mu}, \\ 1 & \text{if } x - z \ge \frac{1}{2\mu}, \end{cases} \tag{B.10}$$

$$\mathrm{d}_y g^\mu \coloneqq \frac{\partial g^\mu}{\partial y} = \begin{cases} 1 & \text{if } \frac{1}{2\mu} \le y - x, \\ 192\mu^5(y-x)^5 - 240\mu^4(y-x)^4 + 80\mu^3(y-x)^3 & \text{if } 0 \le y - x \le \frac{1}{2\mu}, \\ 0 & \text{otherwise,} \end{cases} \tag{B.11}$$

and

$$\mathrm{d}_z g^\mu \coloneqq \frac{\partial g^\mu}{\partial z} = \begin{cases} -1 & \text{if } x - z \ge \frac{1}{2\mu}, \\ -192\mu^5(y-x)^5 - 240\mu^4(y-x)^4 - 80\mu^3(y-x)^3 & \text{if } 0 \le x - z \le \frac{1}{2\mu}, \\ 0 & \text{otherwise.} \end{cases} \tag{B.12}$$

Since the regularized distance also involves the rotation (about the origin) vector $\tilde{q}$, we also consider its derivatives. Let us recall the definition of the rotation vector about the origin.

$$\tilde{q} = \begin{bmatrix} \cos(\omega) & \sin(\omega) \\ -\sin(\omega) & \cos(\omega) \end{bmatrix} q, \tag{B.13}$$

where $\omega(i_0;[j_0,j_1]) = \arctan\big((q^2_{j_1}-q^2_{j_0})/(q^1_{j_1}-q^1_{j_0})\big)$. Then, the derivatives of rotation angle ω are specified in what follows:

$$\frac{\partial\,\omega(i_0;[j_0,j_1])}{\partial(\operatorname{vec} Q)_i} = \begin{cases} \frac{q^2_{j_1}-q^2_{j_0}}{\|q_{j_1}-q_{j_0}\|^2} & \text{if } i = 2\,j_0-1, \\ -\frac{q^1_{j_1}-q^1_{j_0}}{\|q_{j_1}-q_{j_0}\|^2} & \text{if } i = 2\,j_0, \\ -\frac{q^2_{j_1}-q^2_{j_0}}{\|q_{j_1}-q_{j_0}\|^2} & \text{if } i = 2\,j_1-1, \\ \frac{q^1_{j_1}-q^1_{j_0}}{\|q_{j_1}-q_{j_0}\|^2} & \text{if } i = 2\,j_1, \\ 0 & \text{otherwise.} \end{cases}$$

By virtue of the chain rule, the first-order derivatives of the regularized distance D^μ_Q, satisfy:

$$\frac{\partial D^\mu_Q(i_0;[j_0,j_1])}{\partial(\operatorname{vec} Q)_i} = \begin{cases} \mathrm{d}_x g^\mu\,\cos(\omega) - (h^\mu)'\,\sin(\omega) & \text{if } i = 2\,i_0-1, \\ \mathrm{d}_x g^\mu\,\sin(\omega) + (h^\mu)'\,\cos(\omega) & \text{if } i = 2\,i_0, \\ \mathrm{d}_y g^\mu\,\cos(\omega) + \frac{\partial\omega}{\partial\operatorname{vec}(Q)_i}[\zeta] & \text{if } i = 2\,j_0-1, \\ \mathrm{d}_y g^\mu\,\sin(\omega) + \frac{\partial\omega}{\partial\operatorname{vec}(Q)_i}[\zeta] & \text{if } i = 2\,j_0, \\ \mathrm{d}_z g^\mu\,\cos(\omega) + (h^\mu)'\,\sin(\omega) + \frac{\partial\omega}{\partial\operatorname{vec}(Q)_i}[\zeta] & \text{if } i = 2\,j_1-1, \\ \mathrm{d}_z g^\mu\,\sin(\omega) - (h^\mu)'\,\cos(\omega) + \frac{\partial\omega}{\partial\operatorname{vec}(Q)_i}[\zeta] & \text{if } i = 2\,j_1, \\ 0 & \text{otherwise,} \end{cases} \tag{B.14}$$

where $\zeta = (\tilde{q}^2_{i_0})\,\mathrm{d}_x g^\mu + (\tilde{q}^2_{j_0})\,\mathrm{d}_y g^\mu + (\tilde{q}^2_{j_1})\,\mathrm{d}_z g^\mu - (\tilde{q}^1_{i_0})(h^\mu)' + (\tilde{q}^1_{j_1})(h^\mu)'$, and

$$(h^\mu)' = \begin{cases} \operatorname{sign}(x) & \text{if } |x| \geq \frac{1}{2\mu}, \\ 240\mu^5|x|^5 - 320\mu^4|x|^4 + 128\mu^3|x|^3 - 12\mu^2|x|^2 + \mu & \text{otherwise.} \end{cases} \tag{B.15}$$

Altogether, allows us to present the first-order derivative of the function f^μ_1 with respect to $(\operatorname{vec} Q)_i$.

$$\begin{aligned} \frac{\partial f^\mu_1}{\partial(\operatorname{vec} Q)_i} = &\sum_{k=1}^{N_T} \alpha_1 \frac{\partial}{\partial(\operatorname{vec} Q)_i}\left(\sum_{\ell=0}^{2} \frac{1}{h^\ell_Q(i^k_0, i^k_1, i^k_2)}\right) \\ &+ \sum_{[j_0,j_1]\in E_\partial} \sum_{\substack{i_0\in V_\partial \\ i_0\neq j_0,j_1}} \frac{-\alpha_2}{D^\mu_Q(i_0;[j_0,j_1])^2} \frac{\partial D^\mu_Q(i_0;[j_0,j_1])}{\partial(\operatorname{vec} Q)_i} \\ &+ \alpha_3[(\operatorname{vec} Q)_i - (\operatorname{vec} Q_{\mathrm{ref}})_i]\,. \end{aligned} \tag{B.16}$$

Now, we proceed to compute the first-order derivatives of the function f_2^μ given in (5.22), with respect to $(\operatorname{vec} Q)_i$. We recall its definition.

$$f_2^\mu(Q;Q_{\mathrm{ref}}) \coloneqq \sum_{k=1}^{N_T} \frac{1}{N_T} \frac{\beta_1}{\psi_Q(i_0^k,i_1^k,i_2^k)} + \frac{\beta_2}{\sum_{k=1}^{N_T} A_Q(i_0^k,i_1^k,i_2^k)} + \sum_{[j_0,j_1]\in E_\partial} \sum_{\substack{i_0\in V_\partial \\ i_0\neq j_0,j_1}} \frac{1}{\#E_\partial \# V_\partial} \frac{\beta_3}{D_Q^\mu(i_0;[j_0,j_1])} + \frac{\beta_4}{2N_V}\|Q-Q_{\mathrm{ref}}\|_F^2,$$

with

$$\frac{1}{\psi_Q(i_0,i_1,i_2)} \coloneqq \frac{\big(E_Q^0(i_0,i_1,i_2)\big)^2 + \big(E_Q^1(i_0,i_1,i_2)\big)^2 + \big(E_Q^2(i_0,i_1,i_2)\big)^2}{4\sqrt{3}\, A_Q(i_0,i_1,i_2)}.$$

Since the derivatives of the terms associated to β_3 and β_4 were already computed in (B.14) and (B.6), respectively. We will focus, on the first-order derivatives of the β_1– and β_2–terms. These terms also involve the area of each triangle, whose derivative was already computed in (B.7). Therefore, it only remains to compute the derivatives of the sum of the squared edge lengths. We denote this quantity as $S_Q(i_0,i_1,i_2) \coloneqq \big(E_Q^0(i_0,i_1,i_2)\big)^2 + \big(E_Q^1(i_0,i_1,i_2)\big)^2 + \big(E_Q^2(i_0,i_1,i_2)\big)^2$. Then, the first-order derivative of the sum of the squared edge lengths of each triangle with respect to $(\operatorname{vec} Q)_i$ is given by the following expression:

$$\frac{\partial S_Q(i_0,i_1,i_2)}{\partial(\operatorname{vec} Q)_i} = \begin{cases} 4\,q_{i_0}^1 - 2\,q_{i_1}^1 - 2\,q_{i_2}^1 & \text{if } i = 2\,i_0 - 1, \\ 4\,q_{i_0}^2 - 2\,q_{i_1}^2 - 2\,q_{i_2}^2 & \text{if } i = 2\,i_0, \\ 4\,q_{i_1}^1 - 2\,q_{i_0}^1 - 2\,q_{i_2}^1 & \text{if } i = 2\,i_1 - 1, \\ 4\,q_{i_1}^2 - 2\,q_{i_0}^2 - 2\,q_{i_2}^2 & \text{if } i = 2\,i_1, \\ 4\,q_{i_2}^1 - 2\,q_{i_0}^1 - 2\,q_{i_1}^1 & \text{if } i = 2\,i_2 - 1, \\ 4\,q_{i_2}^2 - 2\,q_{i_0}^2 - 2\,q_{i_1}^2 & \text{if } i = 2\,i_2, \\ 0 & \text{otherwise.} \end{cases} \tag{B.17}$$

Thus, the derivative of the function $\frac{1}{\psi_Q}$ satisfies:

$$\frac{\partial}{\partial(\operatorname{vec} Q)_i}\left(\frac{1}{\psi_Q(i_0,i_1,i_2)}\right) = \frac{1}{4\sqrt{3}A_Q(i_0,i_1,i_2)} \frac{\partial S_Q[i_0,i_1,i_2]}{\partial(\operatorname{vec} Q)_i} - \frac{S_Q[i_0,i_1,i_2]}{4\sqrt{3}A_Q(i_0,i_1,i_2)^2} \frac{\partial A_Q(i_0,i_1,i_2)}{\partial(\operatorname{vec} Q)_i}. \tag{B.18}$$

The derivative of the total area is then given by:

$$\frac{\partial}{\partial(\operatorname{vec} Q)_i}\left(\frac{1}{\sum_{k=1}^{N_T} A_Q(i_0^k,i_1^k,i_2^k)}\right) = -\frac{1}{\left(\sum_{k=1}^{N_T} A_Q(i_0^k,i_1^k,i_2^k)\right)^2} \sum_{k=1}^{N_T} \frac{A_Q(i_0^k,i_1^k,i_2^k)}{\partial(\operatorname{vec} Q)_i}. \tag{B.19}$$

Altogether implies that the first-order derivative of the function $f_2^\mu(Q;Q_{\mathrm{ref}})$ given in (5.22) with respect to $(\operatorname{vec} Q)_i$ satisfies:

$$\frac{\partial f_2^{\mu}}{\partial (\operatorname{vec} Q)_i} = \sum_{k=1}^{N_T} \frac{\beta_1}{N_T} \frac{\partial}{\partial (\operatorname{vec} Q)_i} \left(\frac{1}{\psi_Q(i_0^k, i_1^k, i_2^k)} \right) + \beta_2 \frac{\partial}{\partial (\operatorname{vec} Q)_i} \left(\frac{1}{\sum_{k=1}^{N_T} A_Q(i_0^k, i_1^k, i_2^k)} \right) + \sum_{[j_0,j_1] \in E_\partial} \sum_{\substack{i_0 \in V_\partial \\ i_0 \neq j_0, j_1}} \frac{-\beta_3}{\#E_\partial \#V_\partial D_Q^{\mu}(i_0; [j_0, j_1])^2} \frac{\partial D_Q^{\mu}(i_0; [j_0, j_1])}{\partial (\operatorname{vec} Q)_i} + \frac{\beta_4}{N_V} [(\operatorname{vec} Q)_i - (\operatorname{vec} Q_{\mathrm{ref}})_i] \,. \tag{B.20}$$

Bibliography

Absil, P.-A.; R. Mahony; R. Sepulchre (2008). *Optimization Algorithms on Matrix Manifolds*. Princeton University Press. DOI: 10.1515/9781400830244.

Afraites, L.; M. Dambrine; D. Kateb (2008). "On second order shape optimization methods for electrical impedance tomography". *SIAM journal on control and optimization* 47.3, pp. 1556–1590. DOI: https://doi.org/10.1137/070687438.

Agricola, I.; T. Friedrich (2008). *Elementary Geometry*. Vol. 43. Student Mathematical Library. American Mathematical Society. DOI: 10.1090/stml/043.

Alexa, M. (2002). "Recent advances in mesh morphing". *Computer Graphics Forum* 21.2, pp. 173–198. DOI: 10.1111/1467-8659.00575.

Alexa, M.; D. Cohen-Or; D. Levin (2000). "As-rigid-as-possible shape interpolation". *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00, pp. 157–164. DOI: 10.1145/344779.344859.

Allaire, G. (2012). *Shape optimization by the homogenization method*. Vol. 146. Springer Science & Business Media. DOI: https://doi.org/10.1007/978-1-4684-9286-6.

Allaire, G.; F. Jouve; A.-M. Toader (2004). "Structural optimization using sensitivity analysis and a level-set method". *Journal of computational physics* 194.1, pp. 363–393. DOI: https://doi.org/10.1016/j.jcp.2003.09.032.

Alsina, C.; R. B. Nelsen (2008). "Geometric proofs of the Weitzenböck and Hadwiger-Finsler inequalities". *Mathematics Magazine* 81.3, pp. 216–219. DOI: https://doi.org/10.1080/0025570X.2008.11953553.

Ambrosio, L.; G. Buttazzo (1993). "An optimal design problem with perimeter penalization". *Calculus of variations and partial differential equations* 1.1, pp. 55–69.

Amenta, N.; C. Rojas (2020). "Dihedral deformation and rigidity". *Computational Geometry* 90. DOI: https://doi.org/10.1016/j.comgeo.2020.101657.

Baghaie, A.; Z. Yu; R. M. D'souza (2014). "Fast mesh-based medical image registration". *Advances in Visual Computing*. Springer International Publishing, pp. 1–10. DOI: 10.1007/978-3-319-14364-4_1.

Bandara, K.; F. Cirak; G. Of; O. Steinbach; J. Zapletal (2015). "Boundary element based multiresolution shape optimisation in electrostatics". *Journal of computational physics* 297, pp. 584–598. DOI: https://doi.org/10.1016/j.jcp.2015.05.017.

Bänsch, E.; P. Morin; R. H. Nochetto (2005). "A finite element method for surface diffusion: the parametric case". *Journal of Computational Physics* 203.1, pp. 321–343. DOI: 10.1016/j.jcp.2004.08.022.

Bartels, S.; G. Wachsmuth (2020). "Numerical approximation of optimal convex shapes". *SIAM Journal on Scientific Computing* 42.2, A1226–A1244. DOI: 10.1137/19m1256853.

Bauer, M.; M. Bruveris; P. W. Michor (2014). "Overview of the geometries of shape spaces and diffeomorphism groups". *Journal of Mathematical Imaging and Vision* 50.1-2, pp. 60–97. DOI: 10.1007/s10851-013-0490-z.

Berggren, M. (2010). "A unified discrete-continuous sensitivity analysis method for shape optimization". *Applied and Numerical Partial Differential Equations*. Vol. 15. Computational Methods in Applied Sciences. Springer, New York, pp. 25–39. DOI: 10.1007/978-90-481-3239-3_4.

Bhatia, R. P.; K. L. Lawrence (1990). "Two-dimensional finite element mesh generation based on stripwise automatic triangulation". *Computers & Structures* 36.2, pp. 309–319. DOI: https://doi.org/10.1016/0045-7949(90)90131-K.

Bırsan, T. (2015). "Bounds for elements of a triangle expressed by R, r, and s". *Forum Geometricorum*. Vol. 15, pp. 99–103.

Bobrowski, K.; E. Ferrer; E. Valero; H. Barnewitz (2017). "Aerodynamic shape optimization using geometry surrogates and adjoint method". *AIAA Journal* 55.10, pp. 3304–3317. DOI: 10.2514/1.J055766.

Bonito, A.; R. H. Nochetto; M. S. Pauletti (2010). "Geometrically consistent mesh modification". *SIAM Journal on Numerical Analysis* 48.5, pp. 1877–1899. DOI: 10.1137/100781833.

Boumal, N. (2020). *An Introduction to Optimization on Smooth Manifolds*. URL: http://www.nicolasboumal.net/book.

Brezis, H. (2011). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York. DOI: 10.1007/978-0-387-70914-7.

Bucur, D.; G. Buttazzo (2005). *Variational methods in some shape optimization problems*. Vol. 65. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Boston. DOI: https://doi.org/10.1007/b137163.

Burger, M.; R. Stainko (2006). "Phase-field relaxation of topology optimization with local stress constraints". *SIAM Journal on Control and Optimization* 45.4, pp. 1447–1466. DOI: https://doi.org/10.1137/05062723X.

Burger, M.; S. Osher (2005). "A survey on level set methods for inverse problems and optimal design". *European Journal of Applied Mathematics* 16.2, pp. 263–301. DOI: 10.1017/S0956792505006182.

Chenais, D. (1975). "On the existence of a solution in a domain identification problem". *Journal of Mathematical Analysis and Applications* 52.2, pp. 189–219.

Ciarlet, P. G. (2002). *The finite element method for elliptic problems*. SIAM.

Ciarlet, P. G. (2013). *Linear and nonlinear functional analysis with applications*. Vol. 130. Siam.

Dapogny, C. (2013). "Shape optimization, level set methods on unstructuredmeshes and mesh evolution". PhD thesis. Pierre and Marie Curie University.

De Gournay, F.; J. Fehrenbach; F. Plouraboué (2014). "Shape optimization for the generalized Graetz problem". *Structural and Multidisciplinary Optimization* 49.6, pp. 993–1008. DOI: 10.1007/s00158-013-1032-4.

De los Reyes, J. C. (2015). *Numerical PDE-Constrained Optimization*. SpringerBriefs in Optimization. Springer, Cham. DOI: 10.1007/978-3-319-13395-9.

De los Reyes, J. C.; E. Loayza-Romero (2019). "Total generalized variation regularization in data assimilation for Burgers' equation". *Inverse Problems and Imaging* 13.4, pp. 755–786. DOI: 10.3934/ipi.2019035.

Deckelnick, K.; P. J. Herbert; M. Hinze (2021). *A novel $W^{1,\infty}$ approach to shape optimisation with Lipschitz domains*. arXiv: 2103.13857 [math.OC].

Delfour, M.; G. Payre; J.-P. Zolesio (1985). "An optimal triangulation for second-order elliptic problems". *Computer Methods in Applied Mechanics and Engineering* 50, pp. 231–261. DOI: 10.1016/0045-7825(85)90095-7.

Delfour, M.; J.-P. Zolésio (2001). *Shapes and Geometries. Analysis, Differential Calculus, and Optimization*. Philadelphia: SIAM.

Delfour, M.; J.-P. Zolésio (2011). *Shapes and Geometries*. 2nd ed. Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9780898719826.

Do Carmo, M. P. (1992). *Riemannian Geometry*. Mathematics: Theory & Applications. Boston, MA: Birkhäuser Boston, Inc.

Doğan, G.; P. Morin; R. H. Nochetto; M. Verani (2007). "Discrete gradient flows for shape optimization and applications". *Computer Methods in Applied Mechanics and Engineering* 196.37–40, pp. 3898–3914. DOI: 10.1016/j.cma.2006.10.046.

Dokken, J. S.; S. W. Funke; A. Johansson; S. Schmidt (2018). *Shape optimization using the finite element method on multiple meshes with Nitsche coupling*. arXiv: 1806.09821.

Dokken, J. S.; S. W. Funke; A. Johansson; S. Schmidt (2019). "Shape optimization using the finite element method on multiple meshes with Nitsche coupling". *SIAM Journal on Scientific Computing* 41.3, A1923–A1948. DOI: 10.1137/18M1189208.

Dubrovin, B. A.; A. T. Fomenko; S. P. Novikov (1992). *Modern Geometry–Methods and Applications (Graduate Texts in Mathematics) (Pt. 1)*. Springer-Verlag.

Edelsbrunner, H.; J. L. Harer (2010). *Computational Topology*. An introduction. American Mathematical Society, pp. xii+241. DOI: 10.1090/mbk/069.

Elman, H. C.; D. J. Silvester; A. J. Wathen (2014). *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*. 2nd ed. Numerical Mathematics and Scientific Computation. Oxford University Press. DOI: 10.1093/acprof:oso/9780199678792.001.0001.

Eppler, K.; H. Harbrecht (2005). "A regularized Newton method in electrical impedance tomography using shape Hessian information". *Control and Cybernetics* 34.1, pp. 203–225.

Etling, T.; R. Herzog; E. Loayza; G. Wachsmuth (2020). "First and second order shape optimization based on restricted mesh deformations". *SIAM Journal on Scientific Computing* 42.2, A1200–A1225. DOI: 10.1137/19m1241465. arXiv: 1810.10313.

Feppon, F.; G. Allaire; F. Bordeu; J. Cortial; C. Dapogny (2018). *Shape optimization of a coupled thermal fluid-structure problem in a level set mesh evolution framework*. HAL: hal-01686770.

Fuchs, M.; B. Jüttler; O. Scherzer; H. Yang (2009). "Shape metrics based on elastic deformations". *Journal of Mathematical Imaging and Vision* 35.1, pp. 86–102. DOI: 10.1007/s10851-009-0156-z.

Gallier, J. (2008). *Notes on convex sets, polytopes, polyhedra, combinatorial topology, Voronoi diagrams and Delaunay triangulations*. arXiv: 0805.0292.

Gangl, P.; U. Langer; A. Laurain; H. Meftahi; K. Sturm (2015). "Shape optimization of an electric motor subject to nonlinear magnetostatics". *SIAM Journal on Scientific Computing* 37.6, B1002–B1025. DOI: 10.1137/15100477X.

Garcke, H.; M. Hinze; C. Kahle; K. F. Lam (2018). "A phase field approach to shape optimization in Navier–Stokes flow with integral state constraints". *Advances in Computational Mathematics* 44.5, pp. 1345–1383. DOI: https://doi.org/10.1007/s10444-018-9586-8.

Geiersbach, C.; E. Loayza-Romero; K. Welker (2021a). *PDE-constrained shape optimization: towards product shape spaces and stochastic models*. arXiv: 2107.07744 [math.OC].

Geiersbach, C.; E. Loayza-Romero; K. Welker (2021b). "Stochastic approximation for optimization in shape spaces". *SIAM Journal on Optimization* 31.1, pp. 348–376. DOI: 10.1137/20M1316111. arXiv: 2001.10786.

Geiersbach, C.; E. Loayza; K. Welker (2019). *Computational aspects for interface identification problems with stochastic modelling*. arXiv: 1902.01160.

Giacomini, M.; O. Pantz; K. Trabelsi (2017). "Certified descent algorithm for shape optimization driven by fully-computable a posteriori error estimators". *ESAIM. Control, Optimisation and Calculus of Variations* 23.3, pp. 977–1001. DOI: 10.1051/cocv/2016021.

Glowinski, R.; J. He (1998). "On shape optimization and related issues". *Computational Methods for Optimal Design and Control*. Springer, pp. 151–179. DOI: 10.1007/978-1-4612-1780-0_10.

Golub, G.; C. van Loan (1983). *Matrix Computations*. Baltimore: Johns Hopkins University Press.

Gordon, W. B. (1973). "An analytical criterion for the completeness of Riemannian manifolds". *Proceedings of the American Mathematical Society* 37, pp. 221–225. DOI: 10.2307/2038738.

Grossmann, C.; H.-G. Roos; M. Stynes (2007). *Numerical Treatment of Partial Differential Equations*. Universitext. Translation and Revision of the 3rd edition of "Numerische Behandlung partieller Differentialgleichungen" published by Teubner, 2005. Berlin: Springer. DOI: 10.1007/978-3-540-71584-9.

Hadamard, J. (1908). *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*. Vol. 33. Imprimerie nationale.

Hairer, E.; C. Lubich; G. Wanner (2003). "Geometric numerical integration illustrated by the Störmer–Verlet method". *Acta numerica* 12, pp. 399–450. DOI: https://doi.org/dp4xt5.

Hardesty, Antil; Kouri; Ridzal (2020). *The strip method for shape derivatives*.

Hardesty, S.; D. P. Kouri; P. Lindsay; D. Ridzal; B. L. Stevens; R. Viertel (2020). *Shape Optimization for Control and Isolation of Structural Vibrations in Aerospace and Defense Applications*. Tech. rep. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Haubner, J.; M. Siebenborn; M. Ulbrich (2021). "A continuous perspective on shape optimization via domain transformations". *SIAM Journal on Scientific Computing* 43.3, A1997–A2018. DOI: 10.1137/20m1332050. arXiv: 2004.06942.

Heeren, B.; M. Rumpf; M. Wardetzky; B. Wirth (2012). "Time-discrete geodesics in the space of shells". *Computer Graphics Forum* 31.5, pp. 1755–1764. DOI: 10.1111/j.1467-8659.2012.03180.x.

Herzog, R.; E. Loayza-Romero (2021). *A Discretize-Then-Optimize Approach to PDE-Constrained Shape Optimization*. arXiv: 2109.00076 [math.OC].

Herzog, R. (2018). *Optimization on Manfiolds*. Lecture notes, SS2018, Chemnitz University of Technology.

Herzog, R.; E. Loayza-Romero (2020). *A manifold of planar triangular meshes with complete Riemannian metric*. arXiv: 2012.05624.

Hintermüller, M.; A. Laurain; I. Yousept (2015). "Shape sensitivities for an inverse problem in magnetic induction tomography based on the eddy current model". *Inverse Problems. An International Journal on the Theory and Practice of Inverse Problems, Inverse Methods and Computerized Inversion of Data* 31.6, pp. 065006, 25. DOI: 10.1088/0266-5611/31/6/065006.

Hinze, M.; R. Pinnau; M. Ulbrich; S. Ulbrich (2009). *Optimization with PDE Constraints*. Berlin: Springer. DOI: 10.1007/978-1-4020-8839-1.

Hiptmair, R.; A. Paganini; S. Sargheini (2015). "Comparison of approximate shape gradients". *BIT. Numerical Mathematics* 55.2, pp. 459–485. DOI: 10.1007/s10543-014-0515-z.

Hiptmair, R.; A. Paganini (2015). "Shape optimization by pursuing diffeomorphisms". *Computational Methods in Applied Mathematics* 15.3, pp. 291–305.

Hiriart-Urruty, J.-B.; C. Lemaréchal (2001). *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin. DOI: 10.1007/978-3-642-56468-0.

Horak, D.; J. Jost (2013). "Spectra of combinatorial Laplace operators on simplicial complexes". *Advances in Mathematics* 244, pp. 303–336. DOI: 10.1016/j.aim.2013.05.007.

Iglesias, J. A.; K. Sturm; F. Wechsung (2018). "Two-dimensional shape optimization with nearly conformal transformations". *SIAM Journal on Scientific Computing* 40.6, A3807–A3830. DOI: 10.1137/17M1152711.

Ito, K.; K. Kunisch; G. H. Peichl (2008). "Variational approach to shape derivatives". *ESAIM. Control, Optimisation and Calculus of Variations* 14.3, pp. 517–539. DOI: 10.1051/cocv:2008002.

Kay, D. C. (2011). *College geometry: a unified development*. CRC Press.

Kendall, D. G. (1984). "Shape manifolds, procrustean metrics, and complex projective spaces". *Bulletin of the London Mathematical Society* 16.2, pp. 81–121. DOI: 10.1112/blms/16.2.81.

Kilian, M.; N. J. Mitra; H. Pottmann (2007). "Geometric modeling in shape space". *ACM Transactions on Graphics (TOG)*. Vol. 26. ACM, p. 64. DOI: 10.1145/1275808.1276457.

Klassen, E.; A. Srivastava; M. Mio; S. H. Joshi (2004). "Analysis of planar shapes using geodesic paths on shape spaces". *IEEE transactions on pattern analysis and machine intelligence* 26.3, pp. 372–383. DOI: 10.1109/TPAMI.2004.1262333.

Koko, J. (2016a). *Fast MATLAB assembling functions for 2D/3D FEM Matrices*. URL: https://www.mathworks.com/matlabcentral/fileexchange/59616.

Koko, J. (2016b). "Fast MATLAB assembly of FEM matrices in 2D and 3D using cell-array approach". *International Journal of Modeling, Simulation, and Scientific Computing* 07.02, p. 1650010. DOI: 10.1142/s1793962316500100.

Laurain, A. (2018). "A level set-based structural optimization code using FEniCS". *Structural and Multidisciplinary Optimization* 58.3, pp. 1311–1334. DOI: 10.1007/s00158-018-1950-2. arXiv: 1705.01442.

Laurain, A.; K. Sturm (2016). "Distributed shape derivative via averaged adjoint method and applications". *ESAIM. Mathematical Modelling and Numerical Analysis* 50.4, pp. 1241–1267. DOI: 10.1051/m2an/2015075.

Lee, J. M. (2011). *Introduction to Topological Manifolds*. 2nd ed. Vol. 202. Graduate Texts in Mathematics. Springer, New York, pp. xviii+433. DOI: 10.1007/978-1-4419-7940-7.

Lee, J. M. (2012). *Introduction to Smooth Manifolds*. 2nd ed. Springer New York. DOI: 10.1007/978-1-4419-9982-5.

Lee, J. M. (2018). *Introduction to Riemannian Manifolds*. Springer International Publishing. DOI: 10.1007/978-3-319-91755-9.

Liu, X.; Y. Shi; I. Dinov; W. Mio (2010). "A computational model of multidimensional shape". *International journal of computer vision* 89.1, pp. 69–83. DOI: https://doi.org/10.1007/s11263-010-0323-0.

Lozano, C. (2017). "On mesh sensitivities and boundary formulas for discrete adjoint-based gradients in inviscid aerodynamic shape optimization". *Journal of Computational Physics* 346, pp. 403–436. DOI: 10.1016/j.jcp.2017.06.025.

Luft, D.; V. Schulz (2021a). *Pre-Shape Calculus: Foundations and Application to Mesh Quality Optimization*. arXiv: 2012.09124.

Luft, D.; V. Schulz (2021b). *Simultaneous Shape and Mesh Quality Optimization using Pre-Shape Calculus*. arXiv: 2103.15109.

Magnot, J.-P. (2016). "Differentiation on spaces of triangulations and optimized triangulations." *Journal of Physics: Conference Series*. Vol. 738. 1. IOP Publishing, p. 012088.

Magnot, J.-P. (2020). "On the differential geometry of numerical schemes and weak solutions of functional equations". *Nonlinearity* 33.12, pp. 6835–6867. DOI: 10.1088/1361-6544/abaa9f. URL: https://doi.org/10.1088/1361-6544/abaa9f.

Michor, P. W.; D. Mumford (2005). "Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms". *Documenta Mathematica* 10, pp. 217–245.

Michor, P. W.; D. Mumford (2007). "An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach". *Applied and Computational Harmonic Analysis* 23.1, pp. 74–113. DOI: 10.1016/j.acha.2006.07.004.

Mio, W.; A. Srivastava; S. Joshi (2006). "On shape of plane elastic curves". *International Journal of Computer Vision* 73.3, pp. 307–324. DOI: 10.1007/s11263-006-9968-0.

Misztal, M. K. (2010). "Deformable Simplicial Complexes". PhD thesis. Technical University of Denmark.

Moakher, M.; M. Zéraï (2011). "The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data". *Journal of Mathematical Imaging and Vision* 40.2, pp. 171–187. DOI: https://doi.org/10.1007/s10851-010-0255-x.

Morin, P.; R. H. Nochetto; M. S. Pauletti; M. Verani (2012). "Adaptive finite element method for shape optimization". *ESAIM. Control, Optimisation and Calculus of Variations* 18.4, pp. 1122–1149. DOI: 10.1051/cocv/2011192.

Müller, P. M.; N. Kühl; M. Siebenborn; K. Deckelnick; M. Hinze; T. Rung (2021). *A Novel p-Harmonic Descent Approach Applied to Fluid Dynamic Shape Optimization*. arXiv: 2103.14735.

Munkres, J. R. (2018). *Elements of algebraic topology*. CRC press.

Murat, F.; J. Simon (1977). "Optimal control with respect to the domain". *These de l'Université Paris VI*.

Nocedal, J.; S. J. Wright (2006). *Numerical Optimization*. 2nd ed. New York: Springer. DOI: 10.1007/978-0-387-40065-5.

Novruzi, A.; J. R. Roche (2000). "Newton's method in shape optimisation: a three-dimensional case". *BIT. Numerical Mathematics* 40.1, pp. 102–120. DOI: 10.1023/A:1022370419231.

Onyshkevych, S.; M. Siebenborn (2021). "Mesh quality preserving shape optimization using nonlinear extension operators". *Journal of Optimization Theory and Applications* 189.1, pp. 291–316. DOI: https://doi.org/10.1007/s10957-021-01837-8.

Osher, S.; J. A. Sethian (1988). "Fronts propagating with curvature-dependent speed". *Journal of Computational Physics* 79, pp. 12–49. DOI: 10.1016/0021-9991(88)90002-2.

Paganini, A.; F. Wechsung; P. E. Farrell (2018). "Higher-order moving mesh methods for PDE-constrained shape optimization". *SIAM Journal on Scientific Computing* 40.4, A2356–A2382. DOI: https://doi.org/10.1137/17M1133956.

Pironneau, O. (1984). *Optimal Shape Design for Elliptic Systems*. New York: Springer. DOI: 10.1007/978-3-642-87722-3.

Quarteroni, A. (2009). *Numerical models for differential problems*. Vol. 2. Springer.

Quarteroni, A.; A. Valli (1994). *Numerical Approximation of Partial Differential Equations*. Berlin: Springer. DOI: 10.1007/978-3-540-85268-1.

Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems*. 2nd ed. Philadelphia: SIAM. DOI: 10.1137/1.9780898718003.

Schillings, C.; S. Schmidt; V. Schulz (2011). "Efficient shape optimization for certain and uncertain aerodynamic design". *Computers & Fluids. An International Journal* 46, pp. 78–87. DOI: 10.1016/j.compfluid.2010.12.007.

Schmidt, S. (2014). *A two stage CVT / eikonal convection mesh deformation approach for large nodal deformations*. arXiv: 1411.7663.

Schmidt, S.; C. Ilic; V. Schulz; N. R. Gauger (2011). "Airfoil design for compressible inviscid flow based on shape calculus". *Optimization and Engineering. International Multidisciplinary Journal to Promote Optimization Theory & Applications in Engineering Sciences* 12.3, pp. 349–369. DOI: 10.1007/s11081-011-9145-3.

Schmidt, S.; C. Ilic; V. Schulz; N. R. Gauger (2013). "Three-dimensional large-scale aerodynamic shape optimization based on shape calculus". *AIAA Journal* 51.11, pp. 2615–2627. DOI: 10.2514/1.j052245.

Schmidt, S.; V. H. Schulz (2009). "Impulse response approximations of discrete shape Hessians with application in CFD". *SIAM Journal on Control and Optimization* 48.4, pp. 2562–2580. DOI: 10.1137/080719844.

Schmidt, S.; V. H. Schulz (2010). "Shape derivatives for general objective functions and the incompressible Navier-Stokes equations". *Control and Cybernetics* 39.3, pp. 677–713.

Schmidt, S.; V. Schulz; C. Ilic; N. Gauger (2011). "Three dimensional large scale aerodynamic shape optimization based on shape calculus". *41st AIAA Fluid Dynamics Conference and Exhibit*. American Institute of Aeronautics and Astronautics. DOI: 10.2514/6.2011-3718.

Schmidt, S.; E. Wadbro; M. Berggren (2016). "Large-scale three-dimensional acoustic horn optimization". *SIAM Journal on Scientific Computing* 38.6, B917–B940. DOI: 10.1137/15M1021131.

Schulz, V.; M. Siebenborn; K. Welker (2014). *Towards a Lagrange-Newton approach for PDE constrained shape optimization*. arXiv: 1405.3266.

Schulz, V. H. (2014). "A Riemannian view on shape optimization". *Foundations of Computational Mathematics. The Journal of the Society for the Foundations of Computational Mathematics* 14.3, pp. 483–501. DOI: 10.1007/s10208-014-9200-5.

Schulz, V. H.; M. Siebenborn (2016). "Computational comparison of surface metrics for PDE constrained shape optimization". *Computational Methods in Applied Mathematics* 16.3, pp. 485–496. DOI: 10.1515/cmam-2016-0009.

Schulz, V. H.; M. Siebenborn; K. Welker (2015a). *Structured inverse modeling in parabolic diffusion problems*. arXiv: 1409.3464.

Schulz, V. H.; M. Siebenborn; K. Welker (2015b). "Structured inverse modeling in parabolic diffusion problems". *SIAM Journal on Control and Optimization* 53.6, pp. 3319–3338. DOI: 10.1137/140985883.

Schulz, V. H.; M. Siebenborn; K. Welker (2015c). "Towards a Lagrange-Newton approach for PDE constrained shape optimization". *New Trends in Shape Optimization*. Ed. by A. Pratelli; G. Leugering. Vol. 166. International Series of Numerical Mathematics. Birkhäuser/Springer, Cham, pp. 229–249. DOI: 10.1007/978-3-319-17563-8_10.

Schulz, V. H.; M. Siebenborn; K. Welker (2016). "Efficient PDE constrained shape optimization based on Steklov-Poincaré type metrics". *SIAM Journal on Optimization* 26.4, pp. 2800–2819. DOI: 10.1137/15M1029369.

Shewchuk, J. R. (2002). *What is a good linear finite element? Interpolation, conditioning, anisotropy, and quality measures*. Tech. rep. Department of Electrical Engineering and Computer Sciences, University of Californa at Berkeley. URL: http://www.cs.berkeley.edu/~jrs/papers/elemj.pdf.

Simon, J. (1980). "Differentiation with respect to the domain in boundary value problems". *Numerical Functional Analysis and Optimization* 2.7-8, pp. 649–687.

Sokołowski, J.; J.-P. Zolésio (1992). *Introduction to Shape Optimization*. New York: Springer. DOI: 10.1007/978-3-642-58106-9.

Souli, M.; J. P. Zolésio (1993). "Shape derivative of discretized problems". *Computer Methods in Applied Mechanics and Engineering* 108.3-4, pp. 187–199.

Sturm, K. (2015). "Minimax Lagrangian approach to the differentiability of nonlinear PDE constrained shape functions without saddle point assumptions". *SIAM Journal on Control and Optimization* 53.4, pp. 2017–2039. DOI: 10.1137/130930807.

Sturm, K. (2016). "Shape optimization with nonsmooth cost functions: from theory to numerics". *SIAM Journal on Control and Optimization* 54.6, pp. 3319–3346. DOI: 10.1137/16M1069882.

Svrtan, D.; D. Veljan (2012). "Non–Euclidean versions of some classical triangle inequalities". *Forum geometricorum*. Vol. 12, pp. 197–209.

Tartar, L. (2000). "An introduction to the homogenization method in optimal design". *Optimal shape design*. Springer, pp. 47–156. DOI: https://doi.org/10.1007/BFb0106742.

Tröltzsch, F. (2010). *Optimal Control of Partial Differential Equations*. Vol. 112. Graduate Studies in Mathematics. Providence: American Mathematical Society. DOI: 10.1090/gsm/112.

Van Keulen, F.; R. T. Haftka; N. H. Kim (2005). "Review of options for structural design sensitivity analysis. Part 1: Linear systems". *Computer methods in applied mechanics and engineering* 194.30-33, pp. 3213–3243. DOI: https://doi.org/10.1016/j.cma.2005.02.002.

Wilke, D. N.; S. Kok; A. A. Groenwold (2005). "A quadratically convergent unstructured remeshing strategy for shape optimization". *International Journal for Numerical Methods in Engineering* 65.1, pp. 1–17. DOI: 10.1002/nme.1430.

Yang, C.-Y.; Y.-L. Chang; J.-S. Chen (2011). "Analysis of nonsmooth vector-valued functions associated with infinite-dimensional second-order cones". *Nonlinear Analysis. Theory, Methods & Applications. An International Multidisciplinary Journal. Series A: Theory and Methods* 74.16, pp. 5766–5783. DOI: 10.1016/j.na.2011.05.068.

Younes, L. (2012). "Spaces and manifolds of shapes in computer vision: An overview". *Image and Vision Computing* 30.6-7, pp. 389–397. DOI: https://doi.org/10.1016/j.imavis.2011.09.009.

Zolésio, J.-P. (2007). "Control of moving domains, shape stabilization and variational tube formulations". *Control of Coupled Partial Differential Equations*. Springer, pp. 329–382. DOI: 10.1007/978-3-7643-7721-2_15.