Numerical Methods for Optimal Transport and Elastic Shape Optimization

Dissertation
zur Erlangung des Doktorgrades (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Stefan Simon
aus Andernach

Bonn, September 2019

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Martin Rumpf
2. Gutachter: Prof. Dr. Patrick Dondl

Tag der Promotion: 26.11.2019
Erscheinungsjahr: 2019

Abstract

In this thesis, we consider a novel unbalanced optimal transport model incorporating singular sources, we develop a numerical computation scheme for an optimal transport distance on graphs, we propose a simultaneous elastic shape optimization problem for bone tissue engineering, and we investigate optimal material distributions on thin elastic objects.

The by now classical theory of optimal transport admits a metric between measures of the same total mass. Various generalizations of this so-called Wasserstein distance have been recently studied in the literature. In particular, these have been motivated by imaging applications, where the mass-preserving condition is too restrictive. Based on the Benamou–Brenier formulation we present a novel unbalanced optimal transport model by introducing a source term in the continuity equation, which is incorporated in the path energy by a squared L^2-norm in time of a functional with linear growth in space. As a key advantage of our model, this source term functional allows singular sources in space. We demonstrate the existence of constant speed geodesics in the space of Radon measures. Furthermore, for a numerical computation scheme, we apply a proximal splitting algorithm for a finite element discretization.

On discrete spaces, Maas introduced a Benamou–Brenier formulation, where a kinetic energy is defined via an appropriate (*e.g.*, logarithmic) averaging of mass on nodes and momentum on edges. Concerning a numerical optimization scheme, this, unfortunately, couples all these variables on the graph. We propose a conforming finite element discretization in time and prove convergence of corresponding path energy minimizing curves. To apply a proximal splitting algorithm, we introduce suitable auxiliary variables. Besides similar projections as for the classical optimal transport distance and additional simple operations, this allows us to separate the nonlinearity given by the averaging operator to projections onto three-dimensional convex sets, the associated (*e.g.*, logarithmic) cones.

In elastic shape optimization, we are usually concerned with finding a subdomain maximizing the mechanical stability w.r.t. given forces acting onto a larger domain of interest. Motivated by a biomechanical application in bone tissue engineering, where recently biologically degradable polymers have been explored as bone substitutes, we propose a simultaneous elastic shape optimization problem to guarantee stiffness of the polymer implant and of the complementary set where new bone tissue will grow first. Under the assumption that the microstructure of the scaffold is periodic, we optimize a single microcell. We define a novel cost functional depending on specific entries of the homogenized elasticity tensors of polymer and regrown bone. Additionally, the perimeter is penalized for regularizing the interface of the scaffold. For a numerical optimization scheme, we choose a phase-field model, which allows a diffuse approximation of the elastic objects and the perimeter by the Modica–Mortola functional. We also incorporate further biomechanically relevant constraints like the diffusivity of the regrown bone.

Finally, we investigate shape optimization problems for thin elastic objects. For a numerical discretization, we take into account the discrete Kirchhoff triangle (DKT) element for parametric surfaces and approximate the material distribution by a phase-field. To describe equilibrium deformations for a given force, we study different corresponding state equations. In particular, we consider nonlinear elasticity combining membrane and bending models. Furthermore, a special focus is on pure bending isometries, which can be efficiently approximated by the DKT element. We also analyze a one-dimensional model of nonlinear elastic planar beams, where our numerical simulations confirm and extend a theoretical classification result of the optimal design.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Dr. Martin Rumpf for his excellent guidance and his constant support throughout my studies and research, sharing his immense knowledge, introducing me into many interesting topics, and putting trust in my mathematical abilities.

I would like to appreciate the willingness of Prof. Dr. Patrick Dondl for co-reviewing my thesis. I would like to thank him and Dr. Patrina Poh for many fruitful discussions on our joint work on bone tissue engineering.

I am grateful to Prof. Dr. Peter Hornung for inviting me to an exciting workshop in Dresden and providing crucial mathematical input on isometric deformations.

I would like to thank Prof. Dr. Carola Schönlieb, Prof. Dr. Jan Maas, Dr. Matthias Erbar, and Dr. Bernhard Schmitzer for interesting collaborations on projects in the field of optimal transport.

I am also indebted to Prof. Dr. Sergio Conti for his excellent training in applied analysis and giving insightful comments and suggestions on my results in elastic shape optimization.

Furthermore, I would like to thank my colleagues at the Institute for Numerical Simulation, University of Bonn, for providing an always pleasant atmosphere, supporting on any hardware and software problems, and various intense conversations, in particular during relaxing and recharging coffee breaks. I would like to offer my special thanks to Dr. Behrend Heeren, Dr. Alexander Effland, Dr. Martin Lenz, Josua Sassen, Marko Rajković, and Kai Echelmeyer for proofreading parts of my thesis. Besides, I would like to thank Dr. Sascha Tölkes, Dr. Ricardo Perl, Benedict Geihe, Gabi Sodoge-Stork, and Carole Rossignol.

Last but not least, I want to thank my family for their honest love and everlasting support. In particular, I would like to thank my beloved wife Kirsten Simon for her patience, encouragement, and motivation, as well as her final proofreading of my thesis.

Funding My work has been supported by the German Research Foundation (DFG) through the Bonn International Graduate School (BIGS) via the Hausdorff Center for Mathematics (Exzellenzcluster 2047) and the Collaborative Research Center (SFB 1060).

Contents

Abstract **iii**

Acknowledgements **v**

1 Introduction **1**
1.1 Introduction to Optimal Transport 1
1.2 Introduction to Elastic Shape Optimization 3

2 Mathematical Preliminaries **7**
2.1 Radon Measures 7
2.2 Function Spaces 8
2.3 Gamma-Convergence 9

I Numerical Methods for Optimal Transport **11**

3 Foundations in Optimal Transport **13**
3.1 The Classical Optimal Transport Problem 13
3.1.1 Monge's Formulation 13
3.1.2 Kantorovich's Relaxation 14
3.1.3 Benamou–Brenier's Fluid Flow Formulation 15
3.1.4 Wasserstein Gradient Flows 16
3.2 Numerical Methods for the Classical Optimal Transport Problem 16
3.2.1 Overview of Numerical Methods for Optimal Transport 17
3.2.2 Convex Optimization 18
3.2.3 Application of Proximal Splitting Methods to the Flow Formulation 20

4 Optimal Transport with Source Term **23**
4.1 A Benamou–Brenier Formula with Source Term 24
4.2 Relation to Previous Work on Optimal Transport with Source Term 25
4.3 Curves of Radon Measures 26
4.4 Existence of Geodesics for a Generalized Optimal Transport Distance 28
4.4.1 Measure-Valued Formulation of the Path Energy Functional 28
4.4.2 Compactness and Existence Result 30
4.5 Finite Element Discretization 34
4.6 Proximal Splitting Algorithm 35
4.6.1 Projection onto the Set of Solutions to the Generalized Continuity Equation 35
4.6.2 Proximal Mappings of Transport and Source Term Cost 36
4.7 Numerical Results for Generalized Optimal Transport Geodesics 37
4.7.1 Comparison with the L2(L2)-Model 37
4.7.2 Effect of the Source Term Penalization Parameter 39
4.7.3 Application to Textures 42
4.8 Conclusion and Outlook 43

5 Optimal Transport on Graphs **45**
5.1 A Benamou–Brenier Formula on Graphs . . . 46
5.2 A priori Bounds for Mass and Momentum . . . 49
5.3 Finite Element Discretization . . . 50
5.4 Gamma-Convergence of Finite Element Discretization . . . 51
5.5 Optimization with Proximal Splitting . . . 57
5.5.1 Relaxation via Slack Variables . . . 57
5.5.2 Projection onto Solutions to the Continuity Equation . . . 59
5.5.3 Proximal Mapping of the Edge-Based Kinetic Energy . . . 61
5.5.4 Projection onto the Edge-Based Set . . . 62
5.5.5 Proximal Mappings of Auxiliary Operators . . . 65
5.6 Numerical Results for Optimal Transport Geodesics on Graphs . . . 67
5.6.1 Comparison with the Exact Solution for the Two-Node Graph . . . 67
5.6.2 Exploring the Diffuse Behavior on Simple Graphs . . . 68
5.6.3 Experimental Convergence Rate in Time . . . 72
5.6.4 Experimental Results Related to the Gromov–Hausdorff Convergence in Space . . . 72
5.6.5 Discrete Geodesics on Triangular Meshes of a Human Hand . . . 75
5.7 Simulation of the Gradient Flow of the Entropy . . . 76
5.7.1 Adaption of the Numerical Scheme . . . 76
5.7.2 Numerical Results for Gradient Flows . . . 78
5.8 Conclusion and Outlook . . . 79

II Numerical Methods for Elastic Shape Optimization **81**

6 Foundations in Elasticity and Shape Optimization **83**
6.1 Elastic Bodies . . . 83
6.1.1 Hyperelastic Materials . . . 86
6.1.2 Linear Elasticity . . . 87
6.2 Homogenization . . . 88
6.3 Elastic Shape Optimization . . . 90
6.3.1 Perimeter Regularization and Phase-Field Approximation . . . 91
6.3.2 Computing the Shape Derivative . . . 92

7 Simultaneous Elastic Shape Optimization **93**
7.1 Simultaneous Elastic Shape Optimization of a Periodic Microcell . . . 94
7.1.1 State Equations . . . 94
7.1.2 Cost Functional . . . 95
7.2 Hard-Soft Approximation and Perimeter Regularization . . . 96
7.3 Phase-Field Approximation and Finite Element Discretization . . . 97
7.4 Numerical Results for Optimal Periodic Microcells . . . 99
7.4.1 Different Load Scenarios for Equal Material Parameters . . . 100
7.4.2 Influence of the Perimeter Term . . . 101
7.4.3 Influence of Weighting Function . . . 101
7.4.4 Varying Young's Modulus . . . 101
7.4.5 Realistic Material Parameters for Bone and Polymer . . . 103
7.4.6 The Two-Dimensional Case . . . 103
7.5 Extensions of the Model by Diffusion and Volume Constraints . . . 104
7.6 Conclusion and Outlook . . . 106

8 Shape Design of Thin Elastic Objects **107**
8.1 Thin Elastic Shells 108
8.1.1 Differential Geometry for Parametric Surfaces 108
8.1.2 Two-Dimensional Models for Elastic Deformations of Thin Shells 111
8.1.3 Overview of Computational Methods for Thin Elastic Shells 113
8.1.4 Discrete Kirchhoff Triangle Element 115
8.2 Shape Design for Mixed Membrane-Bending Models 116
8.2.1 Shape Optimization Problem for a Phase-Field Approximation 116
8.2.2 Finite Element Discretization for Mixed Membrane-Bending Models 117
8.2.3 Numerical Results for Mixed Membrane-Bending Models 119
8.3 Shape Design for Nonlinear Elastic Beams in 2D 125
8.3.1 State Equation for Nonlinear Elastic Beams in 2D 125
8.3.2 Shape Optimization for Nonlinear Elastic Beams in 2D 126
8.3.3 Phase-Field Approximation and Finite Element Discretization 127
8.3.4 Numerical Results for Nonlinear Elastic Beams in 2D 128
8.4 Shape Design for Bending Isometries of Plates 130
8.4.1 Shape Optimization Problem for Bending Isometries 130
8.4.2 Finite Element Discretization for Bending Isometries 131
8.4.3 Numerical Results for Bending Isometries 132
8.5 Conclusion and Outlook 135

Bibliography **137**

Chapter 1

Introduction

This thesis contains several contributions, which can be categorized into two mathematical research areas, namely optimal transport and shape optimization of elastic objects. Later, the rigorous mathematical foundations, which are in particular required for these specific projects, are discussed in detail in Chapter 3 and Chapter 6. Furthermore, we summarize in Chapter 2 commonly used definitions and well-known theorems, also intending to fix a consistent notation. In the following, we briefly introduce into both fields to give short overviews, including recent developments primarily related to the corresponding topics of this thesis.

1.1 Introduction to Optimal Transport

A Brief History of Optimal Transport. Roughly speaking, the theory of optimal transport is concerned with seeking for the most cost-efficient distribution from a set of sinks to a set of sources. Monge [Mon81] formulated such a problem by asking for the transport with minimal cost of a pile of sand into a hole of the same volume. For a general mathematical formulation, sinks and sources are modeled by probability measures. A relaxed formulation proposed by Kantorovich [Kan42, Kan48] guarantees existence for a certain class of transport cost functions and allows defining the so-called Wasserstein metric on the space of probability measures. Benamou and Brenier [BB00] figured out a dynamical formulation, which can be interpreted as the geodesic equation on the Wasserstein space and thus allows considering it as an infinite-dimensional Riemannian manifold. A groundbreaking result linking the geometry of the Wasserstein space with a partial differential equation was established by Jordan, Kinderlehrer, and Otto [JKO98, Ott01], who demonstrated that the heat equation can be understood as the gradient flow of the entropy functional w.r.t. the Wasserstein distance. Further partial differential equations were characterized via gradient flows of suitable energy functionals w.r.t. the Wasserstein distance, *e.g.*, the Keller-Segel equation [BCC08] or the crowd motion model [MRCS10]. For the incompressible Euler equation, considering a relaxation of Arnold's [Arn66] geodesic formulation in the space of measure-preserving diffeomorphisms, Brenier [Bre89] showed that a midpoint of such a geodesic can be found by solving an optimal transport problem. Furthermore, at first glance, unexpected connections of optimal transport to geometrical questions have emerged. On a Riemannian manifold, the displacement convexity of an appropriate entropy functional along Wasserstein geodesics is equivalent to a nonnegative Ricci curvature. Based on this observation, in the independent works of Lott–Villani [LV09] and Sturm [Stu06a, Stu06b], a meaning of a lower Ricci curvature bound on metric measure spaces was given. Besides numerous proofs, the classical isoperimetric inequality was verified by using techniques from optimal transport [Kno57], which can be applied to prove generalized versions (see, *e.g.*, Figalli and coworkers [FMP10]). Moreover, the optimal transport problem admits a huge variety of applications in the field of mathematical imaging. For image interpolation, it was considered, *e.g.*, for brains and clouds [HZTA04] and in oceanography [HMP15]. By taking into account an appropriate kernel density estimator, it was used for image segmentation in [PFR12]. In [PPC11], the color transfer of images via optimal transport was studied. A decomposition of an image into cartoon, texture, and noise part was investigated in [BL15]. Many problems arising in economy can also be interpreted in the context of optimal transport, *e.g.*, delivering newspapers or matching between job seekers and jobs [Gal16]. Further applications are related to the classification of texts [KSKW15] or an urban planning model [BS05, BW16].

Numerical Methods for Optimal Transport. Computing optimal transport geodesics in its full generality is a quite challenging task. Therefore it has been solved for numerous special cases. In particular, Wasserstein geodesics between probability measures on the real line can be computed explicitly. For discrete measures, Kantorovich's problem becomes a linear program, which can be efficiently solved by the Auction algorithm [BE88]. Benamou and Brenier [BB00] applied duality techniques from convex analysis to compute solutions to their reformulated dynamic problem between density functions. For the so-called semi-discrete optimal transport between a density and a discrete measure, methods from algorithmic geometry were investigated in [Mér11, Lév15]. Further computational methods based on properties of Wasserstein geodesics have been proposed, *e.g.*, in [HZTA04], the polar factorization result by Brenier [Bre91] was used, and in [LR05, BFO10, BFO10], the Monge–Ampère equation was solved. In [Sch16a, Sch16b], a sparse multiscale algorithm was developed by incorporating the cyclical monotonicity property. Recently, entropy regularization methods [BCC$^+$15] to compute approximative solutions have turned out to provide an enormous speedup. Overall, most of the equivalent reformulations of the optimal transport problem can be converted into convex optimization problems. Thus, in this thesis, we intensively apply proximal splitting algorithms based on methods from convex analysis.

Optimal Transport with Source Term. Naturally, the classical optimal transport distance is defined between two measures of the same total mass, which is for example in the Benamou–Brenier formulation encoded via a continuity equation. This mass preserving property is often too restrictive, *e.g.*, in the context of image warping, where images of different total mass have to be compared. Moreover, an extension of the optimal transport distance to arbitrary positive measures is an interesting question from a theoretical point of view, which has been intensively studied in the literature during the last few years. In general, the resulting problems are often referred to as unbalanced optimal transport. One possibility was studied in [CM10], where the marginal constraints in the Kantorovich formulation were relaxed. For absolutely continuous masses a source term in the continuity equation for the Benamou–Brenier formulation was included in [PR16, PR14]. In [CPSV15] and [LMS15], an interpolating distance between the Wasserstein distance and the Fisher–Rao distance was proposed. Recently, in [CPSV18], an equivalence between such generalized optimal transport models based on the Benamou–Brenier formulation and the Kantorovich formulation was demonstrated for a large class of cost functions. In Chapter 4, we study a novel unbalanced optimal transport model on the space of positive Radon measures. There, we adapt the Benamou–Brenier formulation by a source term in the continuity equation, which is appropriately penalized in addition to the kinetic energy, s.t. we can allow singular sources in space. An example of a transport between measures of different total mass is depicted in Figure 1.1.

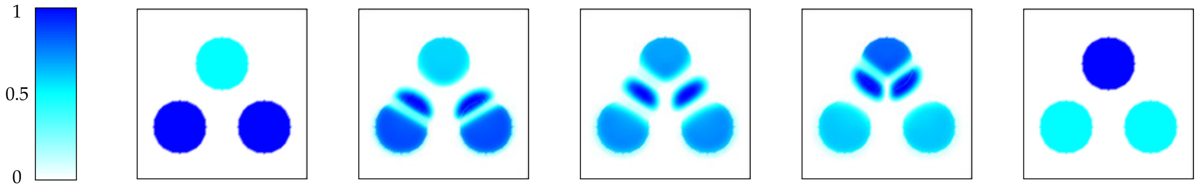


Figure 1.1: Geodesic between densities of different total mass for an optimal transport model with source term. The mass variable is color-coded in a blue-scale (left).

Optimal Transport on Discrete Spaces. The formulations of the optimal transport distance of Monge, Kantorovich, and Benamou–Brenier can be defined without any additional effort between Borel probability measures on complete and separable metric spaces, so-called Polish spaces, and are equivalent under certain conditions. Furthermore, Monge's problem can be considered on more abstract spaces, as far as there is a notion of measures and distance. On discrete spaces described by an irreducible and reversible Markov transition kernel, Maas [Maa11] proposed a Benamou–Brenier formulation, which also allows understanding the heat equation on a finite Markov chain as the gradient flow of a corresponding entropy functional. The associated discrete optimal transport metric does not coincide with Monge's formulation. In Chapter 5, we develop a numerical scheme to compute geodesics and gradient flows for this optimal transport distance on finite Markov chains. For appropriate finite element spaces, we prove convergence of minimizing paths for vanishing step size. In Figure 1.2, we depict an example of such an optimal transport geodesic on a discrete space.

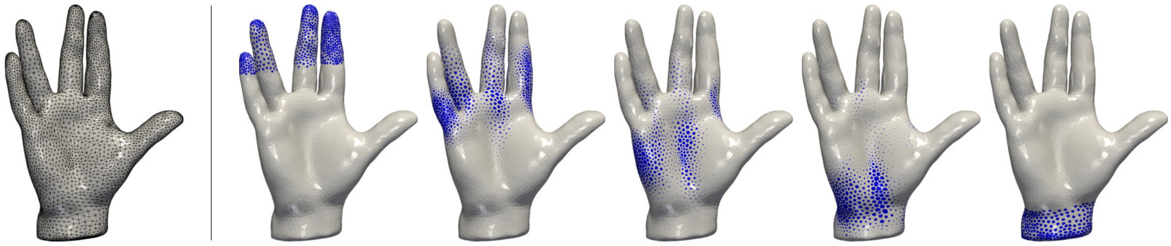


Figure 1.2: Geodesic between discrete measures on a triangular mesh of a human hand (left) for an optimal transport model on graphs. The mass variable, which is actually defined on nodal positions, is represented by blue neighborhoods with an area of a proportional size.

Further Related Work. As we have already mentioned, the fluid flow reformulation by Benamou–Brenier can be interpreted as the geodesic equation on the Wasserstein space. Rumpf and Wirth [RW15] introduced a powerful framework for a time discrete geodesic calculus on Banach manifolds, which allows to approximate geodesics and further differential geometric quantities, like the exponential map, parallel transport, and the Riemannian curvature tensor. This approach was, *e.g.*, applied to the space of viscous fluid objects [RW13], the space of images in the context of the metamorphosis model [BER15], and the space of discrete shells [HRWW12]. In [MRSS15], the general framework by Rumpf and Wirth was used to compute optimal transport geodesics for a viscous optimal transport model with density modulation.

1.2 Introduction to Elastic Shape Optimization

An Overview of Elastic Shape Optimization Problems. Optimizing the mechanical stability of an object is a desirable property in numerous engineering applications. In a general framework of mathematical shape optimization, we ask for the optimal domain within an admissible set, which minimizes a suitable cost functional. Possible applications range from heat diffusion [All02] to fluid dynamics [GHHK15]. Also, the isoperimetric problem can be interpreted as a shape optimization problem, where the area functional has to be minimized over all domains with a fixed volume. In this thesis, we focus on elastic shape optimization problems, where forces are acting on the reference domain of an elastic object and deformations are described via partial differential equations, the so-called state equations. Typical examples of cost functionals studied in the literature are the potential energy [ABFJ97, AJT04], the least square error compared to a target displacement [AJT04], and shape eigenfrequencies [Ped00]. For computational simplicity, in most cases, linear elasticity is taken into account, s.t. the stored elastic, the potential and the free energy coincide for the equilibrium displacement. These three functionals were compared in [PRW12] for nonlinear elasticity, where in particular global minimizers of the free energy do not have to be unique. A worst-case scenario is given by choosing the most expensive of these equilibrium deformations. Usually, the volume of the elastic object is additionally penalized in the cost functional, or a constraint on the maximal amount of volume is integrated into the optimization problem. Nevertheless, such shape optimization problems are in general ill-posed because a minimizing sequence of characteristic functions does not necessarily converge to a characteristic function, and thus the limiting object cannot be characterized as a set. A possible relaxation is based on the theory of homogenization [ABFJ97], where a composite structure determined by its local volume fraction and the effective elasticity tensor is taken into account. Alternatively, in [PRW12], the perimeter of the domain was added to the cost functional. Such a regularization was originally proposed in [AB93] for a scalar-valued problem. A worst-case scenario concerning the uncertainty of a single force acting on the elastic object was studied in [AD14]. For a scenario where multiple loads are acting on the elastic object, several stochastic interpretations to define an associated average cost functional are considered. In the context of a two-stage stochastic programming formulation, in [CHP$^+$08], the expected value was used as compliance functional. Nonlinear risk measures like the expected excess, or the excess probability were investigated in [CHP$^+$11]. In [CRST18], the concept of stochastic dominance was transferred to elastic shape optimization by asking for an object with minimal volume s.t. compared to a given benchmark shape the stochastic dominance constraints given by nonlinear risk measures are satisfied.

Numerical Methods for Elastic Shape Optimization. For a numerical solution scheme to compute an optimal shape, we have to choose a finite-dimensional representation of the elastic object and a corresponding optimization algorithm. A discretization of the elastic object with a finite mesh was, *e.g.*, implemented in [SSW15]. Unfortunately, this requires a remeshing in each optimization step, which is algorithmically quite demanding, especially if the topology of the mesh should change during the optimization process. Level-set functions [CHP$^+$08, CHP$^+$11] to represent the domain by the zero-level set, or phase-field functions [PRW12], which are in particular advantageous to approximate the perimeter functional, have turned out to be more practicable. If the optimal shape is expected to be of a special structure, determining an appropriate set of parameters could simplify the optimization. For example, in [JKZ98], a simple truss model was investigated. For the optimization algorithm, a naive solution scheme is the so-called evolutionary structural optimization (ESO) method [XS93], where, starting on a fixed finite element mesh, those elements with the least contribution to the stiffness are successively removed. Besides, the bi-directional ESO (BESO) [HX10] also allows inserting elements, which might be useful for a fixed volume constraint. However, there is no guarantee that these schemes provide an optimal shape, and, in particular, the solution is mesh-dependent. The homogenization method [ABFJ97] makes use of an explicit formula for an optimal microstructure in linear elasticity, which is given by sequential laminates. Algorithmically, homogenization was used to alternatingly optimize the microstructure and the density on the macroscale. Instead of using the optimal laminate microstructure, the solid isotropic material with penalization (SIMP) method [Ben89] interpolates the material value on each element depending on the density function. In this thesis, we make use of first-order methods, which require to compute the first derivative of the cost functional w.r.t. the elastic object, the so-called shape derivative. This approach was, *e.g.*, applied in [PRW12] for a phase-field model, which we also take into account to discretize the corresponding elastic objects appearing in the specific applications. For a volume constraint, in [AJT04], a Lagrange multiplier was used. A Cahn–Hillard gradient flow with a volume constraint was considered in [ZW07] for a multiphase model. Additional inequality constraints were treated by interior-point methods, *e.g.*, the thickness of trusses in [JKZ98].

Simultaneous Elastic Shape Optimization. The shape optimization problems described above aim to find an optimal subdomain representing the elastic object within a larger domain of interest, which then automatically defines a domain splitting, where the complementary set is considered as void material. In [THD02, TD04], a simultaneous shape optimization problem was investigated by considering the heat conductivity on a subset and the electrical conductivity on the complementary set. More precisely, the optimal domain splitting was sought, s.t. the sum of the traces of the associated homogenized tensors is optimized. For this scalar case, it was conjectured that optimizers are given by domains bounded by periodic minimal surfaces, *e.g.*, the Schwarz P surface. However, in [Sil07] an upper bound for the sum of the traces of the homogenized tensors was derived, which was numerically compared with the corresponding value for a Schwarz P surface and a significant difference to this upper bound was experimentally obtained. In Chapter 7, we propose a similar simultaneous shape optimization problem by taking into account a novel cost functional depending on specific entries of the homogenized elasticity tensors of both subdomains. This formulation is motivated by an application in bone tissue engineering, where biologically degradable polymer implants with a certain microstructure are used as bone substitutes. Incorporating the stiffness of both subdomains in the optimization process guarantees mechanical stability of the polymer implant as well as the regeneration of bone on the complementary set. Furthermore, we adapt the model by additional biologically relevant constraints. In particular, we enforce diffusion constraints on the regrown bone. We show possible optimized periodic microstructures in Figure 1.3.

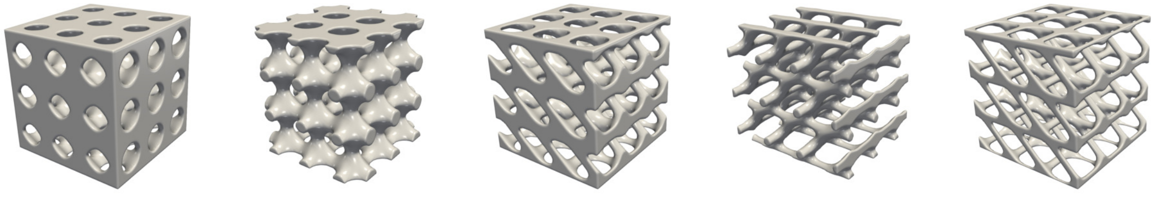


Figure 1.3: Optimized periodic microstructures for bone tissue engineering (here for different material parameters of regrown bone).

Shape Design of Thin Elastic Objects. Thin elastic objects are a special class of curved elastic bodies, which are significantly smaller in one direction. Such thin structures frequently appear in aerodynamics [HZ14, KPRA18, SS13], where in particular airfoils are optimized w.r.t. the aerodynamic drag. Further applications can be found in electrostatics [BCO$^+$15] and automotive engineering [Ble14]. From a theoretical point of view, the behavior of these thin elastic objects has been well-understood via Γ-convergence results for vanishing thickness. Different scalings lead to a membrane theory [LDR95, LDR96] describing tangential distortion on the surface and a bending theory [FJM02, FJMM03] taking into account isometric deformations. In numerical simulations, the corresponding elastic energies have often been combined. Numerous discretization methods have been proposed to approximate thin elastic objects and their deformations, where the essential difficulty is due to curvature terms in the bending energy functional involving second derivatives of the deformation. On quadrilateral meshes, nonuniform rational B-splines (NURBS) [HCB05] allow arbitrary regularity. A fully conforming discretization on triangular meshes is given by loop subdivision finite elements [COS00]. In practice, methods from discrete differential geometry have turned out to be extremely efficient [GHDS03]. To simulate pure bending isometries on plates, in [Bar13] a numerical approximation scheme was provided by making use of the discrete Kirchhoff triangle (DKT) element [BBH80]. The optimal design of shells via composite material lamination was considered in [SL05]. The finite mesh itself was optimized in [BC18] by taking into account loop subdivision surfaces and linear elasticity as in [COS00]. In [VHWP12], NURBS were investigated to construct self-supporting surfaces. In Chapter 8, we study shape optimization problems for thin elastic objects. To describe a material distribution, we use a phase-field discretization. Then we investigate different elastic energies, in particular, nonlinear elasticity and an isometry constraint. In Figure 1.4, we depict optimal designs.

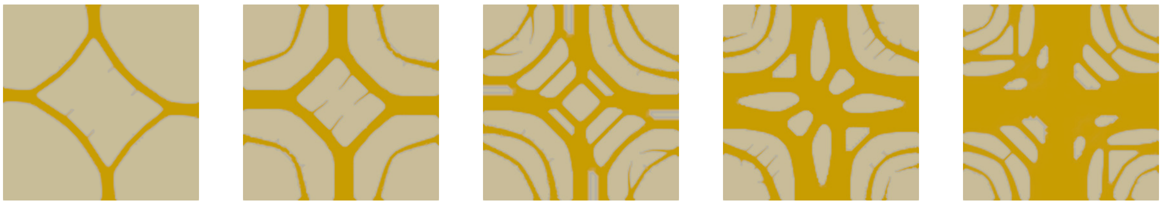


Figure 1.4: Optimal material distributions on a thin plate under certain volume conditions (here for different volume constraints). The hard material is colored in orange.

Chapter 2

Mathematical Preliminaries

The following chapter is mainly considered to fix overall terminology and notation. In this thesis, we investigate many different objects, *e.g.*, images, graphs, rods, plates, shells, and solids. For mathematical modeling of these objects, we take into account different function spaces, possibly also including a time component. Especially, we make use of the space of Radon measures, which we introduce in Section 2.1. Further relevant function spaces are defined in Section 2.2. Finally, we consider the concept of Γ-convergence in Section 2.3, which plays an important role throughout this thesis. For a more detailed introduction, we refer the reader to the books [FL07], [EG15], [AFP00], [Alt16] for functional analysis and [Bra06], [DM93] for Γ-convergence.

2.1 Radon Measures

In the following, we define the space of Radon measures and summarize some essential properties. We start to recall basics from measure theory. In particular, we define measures on a generic set X with a σ-algebra $\mathcal{E} \subset \mathcal{P}(X)$.

Definition 2.1.1 (Measures and Total Variation)**.** Let X be a nonempty set.

1. On a measure space $(X, \mathcal{E})$ a map $\mu\colon \mathcal{E} \to [0,\infty]$ is a positive measure if $\mu(\emptyset) = 0$ and μ is σ-additive on $\mathcal{E}$. If the same holds for a map $\nu\colon \mathcal{E} \to \mathbb{R}$, we call it a signed measure. Moreover, $\nu\colon \mathcal{E} \to \mathbb{R}^m$ with $m \in \mathbb{N}_+$ is a vectorial measure if each component is a signed measure.

2. Let $\nu\colon \mathcal{E} \to \mathbb{R}$ be a signed measure. Then the total variation $|\nu|_{TV}$ for $E \in \mathcal{E}$ is given by

$$|\nu|_{TV}(E) := \sup \left\{ \sum_{n\in\mathbb{N}} |\nu(E_n)| \ : \ E = \bigcup_{n\in\mathbb{N}} E_n \text{ for } E_n \in \mathcal{E} \text{ pairwise disjoint} \right\}$$

   and defines a positive and finite measure (see [AFP00, Theorem 1.6]). For a vectorial measure $\nu\colon \mathcal{E} \to \mathbb{R}^m$ we define the total variation by $|\nu|_{TV}(E) := \sum_{i=1}^m |\nu_i|_{TV}(E)$.

We remark that there are different terminologies used in the literature, where a measure might denote either a positive or a signed measure. Furthermore, some approaches are based on so-called outer measures, which are defined on arbitrary subsets (*e.g.*, in [EG15]).

Now, to define Radon measures, some topological information on the set X is required. Then, we denote by $\mathscr{B}(X)$ the Borel σ-algebra, which is defined as the smallest σ-algebra on X containing all open sets. Due to our applications, we restrict to the case that $X = D \subset \mathbb{R}^d$ is a subset of $\mathbb{R}^d$.

Definition 2.1.2 (Radon Measures)**.** Consider the measure space $(D, \mathscr{B}(D))$ for $D \in \mathscr{B}(\mathbb{R}^d)$.

1. A positive measure $\mu\colon \mathscr{B}(D) \to [0,\infty]$ is a positive Radon measure if $\mu(K) < \infty$ for all $K \subset D$ compact. A signed measure $\nu\colon \mathscr{B}(D) \to \mathbb{R}$ is a signed Radon measure if $|\nu|_{TV}$ is a positive Radon measure and a vectorial measure $\nu\colon \mathscr{B}(D) \to \mathbb{R}^m$ is a vectorial Radon measure if each component is a signed Radon measure.

2. We denote by
   (a) $\mathscr{M}^+(D)$ the set of all positive Radon measures,
   (b) $\mathscr{M}(D)$ the set of all signed Radon measures, and
   (c) $\mathscr{M}(D, \mathbb{R}^m)$ the set of all vectorial Radon measures.

In the following, we further restrict to a compact set $D \subset \mathbb{R}^d$. Then a positive Radon measure is just a finite Borel measure and thus a signed measure. An important characterization of Radon measures is given by the following duality result, which allows us to identify the space of signed Radon measures $\mathscr{M}(D)$ as the topological dual of the space of continuous functions $C(D)$ endowed with the norm $\|f\|_{C(D)} := \sup_{x \in D} |f(x)|$.

Theorem 2.1.3 (Duality of Radon Measures)**.** *Let $D \subset \mathbb{R}^d$ be a compact set. Then every bounded linear functional $L\colon C(D) \to \mathbb{R}$ is represented by a unique signed Radon measure $\nu \in \mathscr{M}(D)$ in the sense that*

$$L(f) = \int_D f \,\mathrm{d}\nu \quad \forall f \in C(D)\,. \tag{2.1}$$

Conversely, every functional L of type (2.1) *for $\nu \in \mathscr{M}(D)$ is a bounded linear functional on $C(D)$.*

Proof. See [FL07, Theorem 1.196]. □

Hence, a sequence of Radon measures $(\nu_n)_{n \in \mathbb{N}} \subset \mathscr{M}(D)$ converges weakly-* to $\nu \in \mathscr{M}(D)$ if

$$\int_D f \,\mathrm{d}\nu_n \to \int_D f \,\mathrm{d}\nu \quad \forall f \in C(D)\,.$$

Furthermore, since $C(D)$ is a separable space, every bounded sequence $(\nu_n)_{n \in \mathbb{N}} \subset \mathscr{M}(D)$ of signed Radon measures has a weakly-* converging subsequence (see [Alt16, Theorem 8.5]).

2.2 Function Spaces

Here, we summarize several properties of Sobolev functions and functions of bounded variation.

In the following, let $D \subset \mathbb{R}^d$ be a domain. First, for k-times continuously differentiable functions $f, g \in C^k(\overline{D})$ with $k \in \mathbb{N}_+$, we define $D^k f \cdot D^k g := \sum_{i_1,\dots,i_k=1,\dots,k} \partial^k_{i_1,\dots,i_k} f \, \partial^k_{i_1,\dots,i_k} g$ and $|D^k f| := \left(D^k f \cdot D^k f\right)^{\frac{1}{2}}$.

Lebesgue and Sobolev Functions For a measurable function $f\colon D \to \mathbb{R}^d$, we recall the norms

$$\begin{aligned}
\|f\|_{L^p(D)} &:= \left(\int_D |f(x)|^p \,\mathrm{d}x\right)^{\frac{1}{p}} && \text{for } p \in [1, \infty)\,,\\
\|f\|_{L^\infty(D)} &:= \operatorname*{ess\,sup}_{x \in D} |f(x)| = \inf\{C \geqslant 0 \ :\ |f(x)| \leqslant C \text{ for } \textit{a.e. } x \in D\}\,, &&\\
\|f\|_{W^{m,p}(D)} &:= \left(\sum_{k=0}^{m} \|D^k f\|^p_{L^p(D)}\right)^{\frac{1}{p}} && \text{for } m \in \mathbb{N},\ p \in [1, \infty)\,,\\
\|f\|_{W^{m,\infty}(D)} &:= \max_{k=0,\dots,m} \|D^k f\|_{L^\infty(D)} && \text{for } m \in \mathbb{N}\,,
\end{aligned}$$

where the derivatives appearing in the definitions of the Sobolev norms $\|\cdot\|_{W^{m,p}(D)}$ for $p \in [1, \infty]$ have to be understood in the distributional sense.

We say that $D \subset \mathbb{R}^d$ has Lipschitz boundary, if for all $x \in \partial D$ there exists a neighborhood U of x and a Lipschitz function $L\colon \mathbb{R}^{d-1} \to \mathbb{R}$ s.t. $D \cap U = \{y = (y_1, \dots, y_d) \in U \ :\ y_d > L(y_1, \dots, y_{d-1})\}$. Then, we define the space $W^{m,p}_0(D)$ as the closure of $C^\infty_c(D)$ w.r.t. the $W^{m,p}(D)$-norm.

Later, we make use of the following two theorems.

Theorem 2.2.1 (Korn's Inequality)**.** *Let $D \subset \mathbb{R}^d$ be a domain with Lipschitz boundary. Then there is a constant $c > 0$ s.t.*

$$\|Du\|^2_{L^2(D)} \leqslant c\left(\|u\|^2_{L^2(D)} + \|\varepsilon(u)\|^2_{L^2(D)}\right) \tag{2.2}$$

for all $u \in W^{1,2}(D, \mathbb{R}^n)$. Here, $\varepsilon(u) := \frac{Du+Du^T}{2}$ denotes the symmetrized gradient.

Proof. See [Nit81]. □

Theorem 2.2.2 (Sobolev Embedding)**.** *Let $D \subset \mathbb{R}^d$ be a domain with Lipschitz boundary.*

1. *Let $m_1 > m_2 \in \mathbb{N}$ and $p_1, p_2 \in [1, \infty)$ with $m_1 - \frac{d}{p_1} > m_2 - \frac{d}{p_2}$.*
   Then the embedding $\mathrm{id}\colon W^{m_1,p_1}(D) \to W^{m_2,p_2}(D)$ is continuous and compact.
2. *Let $m \in \mathbb{N}_+$, $k \in \mathbb{N}$, $p \in [1, \infty)$, and $\alpha \in [0, 1]$ s.t. $m - \frac{d}{p} > k + \alpha$.*
   Then the embedding $\mathrm{id}\colon W^{m,p}(D) \to C^{k,\alpha}(\overline{D})$ is continuous and compact.

Proof. See [Alt16, Theorem 10.9 and Theorem 10.13]. □

Functions of Bounded Variation Next, we introduce the space of functions of bounded variation.

Definition 2.2.3 (Functions of Bounded Variation)**.** Let $D \subset \mathbb{R}^d$ be a domain.

1. The space of functions of bounded variation is defined by

$$BV(D) := \{u \in L^1(D) \;:\; Du \in \mathscr{M}(D, \mathbb{R}^d) \text{ for the distributional gradient}\}.$$

2. For $u \in BV(D)$ the norm is given by $\|u\|_{BV(D)} := \|u\|_{L^1(D)} + |Du|_{TV}(D)$.
3. For a sequence $u_k \in BV(D)$ and $u \in BV(D)$ we say that u_k converges weak-* to u in BV if $u_k \to u$ strongly in $L^1(D)$ and $Du_k \overset{*}{\rightharpoonup} Du$ in $\mathscr{M}(D, \mathbb{R}^d)$.

Then the following embedding theorem holds.

Theorem 2.2.4 (Embedding in BV)**.** *Let $D \subset \mathbb{R}^d$ be a domain with Lipschitz boundary and let $1 \leqslant p < \frac{d}{d-1}$. Then the embedding $\mathrm{id}\colon BV(D) \to L^p(D)$ is continuous and compact.*

Proof. See [AFP00, Theorem 3.47]. □

2.3 Γ-Convergence

Many problems appearing in this thesis result in minimizing an energy functional $\mathcal{E}\colon X \to \mathbb{R} \cup \{\infty\}$ on some metric space X. Typically, to approximate a minimizer of $\mathcal{E}$, we take into account a finite space $X_h \subset X$ and a suitable functional $\mathcal{E}_h\colon X_h \to \mathbb{R} \cup \{\infty\}$, s.t. we can numerically compute a minimizer of $\mathcal{E}_h$. Further functionals considered in this thesis similarly arise as limits of functionals $\mathcal{E}_h : X_h \to \mathbb{R} \cup \{\infty\}$ for $h \to 0$. However, the convergence of $\mathcal{E}_h \to \mathcal{E}$ in a common topology of the functionals is a too strong requirement, but we are only interested in the convergence of the minimizers of $\mathcal{E}_h$ to the minimizer of $\mathcal{E}$. This can be established by using the concept of Γ-convergence.

Definition 2.3.1 (Γ-Convergence)**.** Let (X, d) be a metric space and $\mathcal{E}_k\colon X \to \mathbb{R} \cup \{\infty\}$ for $k \in \mathbb{N}$. We say that the sequence of functionals $(\mathcal{E}_k)_{k\in\mathbb{N}}$ Γ-converges to a functional $\mathcal{E}\colon X \to \mathbb{R} \cup \{\infty\}$ if

1. the Γ-liminf condition holds, *i.e.*, for all $(x_k)_{k\in\mathbb{N}} \subset X$ with $x_k \to x \in X$ we have

$$\mathcal{E}(x) \leqslant \liminf_{k\to\infty} \mathcal{E}_k(x_k)\,, \tag{2.3}$$

2. and the Γ-limsup condition holds, *i.e.*, for all $x \in X$ there exists a sequence $(x_k)_{k\in\mathbb{N}} \subset X$ with $x_k \to x$ s.t.

$$\mathcal{E}(x) \geqslant \limsup_{k\to\infty} \mathcal{E}_k(x_k)\,. \tag{2.4}$$

Note that (2.3) implies that we actually have equality in (2.4). The sequence satisfying $\mathcal{E}(x) = \lim_{k\to\infty} \mathcal{E}_k(x_k)$ is called a recovery sequence.

Definition 2.3.2 (Equicoercivity)**.** Let (X, d) be a metric space and $\mathcal{E}_k \colon X \to \mathbb{R} \cup \{\infty\}$ for $k \in \mathbb{N}$. The sequence of functionals $(\mathcal{E}_k)_{k\in\mathbb{N}}$ is equicoercive if for all $r \in \mathbb{R}$ there is a compact set $K_r \subset X$ s.t. $\{x \in X \,:\, \mathcal{E}_k(x) \leqslant r \ \forall k \in \mathbb{N}\} \subset K_r$.

Thus, for a sequence $(x_k)_{k\in\mathbb{N}}$ with uniformly bounded energy $\mathcal{E}_k(x_k) \leqslant r$, the equicoercivity condition implies convergence of a subsequence $x_{k_l} \to x \in X$. Together with the Γ-convergence, this guarantees that minimizers of $\mathcal{E}_k$ converge to a minimizer of $\mathcal{E}$.

Theorem 2.3.3 (Fundamental Theorem of Γ-Convergence)**.** *Let (X, d) be a metric space and $\mathcal{E}_k \colon X \to \mathbb{R} \cup \{\infty\}$ for $k \in \mathbb{N}$. We assume that the sequence $(\mathcal{E}_k)_{k\in\mathbb{N}}$ is equicoercive and Γ-converges to $\mathcal{E} \colon X \to \mathbb{R} \cup \{\infty\}$. Then*

$$\min_{x\in X} \mathcal{E}(x) = \lim_{k\to\infty} \inf_{x\in X} \mathcal{E}_k(x)\,.$$

Proof. See [Bra06, Theorem 2.10]. □

Consequently, if a sequence $x_k \to x^*$ is asymptotically minimizing, *i.e.*, it satisfies $\mathcal{E}_k(x_k) = \inf_{x\in X} \mathcal{E}_k(x) + o(1)$, then x^* is a minimizer of $\mathcal{E}$.

Later, we make use of the following lower semi-continuity result, which allows proving the Γ-liminf inequality (2.3) for a large class of functionals.

Theorem 2.3.4 (Ioffe)**.** *Let $D \subset \mathbb{R}^d$ be open and bounded. Furthermore, let $f \colon D \times \mathbb{R}^{p+q} \to [0, \infty]$ be a measurable function, s.t. $(s, z) \mapsto f(x, s, z)$ is lower semi-continuous for a.e. $x \in D$ and $z \mapsto f(x, s, z)$ is convex for any $x \in D$ and any $s \in \mathbb{R}^p$. Then, for sequences $u_h \to u$ strongly in $L^1(D, \mathbb{R}^p)$ and $v_h \to v$ weakly in $L^1(D, \mathbb{R}^q)$, we have*

$$\int_D f(x, u, v)\,\mathrm{d}x \leqslant \liminf_{h\to 0} \int_D f(x, u_h, v_h)\,\mathrm{d}x\,.$$

Proof. See [AFP00, Theorem 5.8]. □

Part I

Numerical Methods for Optimal Transport

Chapter 3

Foundations in Optimal Transport

The first part of this thesis is concerned with two different types of optimal transport distances, where we primarily focus on numerical methods to compute corresponding geodesic interpolation paths. In this chapter, we first give an introduction to the basic theory of optimal transport and related mathematical foundations. In particular, in Section 3.1, we define the L^2-Wasserstein distance on the space of Borel probability measures. To compute solutions for this classical optimal transport distance, various algorithms have been developed, which we partially summarize in Section 3.2. Furthermore, since in Chapter 4 and Chapter 5 we intensively make use of so-called proximal splitting methods, we collect the related concepts from convex analysis.

3.1 The Classical Optimal Transport Problem

In the following, we introduce three different formulations of the optimal transport problem, namely those of Monge, Kantorovich, and Benamou–Brenier. For transport costs given by the Euclidean distance, this leads us to the corresponding Wasserstein metric on the space of Borel probability measures. Moreover, we briefly discuss gradient flows on the Wasserstein space and the fundamental connection to the heat equation. For a more general overview of the theory of optimal transport, we refer the reader to the well-established books [AGS08, San15, Vil03, Vil09].

3.1.1 Monge's Formulation

A first version of the optimal transport problem was already formulated in 1781 by Monge [Mon81], who asked for the minimal cost to transport a pile of sand into a hole of the same volume. For a mathematical model, source and sink are described by Borel probability measures $\mu_A \in \mathscr{P}(X)$ and $\mu_B \in \mathscr{P}(Y)$, where we restrict to the case that $X, Y \subset \mathbb{R}^d$ are compact sets. To define a transport of the mass represented by the measure μ_A, we take into account a transport map $T\colon X \to Y$. Then, to guarantee that the mass is transported by T to a distribution corresponding to the measure μ_B, a matching condition is required.

Definition 3.1.1 (Pushforward)**.** Let $\mu \in \mathscr{P}(X)$ and $T\colon X \to Y$ Borel measurable. We define the pushforward $T_\# \mu$ of μ through T as

$$T_\#\mu(E) := \mu(T^{-1}(E)) \quad \text{for all } E \in \mathscr{B}(Y)\,. \tag{3.1}$$

We say that a transport map T matches μ_A to μ_B if $T_\#\mu_A = \mu_B$. Moreover, a transport cost function $c\colon X \times Y \to [0,\infty]$ describes the cost to move a particle from a position $x \in X$ to a position $y \in Y$. Then Monge's problem in its general formulation is to find a transport map T having minimal transport cost, which is given by

$$\inf\left\{ \int_X c(x, T(x))\,\mathrm{d}\mu_A(x) \;:\; T\colon X \to Y \text{ Borel measurable}\,, T_\#\mu_A = \mu_B \right\}. \tag{3.2}$$

We focus on the case that $X = Y = D$ for a compact and convex domain $D \subset \mathbb{R}^d$. Note that the set of all Borel probability measures on D is defined as a subset of positive Radon measures

$$\mathscr{P}(D) = \{\mu \in \mathscr{M}^+(D) \;:\; \mu(D) = 1\}\,.$$

By duality of Radon measures (see Theorem 2.1.3), the matching condition (3.1) is equivalent to

$$\int_D f(T(x))\,\mathrm{d}\mu_A(x) = \int_D f(x)\,\mathrm{d}\mu_B(x) \quad \forall f \in C(D)\,.$$

Furthermore, we restrict the transport cost function to the Euclidean distance $c(x,y) = |x-y|^2$. Because of the convexity assumption on the domain D, the distance on D is induced from the distance on $\mathbb{R}^d$.

Remark 3.1.2. For nonconvex domains, we could take into account the path length from x to y. More generally, we could define Monge's problem for underlying smooth manifolds or even on separable complete metric spaces, so-called Polish spaces, by using the squared distance as transport cost. For noncompact domains, we have to restrict the space of Borel probability measures to guarantee that the integral $\int_X c(x,T(x))\,\mathrm{d}\mu_A(x)$ is finite. A sufficient condition in the case of the cost function $c(x,y) = |x-y|^2$ is to require bounds on the second moment, *i.e.*, $\mu_A, \mu_B \in \mathscr{P}_2(X) := \{\mu \in \mathscr{P}(X) \,:\, \int_X |x|^2\,\mathrm{d}\mu(x) < \infty\}$.

Unfortunately, Monge's problem (3.2), in general, does not admit existence nor uniqueness.

Example 3.1.3 (Nonexistence and Nonuniqueness for Monge's Problem)**.**

1. Let $D = [-1,1]$, $\mu_A = \delta_0$ and $\mu_B = \frac{1}{2}(\delta_{-1} + \delta_1)$, where δ_p denotes the Dirac measure at the point $p \in D$. Then there does not exist a transport map T between μ_A and μ_B, since otherwise

$$f(T(0)) = \int_D f(x)\,\mathrm{d}T_\#\mu_A(x) = \int_D f(x)\,\mathrm{d}\mu_B(x) = \frac{1}{2}(f(-1) + f(1))$$

   for all $f \in C(D)$. In other words, we cannot split a single point.

2. Let $D = [0,1]^2$, $\mu_A = \frac{1}{2}(\delta_{(0,0)} + \delta_{(1,1)})$ and $\mu_B = \frac{1}{2}(\delta_{(1,0)} + \delta_{(0,1)})$. Then an optimal transport map could map $(0,0)$ to $(0,1)$ and $(1,1)$ to $(1,0)$, but also the opposite way is optimal.

3.1.2 Kantorovich's Relaxation

To cope with the existence problem, Kantorovich [Kan42, Kan48] proposed a relaxation of Monge's formulation by embedding the transport map $T\colon D \to D$ between μ_A and μ_B into the product space $D \times D$ by considering a so-called transport plan $\pi = (\mathrm{id} \times T)_\#\mu_A$. Since T fulfills the pushforward matching condition (3.1), the transport plan π satisfies the marginal constraints

$$(\mathrm{proj}_1)_\#\pi = \mu_A \quad \text{and} \quad (\mathrm{proj}_2)_\#\pi = \mu_B\,,$$

where proj_i for $i = 1,2$ denotes the projection on the i-th component. More generally, we define the set of all Borel probability measures on the product space with marginal constraints by

$$\Pi(\mu_A,\mu_B) = \left\{\pi \in \mathscr{P}(D\times D) \,:\, (\mathrm{proj}_1)_\#\pi = \mu_A\,,\ (\mathrm{proj}_2)_\#\pi = \mu_B\right\}\,.$$

Then Kantorovich's problem is given by

$$\inf\left\{\int_{D\times D} c(x,y)\,\mathrm{d}\pi(x,y) \,:\, \pi \in \Pi(\mu_A,\mu_B)\right\} \tag{3.3}$$

and the following existence result holds.

Theorem 3.1.4 (Existence of Solutions)**.** *Suppose that $c\colon D\times D \to \mathbb{R} \cup \{\infty\}$ is lower semi-continuous and bounded from below. Then Kantorovich's problem* (3.3) *admits a solution.*

Proof. See [San15, Theorem 1.5]. □

Under the condition that the initial measure μ_A is absolutely continuous w.r.t. the Lebesgue measure on D, uniqueness of the optimal transport plan can be established by applying Brenier's polar factorization result [Bre91], which allows decomposing a density function into a gradient of a convex function up to a concatenation with a measure-preserving map. In this case, the solution to Monge's and Kantorovich's problem coincide.

Theorem 3.1.5 (Brenier's Polar Factorization). *We consider the specific transport cost function $c(x,y) = |x-y|^2$. Let $\mu_A, \mu_B \mathscr{P}(D)$ with $\mu_A = \rho_A \mathscr{L}_D$ for a density function ρ_A. Then there exists a unique optimal transport map T solving Monge's problem, and $T = \nabla\psi$ is the μ_A-a.e. unique gradient of a convex function ψ. Moreover, the unique optimal transport plan solving Kantorovich's problem is given by $\pi = (\mathrm{id} \times \nabla\psi)_\# \mu_A$.*

Proof. See [San15, Theorem 1.22]. □

In the case $c(x,y) = |x-y|^2$, the relaxed problem (3.3) defines a metric on the space of Borel probability measures, the so-called L^2-Wasserstein distance.

Definition 3.1.6 (Wasserstein Distance). Let $\mu_A, \mu_B \in \mathscr{P}(D)$ be two Borel probability measures. We define the L^2-Wasserstein distance $\mathcal{W}$ between μ_A and μ_B by

$$\mathcal{W}(\mu_A, \mu_B) := \inf \left\{ \int_{D\times D} |x-y|^2 \,\mathrm{d}\pi(x,y) \ : \ \pi \in \Pi(\mu_A, \mu_B) \right\}^{\frac{1}{2}} . \tag{3.4}$$

We refer the reader to [San15, Proposition 5.10] for a proof that $\mathcal{W}$ is indeed a metric on $\mathscr{P}(D)$. Moreover, $\mathcal{W}$ metrizes weak-*-convergence on $\mathscr{P}(D)$ (see [San15, Theorem 5.10]). Regarding a numerical optimization scheme to compute an optimal transport plan solving Kantorovich's problem (3.3), it is useful to consider the corresponding dual formulation

$$\sup \left\{ \int_D f(x) \,\mathrm{d}\mu_A(x) + \int_D g(y) \,\mathrm{d}\mu_B(y) \ : \ (f,g) \in C(D) \times C(D),\ f(x) + g(y) \leqslant c(x,y) \right\} .$$

3.1.3 Benamou–Brenier's Fluid Flow Formulation

In [BB00], Benamou and Brenier transferred Monge's problem into a continuum mechanics framework and derived an equivalent representation of the Wasserstein distance (3.4) heuristically. This dynamical formulation takes into account a curve of probability measures $\mu\colon [0,1] \to \mathscr{P}(D)$ connecting $\mu(0) = \mu_A$ with $\mu(1) = \mu_B$ and a corresponding Eulerian velocity field $v\colon [0,1] \times D \to \mathbb{R}^d$. Here, we assume that μ is a curve of probability densities ρ, *i.e.*, $\mu(t) = \rho(t)\mathscr{L}$ for all $t \in [0,1]$. Then we can formally define the kinetic energy

$$\mathcal{E}_{\text{trans}}(\rho, v) = \int_0^1 \int_D \rho(t,x) |v(t,x)|^2 \,\mathrm{d}x \,\mathrm{d}t .$$

Furthermore, a mass-preserving condition is given by the continuity equation $\partial_t \rho + \mathrm{div}(\rho v) = 0$, *i.e.*, solutions to this equation satisfy $\int_D \rho(t,x) \,\mathrm{d}x = \int_D \rho(0,x) \,\mathrm{d}x$ for all $t \in [0,1]$. We denote by $\mathcal{CE}(\rho_A, \rho_B)$ the set of all weak solutions (ρ, v) of the continuity equation with initial condition $\rho(0) = \rho_A$ and final condition $\rho(1) = \rho_B$. It turns out that Monge's formulation (3.2) of the optimal transport problem can be rewritten by minimizing the kinetic energy over all corresponding curves of mass and velocity, which solve the continuity equation, *i.e.*,

$$\mathcal{W}(\rho_A \mathscr{L}, \rho_B \mathscr{L}) = \inf \{ \mathcal{E}_{\text{trans}}(\rho, v) \ : \ (\rho, v) \in \mathcal{CE}(\rho_A, \rho_B) \}^{\frac{1}{2}} . \tag{3.5}$$

To rigorously formulate (3.5) on appropriate function spaces, the curve μ is required to be absolutely continuous in time. Moreover, the continuity equation has to be defined in a weak sense. Then, one possibility (see, *e.g.*, [AGS08, Chapter 8]) is to define the velocity at time t in a the function space depending on the measure $\mu(t)$ at the specific time. Later, we apply a different approach by making use of a change of variables. Instead of the pair mass and velocity (ρ, v), we consider the pair mass and momentum $(\rho, m = \rho v)$. Then it can be shown that the distance defined by the Benamou–Brenier formulation coincides with the Wasserstein distance (see [San15, Theorem 5.28]). Furthermore, for an absolutely continuous initial measure $\mu_A = \rho_A \mathscr{L}_D$ and an optimal transport plan $\pi = (\mathrm{id} \times \nabla\psi)_\# \mu_A$ as in Theorem 3.1.5, the linear interpolation of the identity and the optimal transport map $\nabla\psi$ under the pushforward w.r.t. μ_A

$$\mu(t) = ((1-t)\mathrm{id} + t\nabla\psi)_\# \,\mu_A = (\mathrm{id} + tv)_\# \,\mu_A$$

is the solution to Benamou–Brenier's problem and satisfies the property of a constant speed geodesic

$$\mathcal{W}(\mu(s), \mu(t)) = |t-s| \,\mathcal{W}(\mu_A, \mu_B) \quad \forall s,t \in [0,1] .$$

For now, we consider the definition of $\mathcal{W}$ in (3.5) just formally and refer to Chapter 4 for a rigorous formulation of a generalized optimal transport distance.

3.1.4 Wasserstein Gradient Flows

In [JKO98], a fundamental connection between gradient flows w.r.t. the Wasserstein metric on $\mathbb{R}^d$ and the heat equation was established. First, we recall that for a function $F \in C^{1,1}(\mathbb{R}^d, \mathbb{R})$, the solution to the Cauchy problem

$$\begin{cases} x'(t) = -\nabla F(x(t)) & \text{for } t > 0\,, \\ x(0) = x_0 \end{cases}$$

can be approximated by an implicit Euler scheme

$$\begin{aligned} x_0^\tau &= x_0\,, \\ x_{k+1}^\tau &\in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} F(x) + \frac{|x - x_k^\tau|^2}{2\tau} \quad \text{for } k \in \mathbb{N}\,, \end{aligned} \tag{3.6}$$

where $\tau > 0$ is a fixed step size. Now, we define the entropy functional $\mathcal{H} \colon L^1(\mathbb{R}^d, [0, \infty]) \to \mathbb{R} \cup \{\infty\}$ by

$$\mathcal{H}(\rho) = \int_{\mathbb{R}^d} \rho(x) \log(\rho(x))\, \mathrm{d}x\,.$$

More generally, for a smooth potential V, we consider the functional $\mathcal{F}(\rho) = \mathcal{H}(\rho) + \int_{\mathbb{R}^d} V(x)\rho(x)\, \mathrm{d}x$. Motivated by the finite-dimensional and smooth case in (3.6), the so-called minimizing movement scheme is defined by the iteration

$$\begin{aligned} \rho_0^\tau &= \rho_0\,, \\ \rho_{k+1}^\tau &\in \operatorname*{arg\,min}_{\mu \in \mathscr{P}_2(\mathbb{R}^d) \,:\, \mu = \rho \mathscr{L}} \mathcal{F}(\rho) + \frac{1}{2\tau} \mathcal{W}(\rho, \rho_k^\tau)^2 \quad \text{for } k \in \mathbb{N}\,. \end{aligned} \tag{3.7}$$

It was shown in [JKO98, Proposition 4.1] that for an absolutely continuous initial condition, there is a unique discrete solution trajectory $(\rho_k^\tau)_{k \in \mathbb{N}}$. Furthermore, in the limit $\tau \to 0$, the following interpretation as the Wasserstein gradient flow of $\mathcal{F}$ was given.

Theorem 3.1.7 (Gradient Flow of Entropy). *Given $\mu_0 \in \mathscr{P}(\mathbb{R}^d)$ with $\mu_0 = \rho_0 \mathscr{L}^d$ and $\mathcal{F}(\rho_0) < \infty$. Let $(\rho_k^\tau)_{k \in \mathbb{N}}$ be the discrete solution trajectory obtained by* (3.7) *and define $\rho^\tau(t, x) = \rho_k^\tau(x)$ for $t \in [k\tau, (k+1)\tau)$. Then $\rho^\tau \rightharpoonup \rho$ in $L^1(\mathbb{R}_+ \times \mathbb{R}^d)$ for $\tau \to 0$, where $\rho \in C^\infty((0, \infty) \times \mathbb{R}^d)$ is the unique solution to the Fokker-Planck equation $\partial_t \rho - \Delta \rho - \operatorname{div}(\rho \nabla V) = 0$ with $\rho(t) \to \rho_0$ in L^1 for $t \to 0$.*

Proof. See [JKO98, Theorem 5.1]. □

Note that in the special case $V = 0$ we recover the heat equation $\partial_t \rho - \Delta \rho = 0$. For a more detailed introduction to Wasserstein gradient flows, we refer the reader to [San15, Chapter 8], where, in particular, further examples of partial differential equations and corresponding energy functionals are summarized.

3.2 Numerical Methods for the Classical Optimal Transport Problem

Numerous applications have led to plenty of computational methods to solve the optimal transport problem at least for some special cases. Here, we first give a brief overview of numerical algorithms and collect the basic ideas corresponding to the different formulations of the optimal transport distance. Later, we study optimal transport distances based on the Benamou–Brenier formulation (3.5), which has already been used in [BB00] for the numerical purpose by applying a suitable change of variables. Then, the optimal transport problem turns into a convex optimization problem, which is solved via an augmented Lagrangian and duality techniques from convex analysis. In [PPO14], it was shown that a proximal splitting algorithm leads in fact to the same optimization scheme, which requires to solve a linear system corresponding to an elliptic problem on the time-space domain and pointwise projections onto a convex set. Here, we introduce the basic concepts from convex analysis, which are necessary for a proximal splitting algorithm.

3.2.1 Overview of Numerical Methods for Optimal Transport

1D Case. In the one-dimensional case, on an interval $[a,b] \subset \mathbb{R}$, the optimal transport map between $\mu_A, \mu_B \in \mathscr{P}([a,b])$ can be computed explicitly. Given any $\mu \in \mathscr{P}([a,b])$, the cumulative distribution function $C_\mu(x) := \int_a^x \mathrm{d}\mu$ is monotone, and thus, has a so-called pseudo-inverse $C_\mu^{-1}(y) := \min\{x \in [a,b] \,:\, y \leqslant C_\mu(x)\}$. Then, an optimal transport map for Monge's problem is given by $T = C_{\mu_B}^{-1} \circ C_{\mu_A}$. We refer the reader to [San15, Chapter 2] for a detailed discussion.

Empirical Measures. Next, we consider the particular case that both measures $\mu_A, \mu_B \in \mathscr{P}(\mathbb{R}^d)$ are finite sums of weighted Dirac measures, *i.e.*, there are finitely many points $x_i \in \mathbb{R}^d$ for $i = 1, \dots, N$ and $y_j \in \mathbb{R}^d$ for $j = 1, \dots, M$ and corresponding weights $\alpha \in \mathbb{R}^N_{\geqslant 0}$, $\beta \in \mathbb{R}^M_{\geqslant 0}$ with $\sum_{i=1}^N \alpha_i = \sum_{j=1}^M \beta_j$ s.t.

$$\mu_A(x) = \sum_{i=1}^N \alpha_i \delta_{x_i}\,, \quad \mu_B(x) = \sum_{j=1}^M \beta_j \delta_{y_j}\,.$$

For a cost function c, we can define an associated cost matrix $C \in \mathbb{R}^{N\times M}$ with entries $C_{ij} = c(x_i, y_j)$. Then solving the Kantorovich problem (3.3) turns into minimizing the Euclidean scalar product $\langle P, C\rangle$ over all couplings $P \in \Pi(\mu_A, \mu_B) = \{P \in \mathbb{R}_+^{N\times M} \,:\, P\mathbf{1}_M = \alpha\,,\ P^T\mathbf{1}_N = \beta\}$, where we denote by $\mathbf{1}_N$ the vector in $\mathbb{R}^N$ with all entries equal 1. Thus, the optimal transport problem becomes a linear program in NM variables with $N + M$ constraints. Note that in the case $N = M$ and $\alpha_i = \beta_j$ for all i, j, this even simplifies to a simple sorting problem. In the general case, the linear program in the dual formulation

$$\max\{\langle f, \alpha\rangle + \langle g, \beta\rangle \,:\, (f,g) \in \mathbb{R}^N \times \mathbb{R}^M \text{ with } f_i + g_j \leqslant C_{ij}\}\,. \tag{3.8}$$

can, *e.g.*, be solved by the Auction algorithm [BE88].

Cyclical Monotonicity. For empirical measures, Schmitzer [Sch16a, Sch16b] proposed a sparse multiscale algorithm by making in addition to the linear program formulation (3.8) use of the cyclical monotonicity property, which states that for an optimal transport plan γ, the support $\mathrm{supp}(\gamma)$ is c-cyclically monotone, *i.e.*, for all $k \in \mathbb{N}$, all permutations σ, and all pairs $(x_i, y_i)_{i=1,\dots,k}$ we have $\sum_{i=1}^k c(x_i, y_i) \leqslant \sum_{i=1}^k c(x_i, y_{\sigma(i)})$.

Entropy Regularization. In [BCC$^+$15], the entropy functional $\mathcal{H}(P) = -\sum_{i=1}^N \sum_{j=1}^M P_{ij}(\log(P_{ij}) - 1)$ was added as a regularizer to the Kantorovich formulation for discrete measures, *i.e.*, for a regularization parameter $\varepsilon > 0$, the optimization problem

$$\min\{\langle P, C\rangle - \varepsilon\mathcal{H}(P) \,:\, P \in \Pi(\mu_A, \mu_B)\} \tag{3.9}$$

was investigated. By considering the associated Gibbs kernel with entries $G_{ij} = e^{-\frac{C_{ij}}{\varepsilon}}$ and defining the Kullback–Leibler divergence as

$$KL(P|G) = \sum_{i=1}^N \sum_{j=1}^M P_{ij}\left(\log\left(\frac{P_{ij}}{G_{ij}}\right) - 1\right)\,,$$

the problem (3.9) can be written as

$$\min\{\varepsilon KL(P|G) \,:\, P \in \Pi(\mu_A, \mu_B)\}\,.$$

Then optimizing the corresponding dual problem was solved by Sinkhorn's algorithm, which only performs matrix-vector-multiplications. Here, the sparsity of the matrix and thus, the speed of convergence depends on the regularization parameter ε. For $\varepsilon \to 0$, it has been shown in [PC17, Proposition 4.1] that solutions to the regularized problem converge to the optimal transport plan with maximal entropy. An entropy regularization was also applied in [Pey15] for the numerical computation of Wasserstein gradient flows and in [PCS16] for the Gromov–Wasserstein distance between two metric spaces, which was introduced by Sturm [Stu06a] using a Kantorovich formulation.

Semi-Discrete Optimal Transport. In the so-called semi-discrete case, we consider the optimal transport problem between a density $\mu_A = \rho_A \mathscr{L}$ and an empirical measure $\mu_B = \sum_{i=1}^N \beta_i \delta_{x_i}$. Based on Monge's formulation, Merigot [Mér11] used a geometric approach by optimizing weighted Voronoi cells. This approach was also applied in [Lév15] for tetrahedral meshes in $3D$.

Polar Factorization. In [HZTA04], the optimal transport map T was computed by making use of the polar factorization result by Brenier [Bre91]. For simplicity, the domain was restricted to be the unit square, where an explicit construction of an initialization T_0 s.t. $\mu_A = \det(DT_0)\mu_B \circ T_0$ is computable. Then, provided that $(T_0)_{\#}\mu_A$ is absolutely continuous, the polar factorization admits a unique decomposition $T_0 = (\nabla \Psi_0) \circ s_0$, where Ψ_0 is a convex function and s_0 is a measure-preserving map. Finally, a gradient descent method was applied to remove the measure-preserving part and consequently obtained an optimal transport map.

Monge–Ampère Equation. Solving the optimal transport problem numerically by solving the Monge–Ampère equation was studied in [LR05, BFO10, BFO10]. Note that in general even for absolutely continuous densities the optimal transport map T does not have to be a homeomorphism. However, under the assumption that T is an orientation-preserving diffeomorphism, the matching condition in (3.1) for measures $\mu_A = \rho_A \mathscr{L}$ and $\mu_B = \rho_B \mathscr{L}$ becomes $\rho_A = \det(DT)\rho_B \circ T$. Using the property of the optimal transport map being a gradient of a convex function $T = \nabla\psi$, we arrive at the Monge–Ampère equation $\rho_A = \det(D^2\psi)\rho_B \circ (\nabla\psi)$.

3.2.2 Convex Optimization

Now, we introduce the basic concepts of convex analysis, where we focus on proximal splitting algorithms. We refer the reader to [BC17] and [ET99] for a more general introduction.

In the following, let H be a Hilbert space. First, we recall basic definitions.

Definition 3.2.1. We say that $f\colon H \to \mathbb{R} \cup \{\infty\}$ is

1. proper if $\operatorname{dom}(f) := \{x \in H \ : \ f(x) < \infty\} \neq \emptyset$,
2. convex if $f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y)$ for all $x, y \in H$, $t \in [0,1]$, and
3. lower semi-continuous if $f(x) \leqslant \liminf_{k\to\infty} f(x_k)$ for all $x_k \to x$.

Furthermore, we denote by $\Gamma_0(H)$ the set of all proper, convex and lower semi-continuous functions on H.

Our main goal is to provide appropriate tools to find a solution to the minimization problem

$$\text{minimize } \mathcal{J}(x) = \mathcal{F}(x) + \mathcal{G}(x) \text{ over all } x \in H,$$

where the algorithm essentially makes use of the splitting of a functional $\mathcal{J}$ into $\mathcal{F} \in \Gamma_0(H)$ and $\mathcal{G} \in \Gamma_0(H)$.

Remark 3.2.2. More generally, we can develop the following concepts for a functional $\mathcal{J}(x) = \mathcal{F}(Kx) + \mathcal{G}(x)$ with a linear operator $K\colon H \to H$. Most of the here presented tools can also be extended to Banach spaces. Since this is not necessary for our applications, we restrict to Hilbert spaces and the case $K = \mathrm{id}$.

We point out that $\mathcal{J}$ does not have to be differentiable, s.t. numerical methods involving a gradient like a gradient descent cannot be applied. Instead, we introduce more general techniques, where it turns out that functions in $\Gamma_0(H)$ are so-called subdifferentiable.

Definition 3.2.3 (Subdifferential)**.** Let $f\colon H \to \mathbb{R} \cup \{\infty\}$ be proper and convex. Then the subdifferential of f in $x \in H$ is defined by

$$\partial f(x) = \{z \in H \ : \ \langle y - x, z\rangle \leqslant f(y) - f(x) \ \forall y \in H\} \ .$$

We call f subdifferentiable at x if $\partial f(x) \neq \emptyset$.

It can be verified that a function $f \in \Gamma_0(H)$ is subdifferentiable (see [BC17, Theorem 9.20]). Then Fermat's rule (see [BC17, Theorem 16.3]) generalizes the necessary condition $Df(x) = 0$ for a minimizer x of a smooth function. Indeed, x is a minimizer of f if and only if $0 \in \partial f(x)$. Since the subdifferential $\partial \mathcal{J}(x)$ might, in general, be challenging to compute, we take into account the so-called proximal mapping.

Definition 3.2.4 (Proximal Mapping)**.** For $f \in \Gamma_0(H)$, the proximal mapping is defined as

$$\operatorname{prox}_f(x) = \operatorname*{arg\,min}_{y \in H} \frac{1}{2}\|x - y\|_H^2 + f(y)\,.$$

Then we have the following relation between the proximal mapping and the subdifferential.

Proposition 3.2.5 (Relation between Proximal Mapping and Subdifferential)**.** *Let $f \in \Gamma_0(H)$ and let $x, p \in H$. Then*

$$p = \operatorname{prox}_f(x) \Leftrightarrow x - p \in \partial f(p)\,.$$

Proof. See [BC17, Proposition 16.44]. □

Now, similar to a gradient descent method, proximal point algorithms iteratively perform proximal operators to obtain a sequence, which converges to a minimizer of $\mathcal{J}$. In many applications, a closed-form expression of the proximal operator of $\mathcal{J}$ is not available, but $\mathcal{J}$ admits a splitting into functions $\mathcal{F}$ and $\mathcal{G}$ as above, s.t. the proximal operators of $\mathcal{F}$ and $\mathcal{G}$ can be computed explicitly. Then, in the optimization scheme, these proximal operators of $\mathcal{F}$ and $\mathcal{G}$ are applied alternatingly, where specific step sizes are given according to an appropriate fixed point map. Here, we present two widespread proximal splitting algorithms, which we use for our applications in Chapter 4 and Chapter 5.

Theorem 3.2.6 (Douglas–Rachford Splitting Algorithm)**.** *Let $a_0 \in H$ be an initial value, $\lambda \in (0, 2)$, and $\gamma > 0$. The iteration of the Douglas–Rachford splitting algorithm is defined for $n \in \mathbb{N}$ as*

$$\begin{aligned} b_{n+1} &= \operatorname{prox}_{\gamma \mathcal{G}}(a_n)\,, \\ a_{n+1} &= a_n + \lambda \left(\operatorname{prox}_{\gamma \mathcal{F}}(2b_{n+1} - a_n) - b_{n+1}\right)\,. \end{aligned} \tag{3.10}$$

Then both sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}_+}$ converge to a minimizer of $\mathcal{J}$.

Proof. See [EB92]. □

Furthermore, the algorithm developed by Chambolle and Pock [CP11] makes use of the convex dual formulation of the actual minimization problem. Therefore, we define the Fenchel conjugate.

Definition 3.2.7 (Fenchel Conjugate)**.** For a function $f \in \Gamma_0(H)$ we define its Fenchel conjugate f^* by

$$f^*(y) = \sup_{x \in H} \langle y, x \rangle_H - f(x)\,.$$

Theorem 3.2.8 (Chambolle–Pock Algorithm)**.** *Let $(a_0, b_0) \in H \times H$ be two initial values and set $c_0 = a_0$. Furthermore, let $\lambda \in [0, 1]$ and $\tau, \sigma > 0$ s.t. $\tau\sigma < 1$. The iteration of the Chambolle–Pock algorithm is defined for $n \in \mathbb{N}$ as*

$$\begin{aligned} b_{n+1} &= \operatorname{prox}_{\sigma \mathcal{F}^*}(b_n + \sigma c_n)\,, \\ a_{n+1} &= \operatorname{prox}_{\tau \mathcal{G}}(a_n - \tau b_{n+1})\,, \\ c_{n+1} &= a_n + \lambda (a_{n+1} - a_n)\,. \end{aligned} \tag{3.11}$$

Then the sequences $(a_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ converge to a minimizer of $\mathcal{J}$.

Proof. See [CP11]. □

A priori, computing $\operatorname{prox}_{f^*}$ might be easier to compute prox_f or vice-versa, but the following theorem allows computing one of these expressions if the other one is known.

Theorem 3.2.9 (Moreau Decomposition)**.** *For $f \in \Gamma_0(H)$ and $\alpha > 0$ we have the following identity*

$$\operatorname{prox}_{\alpha f}(x) + \operatorname{prox}_{\frac{1}{\alpha} f^*}\left(\frac{x}{\alpha}\right) = x\,.$$

Proof. See [BC17, Theorem 14.3]. □

Now, we discuss the example of an indicator function of a convex set, which we frequently apply in the sequel.

Example 3.2.10 (Proximal Map of Indicator Function)**.** Let $K \subset H$ be a closed and convex set. Recall that the indicator function is given by

$$\mathcal{I}_K(x) = \begin{cases} 0 & \text{if } x \in K\,, \\ \infty & \text{if } x \notin K\,. \end{cases}$$

Then

$$\operatorname{prox}_{\mathcal{I}_K}(x) = \operatorname*{arg\,min}_{y \in K} \frac{1}{2}\|x - y\|_H^2 = \operatorname{proj}_K(x)\,,$$

where proj_K denotes the orthogonal projection on K w.r.t. the norm $\|\cdot\|_H$ on H.

Moreover, we give a characterization of the projection onto a convex set by taking into account the so-called normal cone.

Lemma 3.2.11 (Characterization of Projection by Normal Cone)**.** *Let $K \subset H$ be a nonempty, closed, and convex set. For $p \in H$ we define the normal cone by*

$$N_K(p) := \begin{cases} \{x \in H \ : \ \langle y - p, x\rangle \leqslant 0 \ \forall y \in K\} & \textit{if } p \in K\,, \\ \varnothing & \textit{otherwise}\,. \end{cases}$$

Then the projection of p onto K is characterized by

$$p^{pr} = \operatorname{proj}_K(p) \quad \Leftrightarrow \quad p - p^{pr} \in N_K(p^{pr})\,.$$

Proof. See [BC17, Proposition 6.47]. □

3.2.3 Application of Proximal Splitting Methods to the Flow Formulation

Now, we demonstrate how a proximal splitting algorithm can be applied to solve the optimal transport problem numerically. Here, we take into account the Benamou–Brenier formulation (3.5), which we first have to transform into a convex optimization problem. Therefore, we make use of a change of variables by considering, instead of the pair mass and velocity (ρ, v), the pair mass and momentum $(\rho, m = \rho v)$. This change of variables was already performed in (3.5) for the numerical purpose. Then the optimization problem (3.5) becomes

$$\mathcal{W}(\rho_A\mathscr{L}, \rho_B\mathscr{L})^2 = \inf\left\{\int_0^1 \int_D \Phi(\rho, m)\,\mathrm{d}x\,\mathrm{d}t \ : \ (\rho, m) \in \mathcal{CE}(\rho_A, \rho_B)\right\}\,. \tag{3.12}$$

Here, the integrand of the kinetic energy $|v|^2\rho$ transforms pointwise into

$$\Phi(\rho, m) = \begin{cases} \dfrac{|m|^2}{\rho} & \text{if } \rho > 0\,, \\ 0 & \text{if } (\rho, m) = 0\,, \\ \infty & \text{otherwise,} \end{cases} \tag{3.13}$$

with the advantage that Φ is lower semi-continuous, convex and 1-homogeneous. Furthermore, the continuity equation $\partial_t\rho+\mathrm{div}(\rho v) = 0$ simplifies to a linear equation $\partial_t\rho+\mathrm{div}(m) = 0$. Thus, the optimization problem (3.12) is convex. Moreover it can be written as minimizing a functional $\mathcal{J} = \mathcal{F} + \mathcal{G}$ with a splitting into the functionals

$$
\begin{aligned}
\mathcal{F}(\rho, m) &= \mathcal{I}_{\mathcal{CE}(\rho_A,\rho_B)}(\rho, m)\,,\\
\mathcal{G}(\rho, m) &= \int_0^1 \int_D \Phi(\rho, m)\,\mathrm{d}x\,\mathrm{d}t = \mathcal{E}_{\mathrm{trans}}(\rho, m)\,.
\end{aligned}
$$

Thus, provided that $\mathrm{prox}_{\mathcal{F}}$ and $\mathrm{prox}_{\mathcal{G}}$ can be computed explicitly, we can apply a proximal splitting algorithm.

For a fully numerical scheme, in [PPO14], a staggered grid discretization was proposed, whereas in Section 4.5, we use a finite element discretization. Here, we do not describe a specific discretization of the functions ρ and m, but rather mention that for the concrete implementation, it is essential that $\mathrm{prox}_{\mathcal{G}}$ can be performed pointwise. We comment on that in the corresponding applications in Chapter 4 and Chapter 5. Furthermore, it turns out hat $\mathrm{prox}_{\mathcal{F}}$ requires to solve an elliptic problem on the time-space domain.

Proximal Map of Kinetic Energy

First, we identify pointwise the Fenchel conjugate of the function Φ.

Proposition 3.2.12 (Fenchel Conjugate of Kinetic Energy)**.** *For the function Φ defined in* (3.13) *we have that $\Phi^* = \mathcal{I}_{\mathcal{B}}$ is an indicator function of the convex set*

$$
\mathcal{B} = \left\{(\rho, m) \in \mathbb{R}\times\mathbb{R}^d \;:\; \rho + \frac{|m|^2}{4} \leqslant 0\right\}. \tag{3.14}
$$

Proof. See [BB00]. □

Then, using Moreau's identity, prox_{Φ} can be computed by projecting onto the convex set $\mathcal{B}$, which we now describe explicitly.

Lemma 3.2.13 (Projection onto $\mathcal{B}$)**.** *The projection of a point $(\rho, m) \in \mathbb{R}\times\mathbb{R}^d$ onto the set $\mathcal{B}$ is given by*

$$
\mathrm{proj}_{\mathcal{B}}(\rho, m) = (\rho^{pr}, m^{pr}) = \begin{cases} (\rho, m) & \textit{if } (\rho, m) \in \mathcal{B}\,,\\ \left(\rho + 1 - \dfrac{1}{\sigma}\,,\,\sigma m\right) & \textit{if } (\rho, m) \notin \mathcal{B}\,,\end{cases}
$$

where $\sigma \in \mathbb{R}$ is defined as the solution of the equation $\sigma^3|m|^2 + 2(1+\rho)\sigma - 2 = 0$.

Proof. In the case that $(\rho, m) \notin \mathcal{B}$, the projection lies on the boundary $\partial\mathcal{B}$, which can be parametrized by a map $\gamma\colon \mathbb{R}^d \to \partial\mathcal{B}$ defined as $\gamma(b) = \left(-\frac{|b|^2}{2}, b\right)$. Hence, the vector $(1, b) \in \mathbb{R}^{d+1}$ spans the normal space at a point $(a, b) \in \partial\mathcal{B}$. Now, for $(\rho, m) \in \mathbb{R}\times\mathbb{R}^d$, we search for the orthogonal projected point $(\rho^{\mathrm{pr}}, m^{\mathrm{pr}}) \in \partial\mathcal{B}$, which satisfies the relation $(\rho^{\mathrm{pr}}, m^{\mathrm{pr}}) + \tau(1, m^{\mathrm{pr}}) = (\rho, m)$ for some $\tau \in \mathbb{R}$. We set $\sigma = (1+\tau)^{-1}$, which leads to $(\rho^{\mathrm{pr}}, m^{\mathrm{pr}}) = (\rho + 1 - \frac{1}{\sigma}, \sigma m)$. Since $(\rho^{\mathrm{pr}}, m^{\mathrm{pr}}) \in \partial\mathcal{B}$, we obtain $\sigma^3|m|^2 + 2(1+\rho)\sigma - 2 = 0$. □

Note that this polynomial equation of order three can be solved by a simple Newton method.

Projection onto Solutions to the Continuity Equation

Next, we show that the projection on the set of solutions to the continuity equation can be computed by solving a Laplace equation on the time-space domain. Here, we do not specify function spaces, s.t. the following statement has to be understood rather formally.

Lemma 3.2.14 (Projection onto $\mathcal{CE}(\rho_A, \rho_B)$)**.** *For $p = (\rho, m) \colon [0,1] \times D \to \mathbb{R} \times \mathbb{R}^d$ the (formal) projection onto the set $\mathcal{CE}(\rho_A, \rho_B)$ can be computed by*

$$p^{pr} = p + \frac{1}{2}\nabla_{(t,x)}\phi^{pr} = (\rho, m) + \frac{1}{2}(\partial_t \phi^{pr}, \partial_x \phi^{pr}),$$

where $\phi^{pr} \colon [0,1] \times D \to \mathbb{R}$ solves the weak Laplace equation

$$\int_0^1 \int_D \frac{1}{2}\nabla_{(t,x)}\phi^{pr}\nabla_{(t,x)}\widehat{\phi}\,\mathrm{d}x\,\mathrm{d}t = \int_D \widehat{\phi}(1)\rho_B - \widehat{\phi}(0)\rho_A\,\mathrm{d}x - \int_0^1 \int_D p\nabla_{(t,x)}\widehat{\phi}\,\mathrm{d}x\,\mathrm{d}t$$

for all test functions $\widehat{\phi} \colon [0,1] \times D \to \mathbb{R}$.

Proof. The proof (for rigorous function spaces) can be obtained as a special case of Proposition 4.6.1. □

Chapter 4

Optimal Transport with Source Term

We have seen in Chapter 3 that the Wasserstein distance defines a metric on the space of Borel probability measures. Now, we are interested in extending this classical optimal transport distance to the space of positive Radon measures and in particular, defining a metric between two measures of possibly different total mass. Such a generalization is for example motivated by imaging applications, where the Wasserstein distance was used for nonrigid image registration (see, *e.g.*, [HZTA04]), but unfortunately input images to be compared are in general not of the same mass given by the intensity of gray values. Thus, for the classical optimal transport distance, a contrast modulation on the input images is required before an optimal matching between the input images can be computed. Even if the total mass of the input images coincides, a global mass redistribution between totally independent image structures is unfavorable, and instead, we desire local intensity modulations to match similar structures.

In this chapter, we present a possible generalization of the Benamou–Brenier formulation [BB00] by introducing a source term in the continuity equation and penalizing the amount of source in addition to the kinetic energy. We studied such a model in [MRSS15], where we proposed a penalization of the source in a squared L^2-norm both in time and space. There, the existence of geodesic paths is established, where the framework of Radon measures is taken into account, and the corresponding measures for mass, momentum, and source term are decomposed into absolutely continuous and singular parts w.r.t. the Lebesgue measure. To ensure that the definitions of the energy functionals do not depend on the decomposition, by a lower semi-continuity result on integral functionals in [BB90, BB92], it turns out that 1-homogeneity of the integrands for the singular measures is a suitable assumption. But then a penalization of the source term in a squared L^2-norm both in time and space does not allow singular sources. Instead, we propose an L^1-norm of the source term in space and an L^2-norm in time to provide an equiintegrability estimate, which guarantees compactness in the space of curves of Radon measures. This is, in particular, desirable in the context of image warping, where, *e.g.*, line segments correspond to singular sources.

This chapter is organized as follows. We formally derive our generalized optimal transport model with source term in Section 4.1. During the last years, a lot of similar approaches have been proposed in the literature, which we summarize and compare with our model in Section 4.2. In Section 4.4, we rigorously define the generalized optimal transport on the space of Radon measures. Here, following [DNS09], we prove the existence of optimal transport geodesics. As a preliminary step, we summarize important results on curves of Radon measures in Section 4.3. In Section 4.5, we present a finite element discretization, and in Section 4.6, we show how the corresponding discrete optimization problem can be solved via proximal splitting methods based on the approach for classic optimal transport in [PPO14]. Finally, in Section 4.7, we present our numerically computed results for selected academic examples to discuss the properties of our generalized model, as well as for real texture images.

Remark 4.0.1 (Collaborations and Publications). All results presented in this chapter are joint work with Jan Maas and Martin Rumpf and have been published in [MRS17]. It is based on a joint work with Jan Maas, Martin Rumpf, and Carola Schönlieb, which has been published in [MRSS15].

4.1 A Benamou–Brenier Formula with Source Term

In this section, we formally derive a generalized optimal transport distance, which relaxes the mass-preserving condition. In the following, let $D \subset \mathbb{R}^d$ be a compact and convex domain. We recall the Benamou–Brenier formulation (3.5) that allows us to compute the L^2-Wasserstein distance $\mathcal{W}(\rho_A, \rho_B)$ between two probability density functions ρ_A and ρ_B by minimizing the path energy

$$\mathcal{E}_{\text{trans}}(\rho, v) = \int_0^1 \int_D \rho |v|^2 \,\mathrm{d}x \,\mathrm{d}t \tag{4.1}$$

over all curves of density functions $\rho\colon [0,1] \times D \to \mathbb{R}_{\geqslant 0}$ with temporal boundary constraints $\rho(0) = \rho_A$ and $\rho(1) = \rho_B$ and corresponding velocity fields $v\colon [0,1] \times D \to \mathbb{R}^d$ s.t. the continuity equation $\partial_t \rho + \operatorname{div}(\rho v) = 0$ is satisfied.

Note that the continuity equation is a mass-preserving condition, which enforces $\rho(t)$ to remain in the space of Borel probability measures for every $t \in [0,1]$. To relax this condition, we introduce a source term $z\colon [0,1] \times D \to \mathbb{R}$ in the continuity equation:

$$\partial_t \rho + \operatorname{div}(\rho v) = z\,. \tag{4.2}$$

Then, the source term is penalized in addition to the kinetic energy. Therefore, we introduce a source term cost functional

$$\mathcal{E}_{\text{source}}(z) = \int_0^1 \left(\int_D r(z) \,\mathrm{d}x \right)^2 \mathrm{d}t\,. \tag{4.3}$$

Here, we propose $r\colon \mathbb{R} \to \mathbb{R}$ to be a nonnegative, convex function satisfying $r(0) = 0$. Moreover, we assume a linear growth condition, *i.e.*, there exists a constant $C_r \in \mathbb{R}$ s.t. $r(s) \leqslant C_r(1 + |s|)$ for all $s \in \mathbb{R}$. Possible choices for r are given in the following example.

Example 4.1.1 (Functions for Source Term Energy)**.**

1. The absolute value $r(s) = |s|$ corresponds to the L^1-norm in space.
2. For some $\beta > 0$, the Huber function

$$r(s) = \begin{cases} \dfrac{1}{2\beta} s^2 & \text{if } s \leqslant \beta\,, \\[2ex] |s| - \dfrac{\beta}{2} & \text{otherwise,} \end{cases} \tag{4.4}$$

has linear growth for large s but is quadratic around zero. In our computations, we choose $\beta = 10^{-4}$.

Altogether, we define a generalized optimal transport path energy functional

$$\begin{aligned} \mathcal{E}_\delta(\rho, v, z) &= \mathcal{E}_{\text{trans}}(\rho, v) + \frac{1}{\delta} \mathcal{E}_{\text{source}}(z) \\ &= \int_0^1 \int_D \rho |v|^2 \,\mathrm{d}x \,\mathrm{d}t + \frac{1}{\delta} \int_0^1 \left(\int_D r(z) \,\mathrm{d}x \right)^2 \mathrm{d}t\,, \end{aligned} \tag{4.5}$$

which has to be minimized over all solutions to the relaxed continuity equation (4.2) and the temporal boundary constraints $\rho(0) = \rho_A$ and $\rho(1) = \rho_B$, where ρ_A and ρ_B are no longer restricted to have equal total mass.

Later, we show that in a mathematically rigorous setup formulated on the space of Radon measures the linear growth condition on r allows singular sources. Moreover, the penalty parameter $\delta > 0$ allows for regulating the mass modulation rate. Note that for $\delta = 0$ a pure blending between ρ_A and ρ_B with zero velocity has minimal energy, whereas for $\delta \to \infty$ transport becomes cheaper. In our computational results in Section 4.7, we verify these effects of the parameter δ.

4.2 Relation to Previous Work on Optimal Transport with Source Term

During the last years, there has been a lot of activity in extending optimal transport distances to spaces of densities or measures with possibly different masses, which we briefly summarize and point out differences to our model.

Partial Optimal Transport. A so-called partial optimal transport model was proposed in [CM10] by relaxing the marginal constraint in the Kantorovich formulation. More precisely, it was asked for transporting a fixed fraction of some initial to a final density function by minimizing the L^2-transport cost. This model was analyzed in [Fig10] by studying the geometry of the subsets which are transported. However, there is no source term involved directly.

Unbalanced Semi-Discrete Optimal Transport. We have discussed in Section 3.2.1 that Wasserstein geodesics between a density and a discrete measure can be computed by using methods from algorithmic geometry. Recently, in [BSW18], this approach was extended to the unbalanced semi-discrete optimal transport problem.

Furthermore, there are some optimal transport distances which are based on minimizing a path energy subject to a continuity equation with a source term and therefore, can be considered as generalized Benamou–Brenier formulations. For an L^p-norm in time and an L^q norm in space, we introduce the notation

$$\mathcal{E}_{\mathrm{source},L^p(L^q)}(z) := \|\|z(t,\cdot)\|_{L^q(D)}\|_{L^p([0,1])} = \left(\int_0^1 \left(\int_D |z(t,x)|^q \,\mathrm{d}x\right)^{\frac{p}{q}} \mathrm{d}t\right)^{\frac{2}{p}} \tag{4.6}$$

and refer a formulation with such a source term as an $L^p(L^q)$-model. In the same manner, for the Huber function (4.4), the source term cost functional (4.3) is denoted by an $L^2(H)$-model.

$L^1(L^1)$**-Model.** In [PR16, PR14], a source term was introduced and minimizers of the path energy

$$\mathcal{E}_{\mathrm{trans}}(\rho,v) + \mathcal{E}_{\mathrm{source},L^1(L^1)}(z) = \int_0^1 \int_D \rho|v|^2 \,\mathrm{d}x\,\mathrm{d}t + \left(\int_0^1 \int_D |z| \,\mathrm{d}x\,\mathrm{d}t\right)^2$$

subject to equation (4.2) were considered. Then it was proven for absolutely continuous measures ρ and absolutely continuous sources z that this geodesic formulation corresponds to solving the problem

$$\inf\{|\tilde{\rho}_A - \rho_A|_{TV} + |\tilde{\rho}_B - \rho_B|_{TV} + \mathcal{W}(\tilde{\rho}_A,\tilde{\rho}_B) \;:\; \tilde{\rho}_A,\tilde{\rho}_B \in \mathcal{M}(D)\,,\, |\tilde{\rho}_A|_{TV} = |\tilde{\rho}_B|_{TV}\}\,,$$

where the classical Wasserstein distance $\mathcal{W}(\tilde{\rho}_A,\tilde{\rho}_B)$ is well-defined since $\tilde{\rho}_A$ and $\tilde{\rho}_B$ have the same mass.

$L^2(L^2)$**-Model.** Instead of the squared L^1-norm for the source term functional in space, we chose in [MRSS15] a penalization in the squared L^2-norm, *i.e.*, the source term was given by $\mathcal{E}_{\mathrm{source},L^2(L^2)}$. Here, for the moment, we neglect the penalty parameter δ.

Wasserstein–Fisher–Rao Distance. In the independent works [CPSV15] and [LMS15], an interpolating distance between the Wasserstein distance and the Fisher–Rao distance was proposed by minimizing the energy

$$\mathcal{E}_{WFR}(\rho,v,z) = \int_0^1 \int_D \rho(|v|^2 + \alpha(z)) \,\mathrm{d}x\,\mathrm{d}t$$

subject to a continuity equation $\partial_t \rho + \mathrm{div}(\rho v) = \rho z$. Note that the source term in this model is integrated w.r.t. the measure given by ρ. Furthermore, in [CPSV15], a static Kantorovich formulation was formulated and it was shown that the distance in [PR16, PR14] arises as a special case.

In the following, we observe that the differences of these extended Benamou–Brenier formulations become crucial by properly extending the energies to the space of Radon measures.

4.3 Curves of Radon Measures

For a rigorous formulation of the energy functional (4.5), we investigate Radon measures on the time-space domain $[0,1] \times D$. We desire that these measures should still represent curves, *i.e.*, at a given time step $t \in [0,1]$ the time-space measure is again a measure in space. This assumption leads us to the concept of the disintegration of measures. Furthermore, we make use of a proper extension result of energy functionals to Radon measure established in [BB90, BB92]. For a short introduction in basic measure theory, we refer the reader to Section 2.1 and the references therein. In the following, let $(X, \mathcal{E})$ be a complete measure space. Later, in our applications, we consider $X = D$ being the space domain or $X = [0,1] \times D$ being the time-space domain, where in both cases X is endowed with the corresponding Lebesgue measure. First, we introduce an advantageous decomposition of measures.

Definition 4.3.1 (Absolute Continuity and Singularity)**.**

1. Let μ be a positive measure and ν be a vectorial measure on $(X, \mathcal{E})$. Then ν is absolutely continuous w.r.t. μ if for any $A \in \mathcal{E}$ with $\mu(A) = 0$ it follows that $|\nu|_{TV}(A) = 0$. In this case we write $\nu \ll \mu$.

2. Two positive measures μ, ν on $(X, \mathcal{E})$ are mutually singular if there exists $E \in \mathcal{E}$ with $\mu(E) = 0 = \nu(X \backslash E)$. In this case we write $\mu \perp \nu$. We say that vectorial measures μ, ν are mutually singular, if $|\mu|_{TV} \perp |\nu|_{TV}$.

Theorem 4.3.2 (Lebesgue Decomposition)**.** *Let μ be a positive and σ-finite measure on $(X, \mathcal{E})$, and let ν be a vectorial measure on $(X, \mathcal{E})$. Then, there are unique vectorial measures ν^a, ν^s s.t. $\nu^a \ll \mu$, $\nu^s \perp \mu$, and $\nu = \nu^a + \nu^s$. Furthermore, there is a unique function $f \in L^1(X, \mu)^m$ called the density of ν w.r.t. μ s.t. $\nu^a = f\mu$.*

Proof. See [AFP00, Theorem 1.28]. □

Next, we recall the disintegration theorem, where we restrict to a disintegration in time of a time-space domain.

Theorem 4.3.3 (Disintegration in Time)**.** *Let $\mu \in \mathscr{M}^+([0,1] \times D)$ be a positive Radon measure. We consider the projection $\mathrm{proj}_{[0,1]} \colon [0,1] \times D \to [0,1]$ on the time interval. If $\tilde{\mu} := \left(\mathrm{proj}_{[0,1]}\right)_\# \mu$ is a positive Radon measure,* i.e., *$\mu(K \times D) < \infty$ for all $K \subset [0,1]$ compact, then there exists a family $(\mu_t)_{t \in [0,1]} \subset \mathscr{M}^+(D)$ s.t.*

1. *$t \mapsto \mu_t$ is $\tilde{\mu}$-measurable,*

2. *$\mu_t(D) = 1$ $\tilde{\mu}$-a.e.,*

3. *for all $\eta \in L^1([0,1] \times D, \mu)$, we have that $\eta(t, \cdot) \in L^1([0,1], \mu_t)$ for $\tilde{\mu}$-a.e. $t \in [0,1]$,*

4. *for all $\eta \in L^1([0,1] \times D, \mu)$, we have that $t \mapsto \int_D \eta(t,x) \,\mathrm{d}\mu_t(x) \in L^1([0,1], \tilde{\mu})$, and*

5. *for all $\eta \in L^1([0,1] \times D, \mu)$, we have that*

$$\int_{[0,1] \times D} \eta(t,x) \,\mathrm{d}\mu(t,x) = \int_0^1 \int_D \eta(t,x) \,\mathrm{d}\mu_t(x) \,\mathrm{d}\tilde{\mu}(t) \,.$$

Proof. See [AFP00, Theorem 2.28]. □

We denote this disintegration by $\mu = \tilde{\mu} \otimes \mu_t$. In analogy, the disintegration result holds for vectorial Radon measures by taking into account the total variation $|\mu|_{TV} = \tilde{\mu} \otimes |\mu_t|_{TV}$. Note that we later omit the normalization $\mu_t(D) = 1$. In the application, we are interested in verifying the disintegration of a weakly-* convergent sequence of measures in the limit, which leads us to the definition of equiintegrability.

Definition 4.3.4 (Equiintegrability)**.** Let μ be a positive measure on $(X, \mathcal{E})$. A family $F \subset L^1(X, \mu)$ is equiintegrable if

1. for any $\varepsilon > 0$ there exists $A \in \mathcal{E}$ with $\mu(A) < \infty$ and $\int_{X \backslash A} |f| \,\mathrm{d}\mu < \varepsilon$ for all $f \in F$, and

2. for any $\varepsilon > 0$ there exists $\delta > 0$ s.t. for all $E \in \mathcal{E}$ with $\mu(E) < \delta$ we have $\int_E |f| \,\mathrm{d}\mu < \varepsilon$ for all $f \in F$.

The second condition is called uniform integrability. Note that in the cases of interest, the first condition is always satisfied, since X is assumed to be compact. Now, there are several possibilities to verify the equiintegrability condition. We use the following characterization.

Proposition 4.3.5 (Characterizations of Equiintegrability)**.** *A family $F \subset L^1(X, \mu)$ is equiintegrable if and only if for any superlinear function S there exists a constant C_S s.t. for all $f \in F$ we have*

$$\int_X S(f(x))\,\mathrm{d}\mu(x) \leqslant C_S < \infty\,.$$

Proof. See [San15, Chapter 8.3]. □

In particular, if for any $p > 1$ a family $F \subset L^1(X, \mu)$ is uniformly bounded in L^p then F is equiintegrable. However, this is no longer true for $p = 1$. This observation is essential for our optimal transport model with source term. Now, we state the connection between disintegration and equiintegrability.

Lemma 4.3.6 (Equiintegrability implies Existence of Disintegration)**.** *Let $(\mu^n)_{n\in\mathbb{N}} \subset \mathscr{M}^+([0,1] \times D)$ be a sequence of positive Radon measures. We assume that $\mu^n = \mathscr{L}_{[0,1]} \otimes \mu^n_t$ has disintegrations in time. More precisely, $(\mu^n_t) \in \mathscr{M}^+(D)$, $t \mapsto \mu^n_t$ is Borel measurable, and for all $\eta \in L^1([0,1] \times D)$ we have $\int_{[0,1]\times D} \eta(t,x)\,\mathrm{d}\mu^n(t,x) = \int_0^1 \int_D \eta(t,x)\,\mathrm{d}\mu^n_t(x)\,\mathrm{d}t$. Furthermore, we assume convergence $\mu^n \overset{*}{\rightharpoonup} \mu$. We define a sequence $(f^n)_{n\in\mathbb{N}} \subset L^1([0,1])$ by $f^n(t) = \mu^n_t(D)$. If $(f^n)_{n\in\mathbb{N}}$ is equiintegrable, then the limit measure $\mu \in \mathscr{M}^+([0,1] \times D)$ has a disintegration $\mu = \mathscr{L}_{[0,1]} \otimes \mu_t$ in time.*

Proof. The statement is, *e.g.*, applied in [DNS09, Lemma 4.5], where a similar result in probability theory to prove the existence of conditional expectation is referred. Here, we briefly collect the arguments in our specific case.
By assumption, the sequence $(f^n)_{n\in\mathbb{N}}$ is equiintegrable and is uniformly bounded in $L^1([0,1])$, since $\|f^n\|_{L^1} = \mu^n([0,1] \times D)$ and μ^n is convergent. By the Dunford–Pettis Theorem (see [AFP00, Corollary 1.33]) there is a subsequence (again indexed by n) s.t. $f^n \rightharpoonup f$ in $L^1([0,1])$. Then for every $\tilde{\eta} \in C([0,1])$ we have that $\int_{[0,1]\times D} \tilde{\eta}(t)\,\mathrm{d}\mu(t,x) = \int_{[0,1]} \tilde{\eta}(t) f(t)\,\mathrm{d}t$. Thus, by Theorem 4.3.3 we obtain a disintegration $\mu = f\mathscr{L}_{[0,1]} \otimes \hat{\mu}_t$ in time, which can be rewritten as $\mu = \mathscr{L}_{[0,1]} \otimes f(t)\hat{\mu}_t =: \mathscr{L}_{[0,1]} \otimes \mu_t$. □

Next, we consider functionals $\mathcal{J} : L^1(X, \mathbb{R}^d) \to \mathbb{R}$ of type

$$\mathcal{J}(u) = \int_X f(u)\,\mathrm{d}\mathscr{L}\,, \tag{4.7}$$

where $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is supposed to be a proper, convex, and lower semi-continuous function (see Definition 3.2.1). In [BB90, BB92], a proper extension of the functional $\mathcal{J}$ onto the space of Radon measures was defined. For this purpose, we need the definition of the recession function.

Definition 4.3.7 (Recession Function)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper, convex, and lower semi-continuous function. The recession function $f_\infty : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is defined by

$$f_\infty(x) = \lim_{t\to\infty} \frac{f(x_0 + tx) - f(x_0)}{t}\,,$$

where $x_0 \in \mathbb{R}^d$ satisfies $f(x_0) < \infty$.

Proposition 4.3.8 (Properties of the Recession Function)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper, convex, and lower semi-continuous function. Then the recession function satisfies the following properties.*

1. *The definition of f_∞ is independent of x_0.*
2. *f_∞ is convex and lower semi-continuous.*
3. *f_∞ is 1-homogeneous.*

Proof. See [AFP00, Chapter 2.6]. □

Then, we can state the following extension result.

Theorem 4.3.9 (Lower Semi-Continuity of the Extended Functional)**.** *Let $f\colon \mathbb{R}^d \to [0,\infty]$ be a proper, convex, and lower semi-continuous function. We consider an open or compact set $X \subset \mathbb{R}^d$. Let $\mu \in \mathscr{M}^+(X)$ be a positive Radon measure and $\nu \in \mathscr{M}(X,\mathbb{R}^d)$ be a vector-valued Radon measures. We define a functional*

$$\mathcal{G}(\nu,\mu) := \int_X f\left(\tfrac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\right)\,\mathrm{d}\mu(x) + \int_X f_\infty\left(\tfrac{\mathrm{d}\nu^s}{\mathrm{d}|\nu^s|_{TV}}(x)\right)\,\mathrm{d}|\nu^s|_{TV}(x)\,.$$

Then $\mathcal{G}$ is weak-$*$ lower semi-continuous, i.e., *for any sequence $(\mu_k)_{k\in\mathbb{N}}$ of positive Radon measures on X with $\mu_k \overset{*}{\rightharpoonup} \mu$ and any sequence $(\nu_k)_{k\in\mathbb{N}}$ of vector-valued Radon measures on X with $\nu_k \overset{*}{\rightharpoonup} \nu$ we have*

$$\mathcal{G}(\nu,\mu) \leqslant \liminf_{k\to\infty} \mathcal{G}(\nu_k,\mu_k)\,.$$

Proof. See [AFP00, Theorem 2.34]. □

Consequently, $\mathcal{J}$ can be extended to a functional $\mathcal{J}\colon \mathscr{M}(X) \to \mathbb{R}\cup\{\infty\}$ by

$$\mathcal{J}(\nu) = \int_X f(\tfrac{\mathrm{d}\nu}{\mathrm{d}\mathscr{L}})\,\mathrm{d}\mathscr{L} + \int_X f_\infty(\tfrac{\mathrm{d}\nu^s}{\mathrm{d}|\nu^s|_{TV}})\,\mathrm{d}|\nu^s|_{TV} = \int_X f(\tfrac{\mathrm{d}\nu}{\mathrm{d}\mathscr{L}})\,\mathrm{d}\mathscr{L} + f_\infty(1)|\nu^s|_{TV}(X)\,, \tag{4.8}$$

where $\nu = \frac{\mathrm{d}\nu}{\mathrm{d}\mathscr{L}} + \nu^s$ is the Lebesgue decomposition of ν (see Theorem 4.3.2). Then $\mathcal{J}$ is weak-* lower semi-continuous on $\mathscr{M}(X)$. Moreover, in the case of an absolutely continuous measure $\nu = u\mathscr{L}$, the functional $\mathcal{J}(\nu)$ in (4.8) coincides with the old definition of $\mathcal{J}(u)$ in (4.7).

4.4 Existence of Geodesics for a Generalized Optimal Transport Distance

Now, we propose a measure-valued setup to rigorously define a set of weak solutions for the continuity equation with source term (4.2) and the energy in (4.5) by taking into account the extension result in (4.8). Moreover, we prove the existence of corresponding generalized optimal transport geodesics. We follow the lines of [DNS09] for more general optimal transport distances and of [MRSS15] for a source term in the $L^2(L^2)$-norm.

4.4.1 Measure-Valued Formulation of the Path Energy Functional

As for the classical L^2-optimal transport problem, we first apply the change of variables $(\rho,v) \mapsto (\rho, m = \rho v)$, where m denotes the momentum. We recall from (3.13) that the integrand of the kinetic energy transforms to

$$\Phi(\rho,m) = \begin{cases} \dfrac{|m|^2}{\rho} & \text{if } \rho > 0\,, \\ 0 & \text{if } (\rho,m) = 0\,, \\ \infty & \text{otherwise,} \end{cases}$$

which is a lower semi-continuous, convex, and 1-homogeneous function.

Additionally to the assumption that D is a bounded, convex domain, we furthermore consider D to be closed, s.t. by Theorem 2.1.3 duality of Radon measures on $[0,1]\times D$ is given by continuous functions on $[0,1]\times D$. Then we introduce Radon measures

$\mu \in \mathscr{M}^+([0,1]\times D)$	for the mass,
$\nu \in \mathscr{M}([0,1]\times D,\mathbb{R}^d)$	for the momentum, and
$\zeta \in \mathscr{M}([0,1]\times D)$	for the source.

We start by formulating a generalized continuity equation with source term in terms of these measure-valued quantities, which are a priori just measures on the time-space domain $[0,1]\times D$, but we desire that these measures represent curves of measures on the space domain D. Thus, we incorporate certain disintegration assumptions on the measures.

Definition 4.4.1 (Weak Continuity Equation with Source Term)**.** Let $\mu_A, \mu_B \in \mathscr{M}^+(D)$ be given. A triple of measures (μ, ν, ζ) in the space $\mathscr{M}^+([0,1] \times D) \times \mathscr{M}([0,1] \times D, \mathbb{R}^d) \times \mathscr{M}([0,1] \times D)$ is said to be a weak solution to the continuity equation with source term

$$\partial_t \mu + \operatorname{div}(\nu) = \zeta, \qquad \mu_0 = \mu_A, \quad \mu_1 = \mu_B,$$

if the following conditions hold:

1. The measures μ, ν and ζ admit disintegrations w.r.t. the Lebesgue measure in time, *i.e.*, there exist measure-valued functions $t \mapsto \mu_t$ weak-*-continuous in $\mathscr{M}^+(D)$, $t \mapsto \nu_t$ Borel measurable in $\mathscr{M}(D, \mathbb{R}^d)$ with $\int_0^1 |\nu_t|(D)\, \mathrm{d}t < \infty$, and $t \mapsto \zeta_t$ Borel measurable in $\mathscr{M}(D)$ with $\int_0^1 |\zeta_t|(D)\, \mathrm{d}t < \infty$, s.t.

$$\int_{[0,1]\times D} \eta(t,x)\, \mathrm{d}\mu(t,x) = \int_0^1 \int_D \eta(t,x)\, \mathrm{d}\mu_t(x)\, \mathrm{d}t \quad \forall \eta \in L^1(\mu),$$
$$\int_{[0,1]\times D} \eta(t,x)\, \mathrm{d}\nu(t,x) = \int_0^1 \int_D \eta(t,x)\, \mathrm{d}\nu_t(x)\, \mathrm{d}t \quad \forall \eta \in L^1(\nu),$$
$$\int_{[0,1]\times D} \eta(t,x)\, \mathrm{d}\zeta(t,x) = \int_0^1 \int_D \eta(t,x)\, \mathrm{d}\zeta_t(x)\, \mathrm{d}t \quad \forall \eta \in L^1(\zeta).$$

2. The continuity equation with source term $\partial_t \mu + \operatorname{div}(\nu) = \zeta$ with boundary values $\mu_0 = \mu_A$ and $\mu_1 = \mu_B$ holds in the sense of distributions, *i.e.*, for all time-space test functions $\eta \in C^1([0,1] \times D)$ we have

$$\begin{aligned} 0 = &\int_0^1 \left(\int_D \partial_t \eta(t,x)\, \mathrm{d}\mu_t(x) + \int_D \nabla_x \eta(t,x)\, \mathrm{d}\nu_t(x) + \int_D \eta(t,x)\, \mathrm{d}\zeta_t(x) \right) \mathrm{d}t \\ &- \int_D \eta(1,x)\, \mathrm{d}\mu_B(x) + \int_D \eta(0,x)\, \mathrm{d}\mu_A(x). \end{aligned} \tag{4.9}$$

Finally, we denote by $\mathcal{CE}(\mu_A, \mu_B)$ the set of all solutions to the weak continuity equation with source term and temporal boundary data μ_A at time $t = 0$ and μ_B at time $t = 1$.

Note that (4.9) means that the continuity equation is implicitly taken with homogeneous Neumann boundary conditions in space. Moreover, we mention that the source terms ζ in Definition 4.4.1 are signed Radon measures, s.t. solutions μ to the weak continuity equation (4.9) a priori could become negative as well. However, since we aim at computing geodesic between nonnegative measures $\mu_A, \mu_B \in \mathscr{M}^+(D)$, we also define the measures μ to be nonnegative.

Next, we define the energy (4.5) in terms of measures. To this end, by using the Lebesgue decomposition (Theorem 4.3.2), we decompose for each $t \in [0,1]$, the triple $(\mu_t, \nu_t, \zeta_t) \in \mathscr{M}^+(D) \times \mathscr{M}(D, \mathbb{R}^d) \times \mathscr{M}(D)$ into

$$\mu_t = \rho_t \mathscr{L} + \mu_t^\perp, \qquad \nu_t = m_t \mathscr{L} + \nu_t^\perp, \qquad \zeta_t = z_t \mathscr{L} + \zeta_t^\perp,$$

s.t. the singular parts $\mu_t^\perp \in \mathscr{M}^+(D)$, $\nu_t^\perp \in \mathscr{M}(D, \mathbb{R}^d)$, and $\zeta_t^\perp \in \mathscr{M}(D)$ are singular with respect to the Lebesgue measure $\mathscr{L}$ on D. Then we define $\mathscr{L}_t^\perp := \mu_t^\perp + |\nu_t^\perp|_{TV} + |\zeta_t^\perp|_{TV} \in \mathscr{M}^+(D)$, s.t. $\mathscr{L}_t^\perp$ is orthogonal to $\mathscr{L}$. By construction, the singular parts admit a density with respect to $\mathscr{L}_t^\perp$:

$$\mu_t^\perp = \rho_t^\perp \mathscr{L}_t^\perp, \qquad \nu_t^\perp = m_t^\perp \mathscr{L}_t^\perp, \qquad \zeta_t^\perp = z_t^\perp \mathscr{L}_t^\perp.$$

Furthermore, we make use of the proper extension of a functional to the space of Radon measures as described in Section 4.3. Note that in our case, the function Φ is 1-homogeneous. By the convexity and linear growth condition of r, the recession function r_∞ is well-defined with $r_\infty(1) = C_r$.

Now, with these decompositions of the measures at hand, we can define the rigorous version of the energy functional (4.5) in the measure-valued setting. The path energy functional for transport is taken from the Benamou–Brenier formulation of the L^2-Wasserstein distance, *i.e.*, for a fixed time t, the kinetic energy in space is given by

$$\mathcal{D}_{\text{trans}}(\mu_t, \nu_t) := \int_D \Phi(\rho_t, m_t)\, \mathrm{d}\mathscr{L} + \int_D \Phi(\rho_t^\perp, m_t^\perp)\, \mathrm{d}\mathscr{L}_t^\perp.$$

To describe the path energy functional involving the source term, we recall that $r\colon \mathbb{R} \to \mathbb{R}$ is a nonnegative, convex function with linear growth satisfying $r(0) = 0$. Then, for a fixed time t, we define the source term energy functional in space by

$$\mathcal{D}_{\text{source}}(\zeta_t) := \left(\int_D r(z_t)\,\mathrm{d}\mathscr{L} + \int_D C_r|z_t^\perp|\,\mathrm{d}\mathscr{L}_t^\perp\right)^2 .$$

Note that we consider a 1-homogeneous integrand for the singular part of the source measure with the aim to allow for singular sources with support of the source measure on a set of zero Lebesgue measure. Actually, $\mathscr{L}_t^\perp$ depends on μ_t, ν_t and ζ_t, but by the 1-homogeneity we have in fact $|z_t^\perp|_{TV}\mathscr{L}_t^\perp = |\zeta_t^\perp|_{TV}$ and $\Phi(\rho_t^\perp, m_t^\perp)\mathscr{L}_t^\perp = \Phi(\rho_t^\perp, m_t^\perp)(\mu_t^\perp + |\nu_t^\perp|_{TV})$. Therefore, $\mathcal{D}_{\text{trans}}$ only depends on (μ_t, ν_t) and $\mathcal{D}_{\text{source}}$ only depends on ζ_t. The total energy functional in space $\mathcal{D}_\delta\colon \mathscr{M}^+(D) \times \mathscr{M}(D, \mathbb{R}^d) \times \mathscr{M}(D) \to [0, \infty]$ is defined as

$$\mathcal{D}_\delta(\mu_t, \nu_t, \zeta_t) := \mathcal{D}_{\text{trans}}(\mu_t, \nu_t) + \frac{1}{\delta}\mathcal{D}_{\text{source}}(\zeta_t) .$$

Finally, corresponding to (4.5), we can rigorously define the total energy functional in time and space for measure-valued quantities by $\mathcal{E}_\delta\colon \mathscr{M}^+([0,1] \times D) \times \mathscr{M}([0,1] \times D, \mathbb{R}^d) \times \mathscr{M}([0,1] \times D) \to \mathbb{R} \cup \{\infty\}$ as

$$\mathcal{E}_\delta(\mu, \nu, \zeta) = \begin{cases} \displaystyle\int_0^1 \mathcal{D}_\delta(\mu_t, \nu_t, \zeta_t)\,\mathrm{d}t & \text{if } (\mu, \nu, \zeta) \in \mathcal{CE}(\mu_A, \mu_B) , \\ \infty & \text{otherwise.} \end{cases}$$

4.4.2 Compactness and Existence Result

Next, we state a compactness result for solutions to the weak continuity equation with source term.

Proposition 4.4.2 (Compactness of Solutions to the Continuity Equation with Source Term with Bounded Energy)**.** *Suppose that a sequence $(\mu^n, \nu^n, \zeta^n)_{n\in\mathbb{N}}$ in $\mathcal{CE}(\mu_A, \mu_B)$ with temporal boundary values μ_A and μ_B has bounded energy,* i.e.*, there exists a constant $C < \infty$ s.t.*

$$\sup_{n\in\mathbb{N}} \mathcal{E}_\delta(\mu^n, \nu^n, \zeta^n) \leqslant C . \tag{4.10}$$

Then, there exists a subsequence (again indexed by n) and a triple $(\mu, \nu, \zeta) \in \mathcal{CE}(\mu_A, \mu_B)$ s.t.

1. *for all $t \in [0,1]$, $\mu_t^n \overset{*}{\rightharpoonup} \mu_t$ in $\mathscr{M}^+(D)$ for $n \to \infty$,*
2. *$\nu^n \overset{*}{\rightharpoonup} \nu$ in $\mathscr{M}([0,1] \times D, \mathbb{R}^d)$ for $n \to \infty$,*
3. *$\zeta^n \overset{*}{\rightharpoonup} \zeta$ in $\mathscr{M}([0,1] \times D)$ for $n \to \infty$, and*
4. *the following lower semi-continuity estimate holds:*

$$\int_0^1 \mathcal{D}_\delta(\mu_t, \nu_t, \zeta_t)\,\mathrm{d}t \leqslant \liminf_{n\to\infty} \int_0^1 \mathcal{D}_\delta(\mu_t^n, \nu_t^n, \zeta_t^n)\,\mathrm{d}t . \tag{4.11}$$

Proof. Note that the set $\mathcal{CE}(\mu_A, \mu_B)$ of solutions to the continuity equation with source term is closed under weak-* convergence. Consequently, the limit measure (μ, ν, ζ) is contained in $\mathcal{CE}(\mu_A, \mu_B)$ if the subsequence $(\mu^n, \nu^n, \zeta^n)_{n\in\mathbb{N}}$ of measures converges as stated above. The crucial part of the proof is to show that the limit measure can be disintegrated and the subsequence converges in the appropriate sense. In the following, C denotes a generic constant, which may change from line to line.

Step 1: Compactness of the Source Term. Since r is of linear growth, we have $|z| \leqslant C(1 + r(z))$, hence

$$|\zeta_t^n(D)| = \int_D |z_t^n|\,\mathrm{d}\mathscr{L} + \int_D |(z_t^n)^\perp|\,\mathrm{d}(\mathscr{L}_t^n)^\perp \leqslant C\left(1 + \sqrt{\mathcal{D}_{\text{source}}(\zeta_t^n)}\right) .$$

Because of the bounded energy assumption (4.10), the function

$$t \mapsto C\left(1 + \sqrt{\mathcal{D}_{\text{source}}(\zeta_t^n)}\right)$$

is bounded in $L^2([0,1])$, uniformly in n. Thus, we obtain a uniform bound for the source term:

$$\sup_{n\in\mathbb{N}} |\zeta^n([0,1]\times D)| = \sup_{n\in\mathbb{N}} \int_0^1 |\zeta_t^n(D)|\,\mathrm{d}t \leqslant \sup_{n\in\mathbb{N}} \int_0^1 C\Big(1 + \sqrt{\mathcal{D}_{\text{source}}(\zeta_t^n)}\Big)\,\mathrm{d}t \leqslant C\,.$$

From this estimate, we deduce that a subsequence of $(\zeta^n)_{n\in\mathbb{N}}$ converges weakly-* to a measure ζ. Crucial for the compactness result is that we can disintegrate ζ with respect to the Lebesgue measure on $[0,1]$ into a family of measures $(\zeta_t)_{t\in[0,1]} \in \mathcal{M}(D)$. Now, the sequence $\big(t \mapsto |\zeta_t^n|_{TV}(D)\big)_{n\in\mathbb{N}}$ is uniformly bounded in $L^2([0,1])$. By Proposition 4.3.5, this implies an equiintegrability estimate for $(t \mapsto \zeta_t^n(D))_{n\in\mathbb{N}}$, and as a consequence, we obtain the requested disintegration $(\zeta_t)_{t\in[0,1]} \in \mathcal{M}(D)$ of the limit measure ζ.

Step 2: Boundedness of the Mass. A standard approximation argument (see [DNS09, Lemma 4.1]) shows that solutions to the continuity equation with source term satisfy, for all $0 \leqslant t_0 \leqslant t_1 \leqslant 1$,

$$\begin{aligned}&\int_D \eta(t_1,x)d\mu_{t_1}(x) - \int_D \eta(t_0,x)d\mu_{t_0}(x)\\ &= \int_{t_0}^{t_1}\int_D \partial_t\eta(t,x)\,\mathrm{d}\mu_t(x)\,\mathrm{d}t + \int_{t_0}^{t_1}\int_D \nabla_x\eta(t,x)\,\mathrm{d}\nu_t(x)\,\mathrm{d}t + \int_{t_0}^{t_1}\int_D \eta(t,x)\,\mathrm{d}\zeta_t(x)\,\mathrm{d}t\end{aligned} \tag{4.12}$$

for all time-space test functions $\eta \in C^1([0,1]\times D)$. In particular, taking $\eta(t,x) \equiv 1$, it follows that

$$\mu_{t_1}(D) - \mu_{t_0}(D) = \int_{t_0}^{t_1} \zeta_t(D)\,\mathrm{d}t\,. \tag{4.13}$$

This formula (4.13) for the change of mass yields a uniform bound

$$\mu_t^n(D) \leqslant \mu_A(D) + \int_0^t |\zeta_s^n|_{TV}(D)\,\mathrm{d}s \leqslant C \tag{4.14}$$

for all $n \in \mathbb{N}$ and $t \in [0,1]$.

Step 3: Compactness of the Momentum. To prove the compactness of the momentum term, we first claim that the maps $\big(t \mapsto |\nu_t^n|_{TV}(D)\big)_{n\in\mathbb{N}}$ are uniformly bounded in $L^2([0,1])$, hence equiintegrable. To see this, we follow [DNS09, Proposition 3.6] to obtain

$$\begin{aligned}|\nu_t^n|_{TV}(D) &= \int_D |m_t^n|\,\mathrm{d}\mathcal{L} + \int_D |(m_t^n)^\perp|\,\mathrm{d}(\mathcal{L}_t^n)^\perp\\ &\leqslant \left(\int_D \Phi(\rho_t^n, m_t^n)\,\mathrm{d}\mathcal{L}\right)^{\frac12}\left(\int_D \rho_t^n\,\mathrm{d}\mathcal{L}\right)^{\frac12} + \left(\int_D \Phi\big((\rho_t^n)^\perp,(m_t^n)^\perp\big)\,\mathrm{d}(\mathcal{L}_t^n)^\perp\right)^{\frac12}\left(\int_D (\rho_t^n)^\perp\,\mathrm{d}(\mathcal{L}_t^n)^\perp\right)^{\frac12}\\ &\leqslant \left(\int_D \Phi(\rho_t^n, m_t^n)\,\mathrm{d}\mathcal{L} + \int_D \Phi\big((\rho_t^n)^\perp,(m_t^n)^\perp\big)\,\mathrm{d}(\mathcal{L}_t^n)^\perp\right)^{\frac12}\left(\int_D \rho_t^n\,\mathrm{d}\mathcal{L} + \int_D (\rho_t^n)^\perp\,\mathrm{d}(\mathcal{L}_t^n)^\perp\right)^{\frac12}\\ &= (\mathcal{D}_{\text{trans}}(\mu_t^n,\nu_t^n))^{\frac12}\,(\mu_t^n(D))^{\frac12}\,,\end{aligned}$$

where we used the scalar inequality $\sqrt{ab} + \sqrt{cd} \leqslant \sqrt{a+c}\sqrt{b+d}$ which holds for $a,b,c,d \geqslant 0$. Then, taking into account (4.10) and (4.14), the uniform bound on $\big(t \mapsto |\nu_t^n|_{TV}(D)\big)_{n\in\mathbb{N}}$ follows. Using the inequality

$$|\nu^n|_{TV}([0,1]\times D) \leqslant \left(\int_0^1 |\nu_t^n|_{TV}(D)^2\,\mathrm{d}t\right)^{\frac12}\,,$$

we infer that the sequence of vectorial Radon measures $(\nu^n)_{n\in\mathbb{N}} \subset \mathcal{M}([0,1]\times D,\mathbb{R}^d)$ has uniformly bounded total variation on $[0,1]\times D$. Therefore, we can extract a subsequence that converges weakly-* to some measure

$\nu \in \mathscr{M}([0,1] \times D, \mathbb{R}^d)$. Since the sequence $\big(t \mapsto |\nu_t^n|_{TV}(D)\big)_{n\in\mathbb{N}}$ is equiintegrable, we obtain a disintegration $(\nu_t)_{t\in[0,1]} \in \mathscr{M}(D\,,\,\mathbb{R}^d)$ of $\nu = \mathscr{L}_{[0,1]} \otimes \nu_t$ in time.
Note that, without any modifications, the here presented compactness estimate of the momentum variable also holds for a source term in the $L^2(L^2)$-norm (*cf*. [MRSS15]).

Step 4: Compactness of the Mass. We show that $\mu_t^n \overset{*}{\rightharpoonup} \mu_t$ in $\mathscr{M}^+(D)$ for $n \to \infty$. Therefore, we fix $\tau \in [0,1]$, take $\eta \in C^1(D)$, and set $\xi(t,x) := \nabla\eta(x)\chi_{[0,\tau]}(t)$. Even though ξ is discontinuous, it follows from general approximation results (see [AGS08, Proposition 5.1.10]) that

$$\int_0^\tau \int_D \nabla\eta \,\mathrm{d}\nu_t^n \,\mathrm{d}t = \int_{[0,1]\times D} \xi \,\mathrm{d}\nu^n \to \int_{[0,1]\times D} \xi \,\mathrm{d}\nu = \int_0^\tau \int_D \nabla\eta \,\mathrm{d}\nu_t \,\mathrm{d}t\,. \tag{4.15}$$

Setting $\iota(t,x) := \eta(x)\chi_{[0,\tau]}(t)$ and arguing as above, we obtain

$$\int_0^\tau \int_D \eta \,\mathrm{d}\zeta_t^n \,\mathrm{d}t = \int_{[0,1]\times D} \iota \,\mathrm{d}\zeta^n \to \int_{[0,1]\times D} \iota \,\mathrm{d}\zeta = \int_0^\tau \int_D \eta \,\mathrm{d}\zeta_t \,\mathrm{d}t\,. \tag{4.16}$$

Now, we can obtain the convergence of a subsequence of $(\mu_t^n)_{n\in\mathbb{N}}$. By the weak continuity equation (4.12), for a triple (μ^n, ν^n, ζ^n), we have for all $\eta \in C^1(D)$ and all $t \in [0,1]$ that

$$\int_D \eta(x) \,\mathrm{d}\mu_t^n(x) = \int_D \eta(x) \,\mathrm{d}\mu_A(x) + \int_0^t \int_D \nabla\eta(x) \,\mathrm{d}\nu_t^n(x) \,\mathrm{d}t + \int_0^t \int_D \eta(x) \,\mathrm{d}\zeta_t^n(x) \,\mathrm{d}t\,.$$

Using (4.15) and (4.16), we can pass to the limit

$$\int_D \eta(x) \,\mathrm{d}\mu_t^n(x) \to \int_D \eta(x) \,\mathrm{d}\mu_A(x) + \int_0^t \int_D \nabla\eta(x) \,\mathrm{d}\nu_t(x) \,\mathrm{d}t + \int_0^t \int_D \eta(x) \,\mathrm{d}\zeta_t(x) \,\mathrm{d}t =: L_t(\eta)\,.$$

The right rand side defines a linear functional $L_t\colon C^1(D) \to \mathbb{R}$. Furthermore, we get from (4.14) the uniform bound

$$\left|\int_D \eta(x) \,\mathrm{d}\mu_t^n(x)\right| \leqslant \|\eta\|_\infty \sup_{t\in[0,1]} |\mu_t(D)| \leqslant C\|\eta\|_\infty \tag{4.17}$$

for all $\eta \in C^1(D)$ and all $n \in \mathbb{N}$. By density of $C^1(D)$ in $C(D)$, we can extend L_t to a linear and bounded functional on $C(D)$. Hence, by duality of Radon measures (Theorem 2.1.3), this defines for every $t \in [0,1]$ a measure μ_t s.t. $(\mu_t^n)_{n\in\mathbb{N}}$ converges weak-* for a subsequence to μ_t. Then, we can define $\mu \in \mathscr{M}^+([0,1] \times D)$ by

$$\int_{[0,1]\times D} \eta(t,x) \,\mathrm{d}\mu(t,x) = \int_0^1 \int_D \eta(t,x) \,\mathrm{d}\mu_t(x) \,\mathrm{d}t \quad \forall\eta \in C([0,1] \times D)\,,$$

and since the constant in (4.17) does not depend on t, we have that μ^n converges weakly-* to μ in $\mathscr{M}^+([0,1] \times D)$.

Step 5: Weak-* Continuity of the Disintegration of Mass. Finally, we have to check that the disintegration $t \mapsto \mu_t$ is weak-* continuous, *i.e.*, for all $\eta \in C([0,1] \times D)$ and for all $t_k \to t$ we have

$$\int_D \eta(t_k,x) \,\mathrm{d}\mu_{t_k}(x) \to \int_D \eta(t,x) \,\mathrm{d}\mu_t(x)\,. \tag{4.18}$$

We use that the continuity equation (4.12) is solved for the minimizing sequence:

$$\begin{aligned}
&\int_D \eta(t_k,x) \,\mathrm{d}\mu_{t_k}^n(x) - \int_D \eta(t,x) \,\mathrm{d}\mu_t^n(x) \\
&= \int_t^{t_k} \int_D \partial_t\eta(s,x) \,\mathrm{d}\mu_s^n(x) \,\mathrm{d}s + \int_t^{t_k} \int_D \nabla_x\eta(s,x) \,\mathrm{d}\nu_s^n(x) \,\mathrm{d}s + \int_t^{t_k} \int_D \eta(s,x) \,\mathrm{d}\zeta_s^n(x) \,\mathrm{d}s\,.
\end{aligned}$$

Since, up to subsequences, for every s, μ_s^n converges weak-* to μ_s for $n \to \infty$, and (μ^n, ν^n, ζ^n) converges weak-* to (μ, ν, ζ), we obtain (4.18) for every $\eta \in C^1([0,1] \times D)$. Then for $\eta \in C([0,1] \times D)$, we choose an approximating sequence $\eta^l \in C^1([0,1] \times D)$ s.t. $\|\eta^l - \eta\|_\infty \to 0$ for $l \to \infty$. Using the triangle inequality, we get

$$\begin{aligned}
&\left|\int_D \eta(t_k, x)\, \mathrm{d}\mu_{t_k}(x) - \int_D \eta(t,x)\, \mathrm{d}\mu_t(x)\right| \\
&\leqslant \left|\int_D \eta(t_k, x)\, \mathrm{d}\mu_{t_k}(x) - \int_D \eta^l(t_k,x)\, \mathrm{d}\mu_{t_k}(x)\right| + \left|\int_D \eta^l(t_k, x)\, \mathrm{d}\mu_{t_k}(x) - \int_D \eta^l(t,x)\, \mathrm{d}\mu_t(x)\right| \\
&+ \left|\int_D \eta^l(t, x)\, \mathrm{d}\mu_t(x) - \int_D \eta(t,x)\, \mathrm{d}\mu_t(x)\right| \\
&\leqslant \|\eta^l - \eta\|_\infty (|\mu_{t_k}(D)| + |\mu_t(D)|) + \left|\int_D \eta^l(t_k, x)\, \mathrm{d}\mu_{t_k}(x) - \int_D \eta^l(t,x)\, \mathrm{d}\mu_t(x)\right| .
\end{aligned}$$

Because of equation (4.14), the masses $|\mu_{t_k}(D)|$ and $|\mu_t(D)|$ are uniformly bounded. Then, by choosing a diagonal subsequence $k(l)$, we obtain weak-* convergence of μ_{t_k} to μ_t.

Step 6: Lower Semi-Continuity Estimate. The lower semi-continuity of $\mathcal{D}_\delta$ directly follows from the general result for integral functionals on measures (see Theorem 4.3.9). More precisely, for weak-* convergent sequences of measures

$$\mu_t^n \overset{*}{\rightharpoonup} \mu_t \in \mathscr{M}^+(D)\,, \qquad \nu_t^n \overset{*}{\rightharpoonup} \nu_t \in \mathscr{M}(D, \mathbb{R}^d)\,, \qquad \zeta_t^n \overset{*}{\rightharpoonup} \zeta_t \in \mathscr{M}(D)\,.$$

we have that

$$\mathcal{D}_\delta(\mu_t, \nu_t, \zeta_t) \leqslant \liminf_{n \to \infty} \mathcal{D}_\delta(\mu_t^n, \nu_t^n, \zeta_t^n)\,.$$

Then the lower semi-continuity estimate (4.11) follows from the last formula. □

Remark 4.4.3. At first glance, the penalty functional $\int_0^1 \left(\int_D |z| d\mathscr{L} + \int_D |z^\perp|\, \mathrm{d}\mathscr{L}^\perp\right)\, \mathrm{d}t$ seems to be an appropriate choice, which allows for singular sources due to the built-in 1-homogeneity of the integrand. However, there is no equiintegrability estimate for a sequence of source terms $\left(t \mapsto \zeta_t^n(D)\right)_{n \in \mathbb{N}}$. Indeed, a uniform bound in L^1 does not suffice to deduce uniform integrability. Thus, the disintegration of the limit measure ζ remains unclear. In other words, there exists a subsequence of an energy minimizing sequence that converges weakly-* to a measure on $[0,1] \times D$, but the limit measure can not necessarily be represented in terms of a curve in $\mathscr{M}(D)$.

Now, we can rigorously define a generalized optimal transport distance $\mathcal{W}_\delta(\mu_A, \mu_B)$ for $\mu_A, \mu_B \in \mathscr{M}^+(D)$ by

$$\mathcal{W}_\delta(\mu_A, \mu_B) := \inf_{(\mu,\nu,\zeta) \in \mathcal{CE}(\mu_A, \mu_B)} \left(\mathcal{E}_\delta(\mu, \nu, \zeta)\right)^{1/2}\,. \tag{4.19}$$

The following result shows in particular that $\mathcal{W}_\delta(\mu_A, \mu_B) \in [0, \infty)$ for all $\mu_A, \mu_B \in \mathscr{M}^+(D)$.

Theorem 4.4.4 (Existence of Geodesics). *Let $\delta \in (0, \infty)$ and take $\mu_A, \mu_B \in \mathscr{M}^+(D)$. Then, there exists a minimizer $(\mu_t, \nu_t, \zeta_t)_{t \in [0,1]}$ that realizes the infimum in* (4.19)*. Moreover, $\mathcal{W}_\delta$ defines a metric on $\mathscr{M}^+(D)$, and the associated curve $(\mu_t)_{t \in [0,1]}$ is a constant speed geodesic for $\mathcal{W}_\delta$,* i.e.,

$$\mathcal{W}_\delta(\mu_s, \mu_t) = |s - t| \mathcal{W}_\delta(\mu_A, \mu_B)$$

for all $s, t \in [0,1]$.

Proof. The linear interpolation $(\mu_t = (1-t)\mu_A + t\mu_B)_{t \in [0,1]}$ together with $\nu = 0$ and $\zeta = \mu_B - \mu_A$ is an admissible triple for the set $\mathcal{CE}(\mu_A, \mu_B)$ with finite energy, since the assumptions on r imply that there exists a constant $C < \infty$ s.t.

$$\mathcal{E}_\delta(\mu, \nu, \zeta) \leqslant C\,(1 + |\mu_B - \mu_A|(D))^2 < \infty\,.$$

It follows that $\mathcal{W}_\delta(\mu_A, \mu_B) < \infty$, and the existence of a minimizer is an immediate consequence of Proposition 4.4.2. The remaining statements follow in analogy to the arguments in [DNS09, Theorem 5.4]. □

4.5 Finite Element Discretization

For the numerical discretization, we suppose for simplicity that the space domain D is polygonal, otherwise, it could be approximated by a polygonal domain. We consider a triangular mesh T_h of D with grid size h. Then, on the time-space domain $[0,1] \times D$, a tetrahedral mesh S_h is generated via subdivision of prisms $(kh,(k+1)h) \times T$ into three tetrahedrons, where $T \in T_h$ is an element of the triangular mesh of the space domain D. On the resulting tetrahedral, we define finite element spaces

$$\begin{aligned} V^1(S_h) &= \{\phi_h \colon [0,1] \times D \to \mathbb{R} \;:\; \phi_h \text{ continuous and piecewise linear on elements in } S_h\}\,, \\ V^0(S_h) &= \{\rho_h \colon [0,1] \times D \to \mathbb{R} \;:\; \rho_h \text{ piecewise constant on elements in } S_h\}\,. \end{aligned}$$

Then, we take into account the following finite element functions to discretize the measures:

$$\begin{aligned} &\rho_h \in V^0(S_h) && \text{for the mass,} \\ &m_h \in \left(V^0(S_h)\right)^d && \text{for the momentum, and} \\ &z_h \in V^1(S_h) && \text{for the source.} \end{aligned}$$

In analogy to Definition 4.4.1, the set of discrete solutions to a continuity equation is defined as follows.

Definition 4.5.1 (Discrete Weak Continuity Equation with Source Term)**.** Let $\rho_A, \rho_B \in V^0(S_h)$ be given. Then, the set $\mathcal{CE}_h(\rho_A,\rho_B)$ of solutions to a weak continuity equation with source term and boundary values ρ_A, ρ_B is given by all triples $(\rho_h, m_h, z_h) \in V_h^0(S) \times V_h^0(S)^d \times V_h^1(S)$ satisfying

$$\int_0^1 \int_D \rho_h \partial_t \phi_h + m_h \nabla_x \phi_h + z \phi_h \,\mathrm{d}x\,\mathrm{d}t = \int_D (\phi_h(1)\rho_B - \phi_h(0)\rho_A)\,\mathrm{d}x \quad \forall \phi_h \in V^1(S_h)\,.$$

Here, we use Neumann boundary condition in space, but the approach can easily be adapted to Dirichlet or periodic boundary conditions.

Next, we introduce discrete versions of the transport cost (4.1) and source cost (4.3). According to our finite element discretization, we need for the source term functional a suitable interpolation of $r(z_h)$. Therefore, we set $\mathcal{R}_h(z_h)(t,x)$ as the piecewise affine interpolation of $r(z_h((k-1)h,\cdot))$ on the triangle T for $(t,x) \in (kh,(k+1)h) \times T$ (one of the prisms underlying the tetrahedral grid). Since ρ_h and m_h are constant on each tetrahedron $S \in S_h$, we can define

$$\begin{aligned} \mathcal{E}_{\mathrm{trans},h}(\rho_h, m_h) &:= \int_0^1 \int_D \Phi(\rho_h, m_h)\,\mathrm{d}x\,\mathrm{d}t\,, \\ \mathcal{E}_{\mathrm{source},h}(z_h) &:= \int_0^1 \left(\int_D \mathcal{R}_h(z_h)\,\mathrm{d}x \right)^2 \mathrm{d}t\,, \\ \mathcal{E}_{\delta,h}(\rho_h, m_h, z_h) &:= \mathcal{E}_{\mathrm{trans},h}(\rho_h, m_h) + \frac{1}{\delta}\mathcal{E}_{\mathrm{source},h}(z_h)\,. \end{aligned}$$

Then, a discrete version of the minimization problem (4.19) is given by

$$\mathcal{W}_{\delta,h}(\rho_A \mathscr{L}, \rho_B \mathscr{L}) := \inf_{(\rho_h, m_h, z_h) \in \mathcal{CE}_h(\rho_A,\rho_B)} \mathcal{E}_{\delta,h}(\rho_h, m_h, z_h)^{\frac{1}{2}}\,. \tag{4.20}$$

Remark 4.5.2. Numerically, we are not able to treat singular measures as presented in Section 4.4. However, such measures can be obtained in the limit for a mesh size $h \to 0$. For example, on a fixed mesh, a line source can be approximated via sources with a support of thickness $2h$.

Remark 4.5.3. In the implementation, we restrict $D = [0,1]^2$ to be the unit square. Then, we use a tetrahedral mesh of the time-space domain by subdividing cubes of side length h into six tetrahedrons.

4.6 Proximal Splitting Algorithm

We recall from Section 3.2.3 that proximal splitting algorithms can be used to solve the classical L^2-optimal transport problem numerically, as it was, *e.g.*, proposed in [PPO14]. Now, we intend to apply a proximal splitting algorithm to solve the fully discrete optimization problem (4.20). Therefore, we define the two functionals

$$\begin{aligned}\mathcal{F}(\rho_h, m_h, z_h) &:= \mathcal{E}_{\delta,h}(\rho_h, m_h, z_h)\,,\\ \mathcal{G}(\rho_h, m_h, z_h) &:= \mathcal{I}_{\mathcal{CE}_h(\rho_A,\rho_B)}(\rho_h, m_h, z_h)\,.\end{aligned}$$

More precisely, by adding the indicator function of the set of solutions to the continuity equation $\mathcal{CE}(\mu_A, \mu_B)$ to the kinetic energy functional, the constrained optimization problem (4.20) can be rewritten as an unconstrained minimization problem. We recall from Definition 3.2.4 that the proximal mapping of a convex and lower semi-continuous function f is given by $\operatorname{prox}_f(x) = \arg\min_{y\in X} f(y) + \frac{1}{2}\|x-y\|_H^2$, where H is a suitable Hilbert space. Here, for a triple $(\rho_h, m_h, z_h) \in V_h^0(S_h) \times V_h^0(S_h)^d \times V_h^1(S_h)$, we choose a weighted L^2-norm

$$\|(\rho_h, m_h, z_h)\| := \left(\int_0^1 \int_D |\rho_h|^2 + |m_h|^2 + \frac{1}{\delta}|z_h|^2 \,\mathrm{d}x\,\mathrm{d}t\right)^{\frac{1}{2}}\,,$$

which can be computed exactly by choosing a quadrature rule of at least second order.

In the following, we compute the proximal mappings of $\mathcal{F}$ and $\mathcal{G}$. We show that the computation of $\operatorname{prox}_{\mathcal{G}}$ requires to solve an elliptic problem on the time-space domain and that the computation of $\operatorname{prox}_{\mathcal{F}}$ is rather simple. Finally, we use the Douglas–Rachford algorithm (3.10), which was also applied in [PPO14] for the classical L^2-optimal transport problem.

4.6.1 Projection onto the Set $\mathcal{CE}_h(\rho_A, \rho_B)$

Since $\mathcal{CE}_h(\rho_A, \rho_B)$ is a convex set, we recall from Lemma 3.2.10 that the proximal mapping of the indicator function of $\mathcal{CE}_h(\rho_A, \rho_B)$ can be computed by the orthogonal projection. More precisely, to project a point $(p_h = (\rho_h, m_h), z_h) \in V^0(S)^{d+1} \times V^1(S)$ onto $\mathcal{CE}_h(\rho_A, \rho_B)$, this requires to solve

$$(p_h^{\mathrm{pr}}, z_h^{\mathrm{pr}}) = \operatorname{proj}_{\mathcal{CE}_h(\rho_A,\rho_B)}(\rho_h, m_h)(p_h, z_h) = \underset{(q_h,w_h)\in\mathcal{CE}_h(\rho_A,\rho_B)}{\arg\min} \|(p_h, z_h) - (q_h, w_h)\|^2\,. \tag{4.21}$$

The solution to this constrained optimization problem is given in the following.

Proposition 4.6.1 (Projection onto Generalized Solutions to the Continuity Equation)**.** *The solution* (p_h^{pr}, z_h^{pr}) *to the projection problem* (4.21) *is given by*

$$p_h^{pr} = p_h + \frac{1}{2}\nabla_{(t,x)}\phi_h^{pr}\,, \quad z_h^{pr} = z_h + \frac{\delta}{2}\phi_h^{pr}\,, \tag{4.22}$$

where $\phi_h^{pr} \in V_h^1(S)$ *is defined by solving*

$$\int_0^1 \int_D \frac{1}{2}\nabla_{(t,x)}\phi_h^{pr}\nabla_{(t,x)}\psi_h + \frac{\delta}{2}\phi_h^{pr}\psi_h \,\mathrm{d}x\,\mathrm{d}t = \int_D \psi_h(1)\rho_B - \psi_h(0)\rho_A \,\mathrm{d}x - \int_0^1\int_D z_h\psi_h + p_h\nabla_{(t,x)}\psi_h \,\mathrm{d}x\,\mathrm{d}t$$

for all $\psi_h \in V_h^1(S)$.

Proof. The associated Lagrangian to the minimization problem (4.21) is given by

$$\mathcal{L}(q_h, w_h, \psi_h) = \|(p_h, z_h) - (q_h, w_h)\|^2 - \int_0^1\int_D q_h \cdot \nabla_{(t,x)}\psi_h + w_h\psi_h \,\mathrm{d}x\,\mathrm{d}t + \int_D \psi_h(1)\rho_B - \psi_h(0)\rho_A \,\mathrm{d}x\,,$$

with a Lagrange multiplier $\psi_h \in V^1(S_h)$. In terms of the Lagrangian, the projection problem can be written as a saddle point problem, where we ask for $(p_h^{\mathrm{pr}}, z_h^{\mathrm{pr}}, \phi_h^{\mathrm{pr}}) \in V^0(S)^{d+1} \times V^1(S) \times V^1(S)$ s.t.

$$\mathcal{L}\left(p_h^{\mathrm{pr}}, z_h^{\mathrm{pr}}, \phi_h^{\mathrm{pr}}\right) = \min_{(q_h,w_h)\in V_h^0(S)^{d+1}\times V_h^1(S)} \max_{\psi_h\in V_h^1(S)} \mathcal{L}(q_h, w_h, \psi_h)\,.$$

The Euler–Lagrange equations corresponding to this saddle point problem are given by

$$\int_0^1 \int_D p_h^{\mathrm{pr}} \cdot \nabla_{(t,x)} \psi_h + z_h^{\mathrm{pr}} \psi_h \,\mathrm{d}x\,\mathrm{d}t = \int_D \psi_h(1)\,\rho_B - \psi_h(0)\,\rho_A \,\mathrm{d}x \qquad \forall \psi_h \in V_h^1(S)\,, \tag{4.23}$$

$$\int_0^1 \int_D q_h \cdot \nabla_{(t,x)} \phi_h^{\mathrm{pr}} \,\mathrm{d}x\,\mathrm{d}t = \int_0^1 \int_D 2(p_h^{\mathrm{pr}} - p_h)\, q_h \,\mathrm{d}x\,\mathrm{d}t \qquad \forall q_h \in V_h^0(S)^{d+1}\,, \tag{4.24}$$

$$\int_0^1 \int_D \phi_h^{\mathrm{pr}} w_h \,\mathrm{d}x\,\mathrm{d}t = \int_0^1 \int_D \frac{2}{\delta}(z_h^{\mathrm{pr}} - z_h)\, w_h \,\mathrm{d}x\,\mathrm{d}t \qquad \forall w_h \in V_h^1(S)\,. \tag{4.25}$$

Testing (4.24) with $q_h = \nabla_{(t,x)}\psi_h$ and then using (4.23) gives

$$\begin{aligned}\int_0^1 \int_D \frac{1}{2} \nabla_{(t,x)} \phi_h^{\mathrm{pr}} \cdot \nabla_{(t,x)} \psi_h \,\mathrm{d}x\,\mathrm{d}t &= \int_0^1 \int_D (p_h^{\mathrm{pr}} - p_h) \cdot \nabla_{(t,x)} \psi_h \,\mathrm{d}x\,\mathrm{d}t \\ &= \int_D \psi_h(1)\rho_B - \psi_h(0)\rho_A \,\mathrm{d}x - \int_0^1 \int_D z_h^{\mathrm{pr}} \psi_h + p_h \cdot \nabla_{(t,x)} \psi_h \,\mathrm{d}x\,\mathrm{d}t\,.\end{aligned}$$

Hence, by applying (4.25), which leads to $z_h^{\mathrm{pr}} = z_h + \frac{\delta}{2}\phi_h^{\mathrm{pr}}$, we obtain for all $\psi_h \in V_h^1(S)$ that

$$\int_0^1 \int_D \frac{1}{2} \nabla_{(t,x)} \phi_h^{\mathrm{pr}} \nabla_{(t,x)} \psi_h + \frac{\delta}{2} \phi_h^{\mathrm{pr}} \psi_h \,\mathrm{d}x\,\mathrm{d}t = \int_D \psi_h(1)\rho_B - \psi_h(0)\rho_A \,\mathrm{d}x - \int_0^1 \int_D z_h \psi_h + p_h \nabla_{(t,x)} \psi_h \,\mathrm{d}x\,\mathrm{d}t\,.$$

This system can be solved in ϕ_h^{pr}. Finally, the solution to the projection problem is given by (4.22). □

4.6.2 Proximal Mappings of Transport and Source Term Cost

The functional $\mathcal{E}_{\delta,h}$ is composed of the transport cost $\mathcal{E}_{\mathrm{trans},h}$, which only depends on ρ_h and m_h, and the source term cost $\mathcal{E}_{\mathrm{source},h}$, which only depends on z_h. Thus, we can compute these proximal mappings separately.

Proximal Mapping of Transport Cost. We note that ρ_h and m_h are constant on each tetrahedron of the simplicial mesh S_h. Thus, as for the classical L^2-optimal transport distance (*cf*. Proposition 3.2.12), the proximal map of the kinetic energy $\mathcal{E}_{\mathrm{trans},h}$ can be computed by projecting for each tetrahedron the associated value onto a convex set $\mathcal{B}$ as defined in (3.14).

Proximal Mapping of Source Term Cost. We discuss different choices for the source term cost functional.

First, we consider an L^2-norm both in time and space, which was studied in [MRSS15]. In this case, for a step size $\gamma > 0$, we easily get a pointwise update

$$\mathrm{prox}_{\frac{\gamma}{\delta}|\cdot|^2}(z)(t,x) = \operatorname*{arg\,min}_{w\in\mathbb{R}} \frac{1}{\delta}|w|^2 + \frac{1}{\delta}|w - z(t,x)|^2 = \frac{1}{1+\gamma} z(t,x)\,.$$

For a source term in the L^1-norm both in time and space, following computations in [Ess09], we also get a pointwise update for the proximal operator of the $L^1(L^1)$-norm, which is given by

$$\mathrm{prox}_{\frac{\gamma}{\delta}|z|}(z)(t,x) = \begin{cases} 0 & \text{if } |z(t,x)| \leqslant \frac{\gamma}{2}\,, \\ z(t,x) - \frac{\gamma}{2}\mathrm{sgn}(z(t,x)) & \text{otherwise.} \end{cases}$$

Thus, a numerical scheme for a source term in $L^1(L^1)$ would be as simple as for a source term in $L^2(L^2)$. However, the existence of geodesics is not guaranteed (see Remark 4.4.3).

In the case of a linear growth function $r(\cdot)$, the minimization problem to compute the proximal map only decouples in time but not in space. More precisely, for each discrete time step k, we have to solve

$$\operatorname*{arg\,min}_{w_h(kh,\cdot)\in V^1(T_h)} \frac{\gamma}{\delta} \left(\int_D \mathcal{R}_h(w_h)(kh,x)\,\mathrm{d}x \right)^2 + \frac{1}{2\delta} \int_D |w_h(kh,x) - z_h(kh,x)|^2 \,\mathrm{d}x\,. \tag{4.26}$$

Then, we solve the minimization problem (4.26) via a gradient descent method, which requires that r is differentiable. For example, this is not the case for a source term in the $L^2(L^1)$-norm, since $r(z) = |z|$ is not differentiable. For our numerical computations, we restrict r to be the Huber function (4.4).

4.7 Numerical Results for Generalized Optimal Transport Geodesics

We present our numerical results for geodesics w.r.t. the generalized optimal transport distance $\mathcal{W}_\delta$. The computational domain is always given by the unit square $D = [0,1]^2$. For the finite element discretization, we choose a grid size of $h = 2^{-7}$ on the time-space domain $[0,1]^3$, which implies a temporal discretization with $\frac{1}{h} = 128$ time steps, and we show extractions at time steps $t = \frac{16i}{128}$ for $i = 0,\dots,8$. We use piecewise linear RGB scales to plot the mass variable (0 (white), 0.5 (light blue), 1 (blue)) and the source term (minimal value (green), 0 (white), maximal value (purple)). From (4.6), we recall the notation $L^2(L^2)$ for an L^2-norm penalization of the source term both in time and space and $L^2(H)$ for the L^2-Huber cost functional with the function r as defined in (4.4), *i.e.*,

$$\mathcal{E}_{\mathrm{source},L^2(L^2)}(z) = \int_0^1 \int_D |z|^2 \,\mathrm{d}x\,\mathrm{d}t\,, \qquad \mathcal{E}_{\mathrm{source},L^2(H)}(z) = \int_0^1 \left(\int_D r(z)\,\mathrm{d}x\right)^2 \mathrm{d}t\,.$$

4.7.1 Comparison with the $L^2(L^2)$-Model

Here, we compare the $L^2(H)$-model with the $L^2(L^2)$-source term functional. Therefore, we consider both singular and absolutely continuous measures. For the penalty parameter for the source term functional, we choose $\delta = 1$.

Generation of Approximatively Singular Measures. The source term cost functional (4.3) for the $L^2(H)$-model has been chosen s.t. singular sources in space are allowed, which is not possible for an $L^2(L^2)$-model, where a singular source always has infinite path energy. However, singular sources cannot be implemented directly with our finite element discretization, but in Figure 4.1, we study the transport between measures supported on a thin rectangular strip as an approximation of a singular measure. The densities ρ_A and ρ_B are constant on this rectangle but have different intensity values. Our model with the $L^2(H)$ cost functional for the source term is able to generate the thin rectangles directly, s.t. the corresponding geodesic is given by a blending of the two measure ρ_A and ρ_B. Instead, for an $L^2(L^2)$ source term, which was proposed in [MRSS15], the generation of mass takes place on a thick superset of the rectangular strip and is then transported towards the strip. In particular, this is visible by considering the geodesic interpolation at intermediate time steps $t \in (0,1)$, where the rectangle is blurred. This effect is furthermore reflected by considering the corresponding source terms.

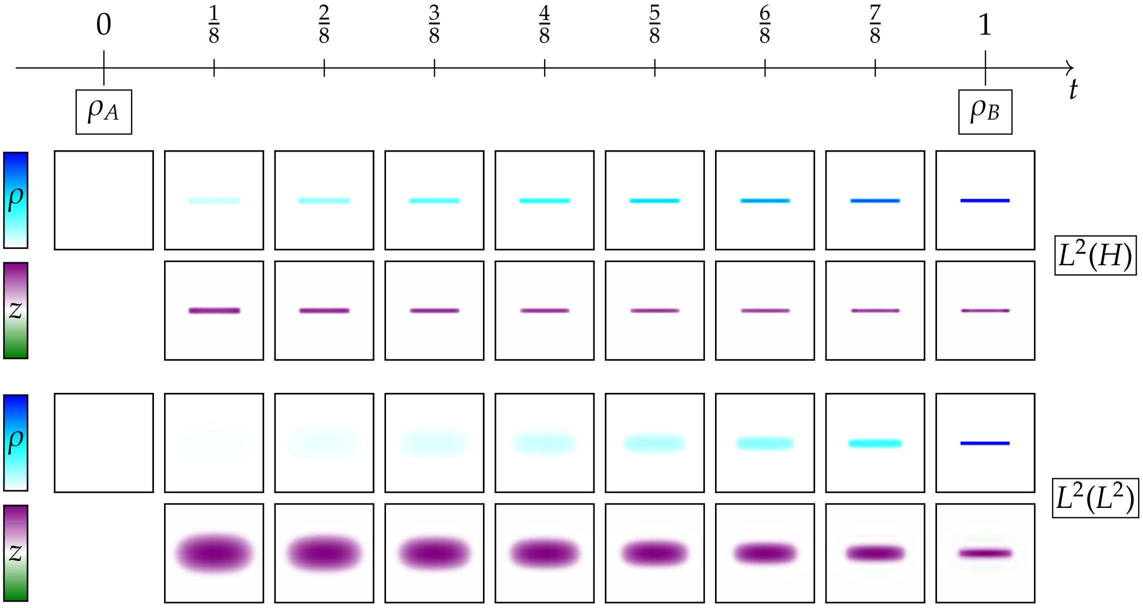


Figure 4.1: Generalized optimal transport geodesic between characteristic functions of thin rectangles with different intensities as approximation of singular measures. Here, the source term parameter is given by $\delta = 10^0$.

Generation of Absolute Continuous Measures. As for the thin rectangle, we observe a similar effect for, now substantially, absolutely continuous measures. In Figure 4.2, we compare the $L^2(H)$ and the $L^2(L^2)$ source term for a geodesic interpolation between differently scaled characteristic functions of a square. Again, the resulting geodesic for the $L^2(H)$-model is given by a blending of the two measure ρ_A and ρ_B, whereas in the $L^2(L^2)$-model the additional mass is generated on a larger support. In Figure 4.3, we show a plot of the map $t \mapsto \int_D |z(t,\cdot)| \, \mathrm{d}x$ for both models, where it turns out that this L^1-norm of the source term in space is constant for the $L^2(H)$-model. Indeed, the larger support is advantageous for the $L^2(L^2)$ source term, since, compared to a pure blending, the L^2-norm in space is smaller. Thus, if transporting mass is comparably cheap, using a constant distribution of the source term on the full domain D becomes more favorable. To balance the interaction between the kinetic energy and the source term cost, we can choose the parameter δ appropriately.

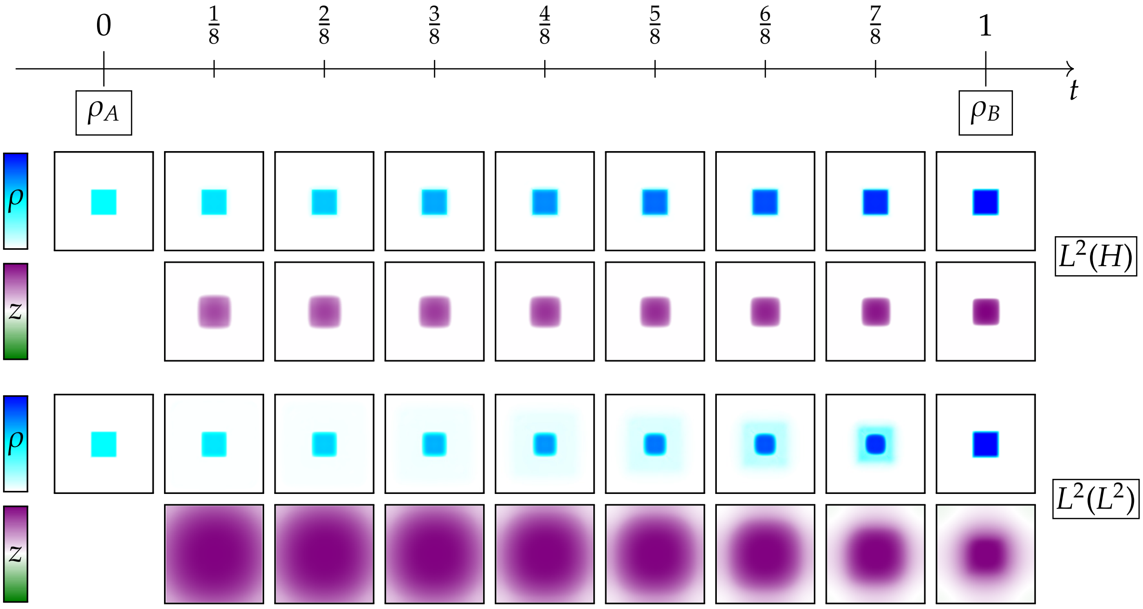


Figure 4.2: Generalized optimal transport geodesic and corresponding source terms between two characteristic functions of squares with different intensities. Here, the source term parameter is given by $\delta = 10^0$.

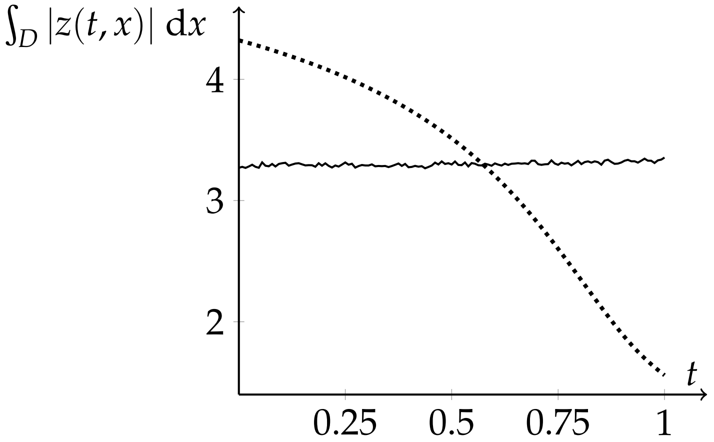


Figure 4.3: Distribution of the L^1-norm of the source term in time for the example in Figure 4.2 of characteristic functions of squares (dotted line: $L^2(L^2)$, continuous line: $L^2(H)$).

4.7.2 Effect of the Source Term Penalization Parameter δ

Next, we investigate the effect of the penalty parameter δ for the source term. Here, we restrict to the $L^2(H)$-model, but we remark that similar effects are obtained for the $L^2(L^2)$-model.

Transport versus Blending. In Figure 4.4, we choose as input data ρ_A at time $t = 0$ a characteristic function of a square and as input data ρ_B at time $t = 1$ a sum of two characteristic functions of squares, where one square is the same as for ρ_A, and the other square is translated. Now, there are two obvious transport paths connecting ρ_A and ρ_B, namely the curve, which blends the second square and the curve, which transports a part of the second square and blends the remaining measure. Indeed, we observe both scenarios as limit cases. For $\delta \to \infty$, transport becomes expensive. In Figure 4.4, we observe a simple blending for large values of δ. In contrast, for $\delta \to 0$ transport becomes cheaper, which is reflected by the computational results for small δ in Figure 4.4. For intermediate values of δ, we obtain transport paths, where only a small part of the second square is transported.

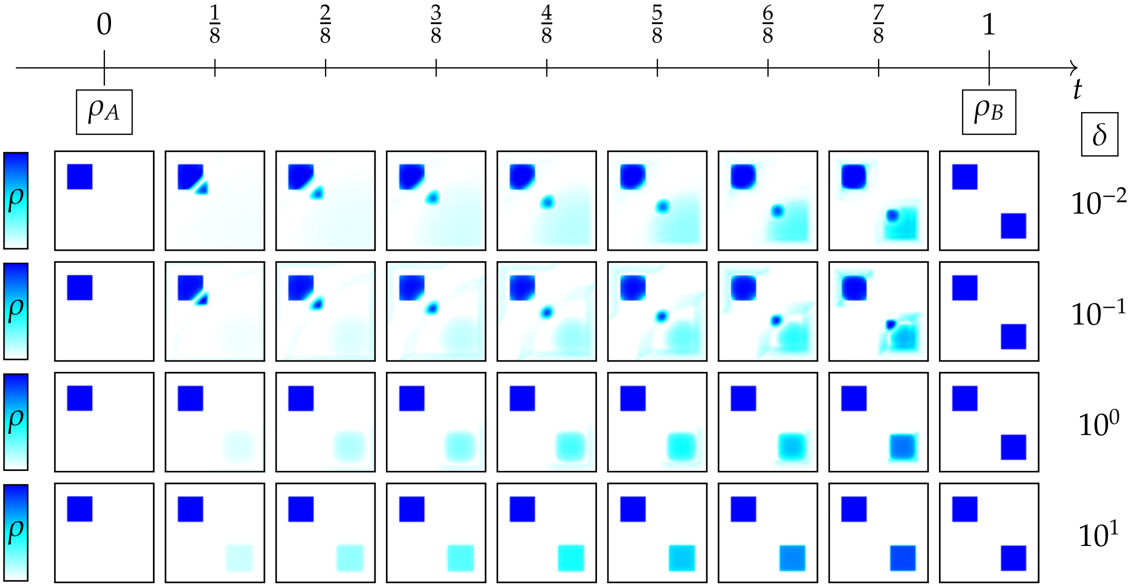


Figure 4.4: Generalized optimal transport geodesic between a characteristic function on a square and a characteristic functions of two squares. We choose (from top to bottom) $\delta = 10^{-2}, 10^{-1}, 10^{0}, 10^{1}$.

Positive and Negative Sources. In Figure 4.5, we show another example to study the effect of the penalty parameter for the source term. Here, the input data ρ_A at time $t = 0$ consists of three scaled characteristic functions of balls, where one of these balls has a smaller density value than the other two. The input data ρ_B at time $t = 1$ is based on the identical geometric configuration, but with swapped densities, *i.e.*, the other two balls have a smaller density value. For the generalized optimal transport geodesic, we observe that mass is transported from the two balls with higher density at time $t = 0$ to the ball with higher density at time $t = 1$. Note that this amount of transported mass depends on the parameter δ. At the same time, a blending of the transported masses as a compensation for the unbalanced total mass can be observed. This example demonstrates that for a geodesic path, the source term variable can achieve both positive and negative values at the same time. Moreover, in Figure 4.6, we show plots of the integrated source term. A striking observation in Figure 4.3 and Figure 4.6 is that $t \mapsto \int_D |z(t,\cdot)| \, \mathrm{d}x$ is approximately constant in time for the $L^2(H)$-model, which is in contrast to the $L^2(L^2)$-model, as indicated in Figure 4.3.

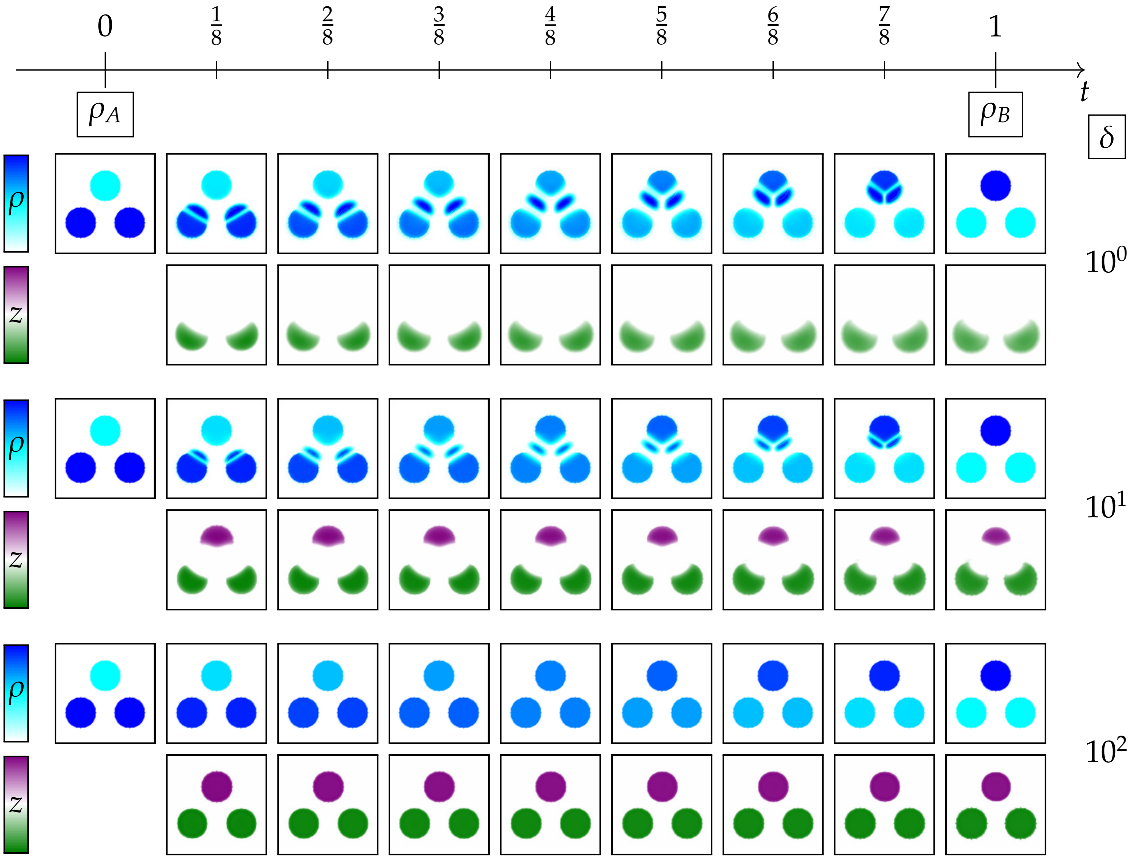


Figure 4.5: Generalized optimal transport geodesic with corresponding distribution of the source term in time between three scaled characteristic functions of balls with different densities. Here, the source term parameters are given (from top to bottom) by $\delta = 10^0, 10^1, 10^2$.

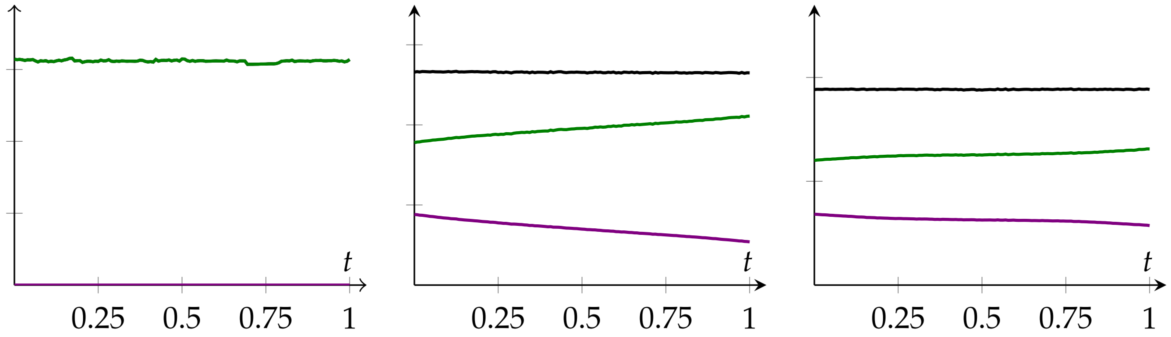


Figure 4.6: Distribution of the L^1-norm of the source term in time for the example in Figure 4.5. We depict $\int_D |z(t,\cdot)|\,\mathrm{d}x$ (black), $\int_D z^+(t,\cdot)\,\mathrm{d}x$ (purple), and $\int_D z^-(t,\cdot)\,\mathrm{d}x$ (green). The source term parameters are (from left to right) $\delta = 10^0, 10^1, 10^2$.

Periodic Boundary Conditions. In Figure 4.7, we investigate periodic boundary conditions in space for different values of the source term parameter. Here, the input data ρ_A at time $t = 0$ is given by a bump function in the center $(0.5, 0.5)$ of the periodic cell, *i.e.*,

$$\rho_A(x_1, x_2) = \begin{cases} \exp\left(1 - \left(1 - \sigma^{-2}(x_1 - 0.5)^2 - \sigma^{-2}(x_2 - 0.5)\right)^{-1}\right) & \text{if } (x_1 - 0.5)^2 + (x_2 - 0.5)^2 < \sigma^2\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

where $\sigma = 0.75$. For the input data ρ_B at time $t = 1$, we choose a bump function with the center at $(0, 0)$ and $\sigma = 0.5$. A similar example was already considered in [BB00], where periodically extended Gaussian probability measures were taken into account. For equal size, the classical optimal transport geodesic is given by splitting the bump of ρ_A into four parts and transporting these parts to the four corners, which has effectively lower kinetic energy as the translation. Also, for the generalized optimal transport distance, we obtain a splitting of the bump function. Moreover, depending on the parameter δ, the unbalance of mass between ρ_A and ρ_B is blended during the transport (for small values of δ) or on the support of ρ_A (for intermediate values of δ). As in Figure 4.4, for even larger values of δ (which we do not show in Figure 4.7), we obtain a pure blending of both bumps.

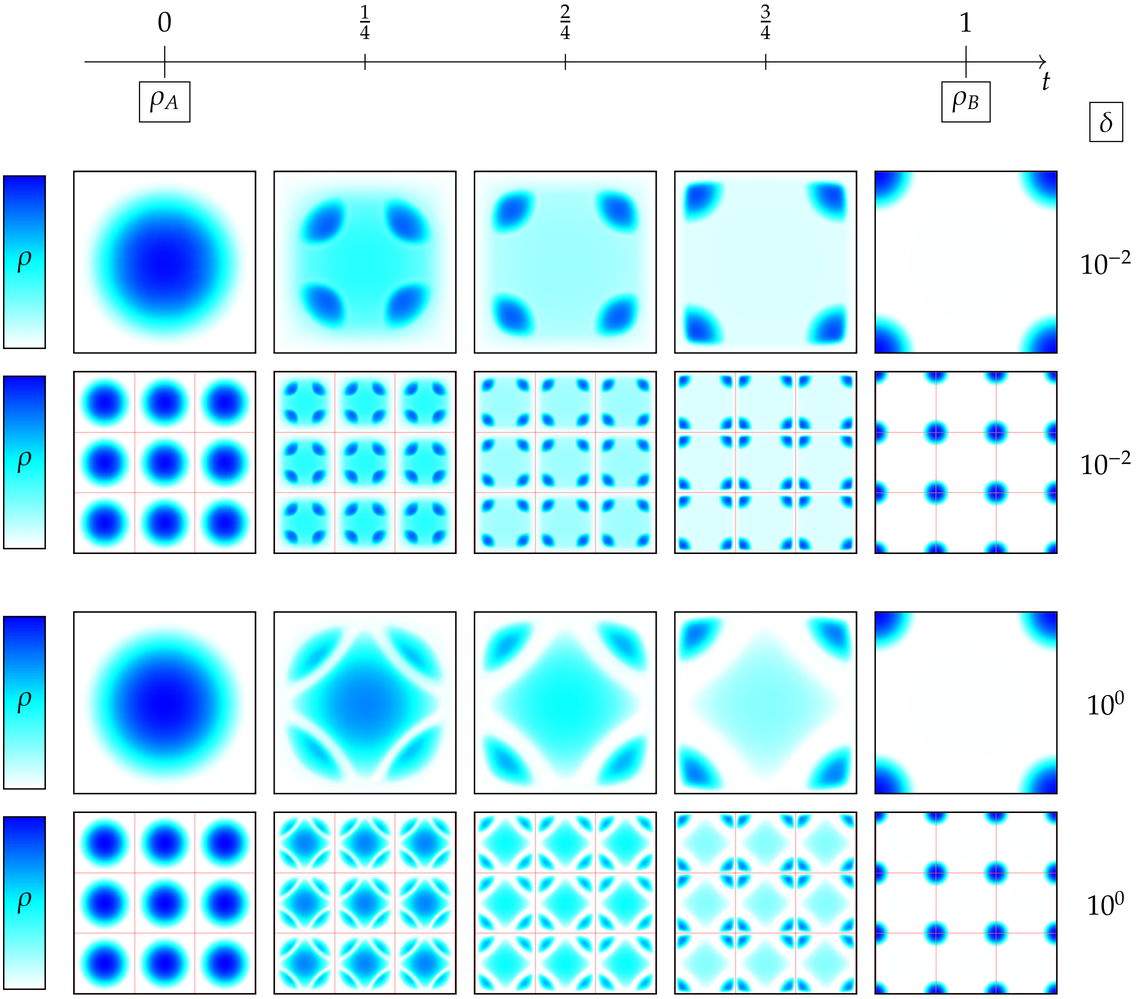


Figure 4.7: Generalized optimal transport geodesic connecting two translated bump functions. We compare two different values of the source term penalty parameter. In both cases, we depict single periodic cells, which is our computational domain. Furthermore, to pronounce the periodicity, we extended the periodic cells to 3 × 3 blocks.

4.7.3 Application to Textures

Finally, in Figure 4.8, we depict examples of generalized optimal transport geodesics between images of wood textures and marble textures. We choose the $L^2(H)$ source term cost functional and $\delta = 10^{-1}$. The grid size is given by $h = 2^{-8}$. In both cases, the interpolated images on the geodesic paths could be interpreted as realistic textures.

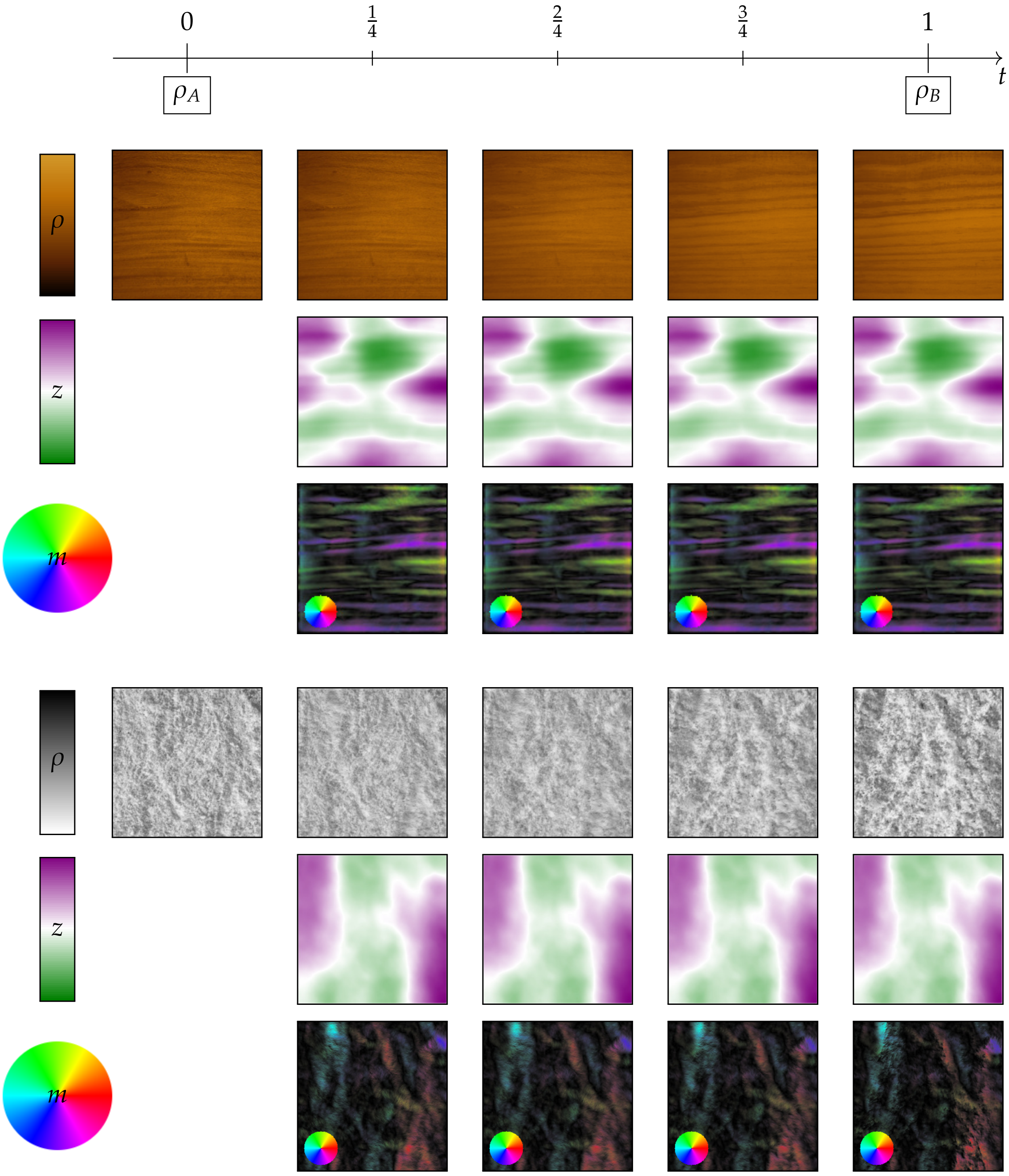


Figure 4.8: Generalized optimal transport geodesics between textures of wood (top) and marble (bottom) with corresponding source terms (positive values in purple and negative values in green) and momenta (color-code given by the wheel on the lower left, which indicates both the direction and the magnitude).

4.8 Conclusion and Outlook

We have developed a new generalized optimal transport model with source term, which is based on the Benamou–Brenier formulation. To incorporate singular sources, we have chosen a functional with linear growth to penalize the source term in space, whereas an L^2-norm in time has allowed an equiintegrability estimate to prove the existence of generalized optimal transport geodesics in the space of Radon measures. Selected numerical test cases have shown strikingly different behavior compared to a source term penalized in an L^2-norm both in time and space.

Note that an extension of our computational method to weighted Riemannian barycenters w.r.t. the generalized optimal transport distance would be straightforward. For the classical L^2-Wasserstein distance, also a discrete geodesic extrapolation (*i.e.*, the time-discrete exponential map) of an initial probability measure in a direction given by another probability measure can be directly obtained from a geodesic interpolation because of the displacement convexity formula. This property is unclear for our generalized model. Furthermore, an extension to Riemannian splines, as for discrete shells [HRW17] based on the general time-discrete framework in [RW15] on Banach manifolds, would be interesting even for the classical Wasserstein distance.

Finally, we want to point out that the range of applications for realistic images seems to be somewhat limited. In particular, an interpolation between images of human faces usually looks quite blurry. Instead, amazing results were obtained for the metamorphosis model [BER15]. Currently, the optimal transport distance has been explored in the quickly developing field of machine learning. In [SHB$^+$18], the reconstruction of images as barycenters of dictionary atoms w.r.t. the entropy regularized Wasserstein distance was performed.

Chapter 5

Optimal Transport on Graphs

In Chapter 3, we have introduced the L^2-Wasserstein distance between probability measures on a convex domain in the Euclidean space. There, we have seen that for absolutely continuous measures, the formulations of Monge, Kantorovich, and Benamou–Brenier coincide and minimizing paths of the Benamou–Brenier functional are constant speed geodesics. This differential geometric interpretation allows, *e.g.*, transferring the concept of gradient flows to the Wasserstein space [JKO98]. Furthermore, the definition of the L^2-Wasserstein distance on more general space domains is often straightforward, as far as there is a notation of measures and distance. However, on a discrete space, the formulations of Monge and Kantorovich imply that constant speed geodesics must be constant paths. Instead, Maas [Maa11] developed an L^2-Wasserstein metric on the space of probability measures on discrete spaces given by an irreducible and reversible Markov transition kernel by taking into account an appropriate Benamou–Brenier formulation. Remarkably, as for the classical optimal transport distance, it was verified that the gradient flow of the entropy can be identified with the heat equation on the Markov kernel.

In this chapter, the main focus lies on investigating a numerical scheme to approximate the Wasserstein distance on discrete spaces. To compute minimizing paths of the classical Benamou–Brenier functional, proximal splitting algorithms have turned out to be an efficient tool [BB00, PPO14], where a suitable discretization of the mass and momentum variables allows decoupling the computation of the proximal operator of the kinetic energy into pointwise projections. Now, for the optimal transport distance on discrete spaces as defined in [Maa11], the mass variable is defined at nodes, and the momentum variable is considered on edges. Then, an averaging operator from a pair of nodes to its common edge is required to define a corresponding kinetic energy, which unfortunately couples all variables in space. In [SRGB16], a similar optimal transport distance on graphs was investigated, where the special structure of the harmonic mean was used, s.t. the proximal operator of the kinetic energy functional can be computed as for the classical optimal transport. Here, we present a fully discrete approximation for a generic class of averaging operators. In particular, to recover the heat equation as a gradient flow, the logarithmic mean has to be taken into account. To decouple the optimization problem, we introduce several auxiliary variables, s.t. the core ingredient of our numerical algorithm is a projection onto a three-dimensional set defined by the respective mean.

This chapter is organized as follows. In Section 5.1, we recall the Wasserstein distance on discrete spaces introduced by Maas. We derive certain a priori bounds for corresponding geodesic paths in Section 5.2. For a fully discrete approximation, we choose a finite element discretization in Section 5.3, for which we prove Γ-convergence in Section 5.4. Then, we investigate a numerical computation scheme via a proximal splitting algorithm in Section 5.5. In Section 5.6, we present our numerical results for geodesic paths. Finally, in Section 5.7, we consider gradient flows w.r.t. the optimal transport distance on graphs. We show that a minimizing movement scheme to compute a gradient step can be solved by a proximal splitting algorithm, and we compare our numerically computed gradient flow trajectory with the solution to the heat equation.

Remark 5.0.1 (Collaborations and Publications). All results presented in this chapter are joint work with Matthias Erbar, Martin Rumpf, and Bernhard Schmitzer and will be published in [ERSS17].

5.1 A Benamou–Brenier Formula on Graphs

Here, we define the Benamou–Brenier formulation for the L^2-Wasserstein metric on the space of probability measures on discrete spaces as introduced in [Maa11].

Irreducible and Reversible Markov Chains. We denote by $\mathcal{X}$ a finite set, which can be interpreted as the set of nodes of a graph. Furthermore, let $Q\colon \mathcal{X} \times \mathcal{X} \to [0,\infty)$ be the transition rate matrix of a Markov chain on $\mathcal{X}$. More precisely, the corresponding graph has a directed edge $(x,y) \in \mathcal{X} \times \mathcal{X}$ if $Q(x,y)$ is positive. Then, the set of edges indicated by nonzero transition probability is given by

$$\mathcal{S} = \{(x,y) \in \mathcal{X} \times \mathcal{X} \ : \ Q(x,y) > 0\}\,.$$

Here, we make the assumption that $Q(x,x) = 0$ for all $x \in \mathcal{X}$, since, for an optimal transport path, a loop would not be taken into account. We suppose that Q is irreducible or equivalently that the corresponding graph is strongly connected. The irreducibility condition implies that there exists a unique stationary distribution $\pi\colon \mathcal{X} \to (0,1]$ of the Markov chain with $\sum_{x\in\mathcal{X}} \pi(x) = 1$. Furthermore, we assume that Q is reversible w.r.t. π, *i.e.*, the detailed balance condition $\pi(x)Q(x,y) = \pi(y)Q(y,x)$ holds for all $x,y \in \mathcal{X}$. The reversibility condition implies that a directed edge $(x,y) \in \mathcal{S}$ has nonzero transition probability if and only if this is the case for the edge $(y,x) \in \mathcal{S}$ in the opposite direction. Later, we make use of the following rates of the Markov kernel:

$$C^* := \max_{x\in\mathcal{X}} \sum_{y} Q(x,y)\,, \tag{5.1}$$

$$C_* := \min_{x,y\in\mathcal{X}, Q(x,y)>0} Q(x,y)\pi(x)\,. \tag{5.2}$$

Now, the set of probability densities on $\mathcal{X}$ w.r.t. π is given by

$$\mathscr{P}(\mathcal{X}) := \left\{\rho\colon \mathcal{X} \to \mathbb{R}_{\geqslant 0} : \sum_{x\in\mathcal{X}} \pi(x)\rho(x) = 1\right\}\,.$$

As for classical optimal transport, the condition $\sum_{x\in\mathcal{X}} \pi(x)\rho(x) = 1$ can be replaced by $\sum_{x\in\mathcal{X}} \pi(x)\rho(x) = c$ for any $c \in \mathbb{R}_+$, but for simplicity we restrict to the case $c = 1$.

Differential Operators on Graphs. Next, we consider functions $\phi\colon \mathcal{X} \to \mathbb{R}$ on nodes and $\Phi\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on edges, which we also identify with vectors in $\mathbb{R}^{\mathcal{X}}$ and $\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, respectively. First, we define inner products on $\mathbb{R}^{\mathcal{X}}$ and $\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ by

$$\langle \phi, \psi\rangle_\pi := \sum_{x\in\mathcal{X}} \phi(x)\psi(x)\pi(x)\,, \qquad \langle \Phi, \Psi\rangle_Q := \frac{1}{2}\sum_{x,y\in\mathcal{X}} \Phi(x,y)\Psi(x,y)Q(x,y)\pi(x)$$

for ϕ, $\psi \in \mathbb{R}^{\mathcal{X}}$ and Φ, $\Psi \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, and denote the corresponding induced norms by $\|\cdot\|_\pi$ and $\|\cdot\|_Q$. Then, we introduce discrete differential operators. A discrete gradient $\nabla_{\mathcal{X}}\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ and a discrete divergence $\mathrm{div}_{\mathcal{X}}\colon \mathbb{R}^{\mathcal{X}\times\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ are given by

$$(\nabla_{\mathcal{X}}\psi)(x,y) := \psi(x) - \psi(y), \qquad (\mathrm{div}_{\mathcal{X}}\Psi)(x) := \frac{1}{2}\sum_{y\in\mathcal{X}} Q(x,y)(\Psi(y,x) - \Psi(x,y)).$$

Note that the discrete integration by parts formula

$$\langle \phi, \mathrm{div}_{\mathcal{X}}\Psi\rangle_\pi = -\langle \nabla_{\mathcal{X}}\phi, \Psi\rangle_Q$$

can easily be verified s.t. duality between discrete gradient and divergence holds. The associated discrete Laplace-operator $\Delta_{\mathcal{X}}\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ is given by

$$\Delta_{\mathcal{X}}\psi(x) := \mathrm{div}_{\mathcal{X}}(\nabla_{\mathcal{X}}\psi)(x) = \sum_{y\in\mathcal{X}} Q(x,y)\,[\psi(y) - \psi(x)] = (Q-D)\psi(x)\,,$$

where $D = \mathrm{diag}(\sum_{y\in\mathcal{X}} Q(x,y))_{x\in\mathcal{X}}$.

Optimal Transport Distance on Graphs. Having these differential operators at hand, we are able to formulate a continuity equation for time-dependent probability densities $\rho\colon [0,1] \times \mathbb{R}^{\mathcal{X}} \to \mathbb{R}_+$ and momenta $m\colon [0,1] \times \mathbb{R}^{\mathcal{X}\times\mathcal{X}} \to \mathbb{R}$ describing the flow of mass along the graph edges. In the following, we frequently identify these functions ρ and m by functions $\rho\colon [0,1] \to \mathbb{R}^{\mathcal{X}}$ and $m\colon [0,1] \to \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$.

Definition 5.1.1 (Continuity Equation on Graphs)**.** The set $\mathcal{CE}(\rho_A, \rho_B)$ of solutions to the continuity equation for given boundary data ρ_A, $\rho_B \in \mathscr{P}(\mathcal{X})$ is defined as the set of all pairs (ρ, m) with $\rho\colon [0,1] \times \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ and $m\colon [0,1] \times \mathbb{R}^{\mathcal{X}\times\mathcal{X}} \to \mathbb{R}$ measurable s.t.

$$\int_0^1 \langle \partial_t \varphi(t,\cdot), \rho(t,\cdot)\rangle_\pi + \langle \nabla_{\mathcal{X}} \varphi(t,\cdot), m(t,\cdot)\rangle_Q \,\mathrm{d}t = \langle \varphi(1,\cdot), \rho_B\rangle_\pi - \langle \varphi(0,\cdot), \rho_A\rangle_\pi \tag{5.3}$$

for all $\varphi \in C^1([0,1], \mathbb{R}^{\mathcal{X}})$.

To define the kinetic energy in terms of a mass density on nodes and a momentum on edges, we introduce an appropriate averaging operator mapping the mass of two neighboring nodes to the common edge.

Definition 5.1.2 (Averaging Operator for Mass on Edges)**.** For an averaging function $\theta\colon (\mathbb{R}_{\geqslant 0})^2 \to \mathbb{R}_{\geqslant 0}$ we require that

1. θ is continuous, concave, 1-homogeneous, and symmetric,
2. $\theta \in C^\infty\left((\mathbb{R}_+)^2, \mathbb{R}_{\geqslant 0}\right)$ and $\theta(s,t) > 0$ if $(s,t) \in \mathbb{R}^2_+$,
3. $\theta(0,s) = \theta(s,0) = 0$ and $\theta(s,s) = s$ for $s \in \mathbb{R}_{\geqslant 0}$, and
4. $s \mapsto \theta(t,s)$ is monotone increasing on $\mathbb{R}_{\geqslant 0}$ for fixed $t \in \mathbb{R}_{\geqslant 0}$.

Note that we can extend θ to a concave function $\theta\colon \mathbb{R}^2 \to \mathbb{R} \cup \{-\infty\}$ by setting $\theta(s,t) = -\infty$ for $(s,t) \notin (\mathbb{R}_{\geqslant 0})^2$.

Example 5.1.3 (Possible Averaging Operators)**.** Possible choices for θ are, *e.g.*, the logarithmic mean $\theta_{\log}$ or the geometric mean θ_{geo} for s, $t \in \mathbb{R}_{\geqslant 0}$:

$$\theta_{\log}(s,t) = \begin{cases} 0, & \text{if } s = 0 \text{ or } t = 0\,, \\ s, & \text{if } s = t\,, \\ \dfrac{t-s}{\log(t) - \log(s)} & \text{otherwise,} \end{cases} \qquad \text{and} \qquad \theta_{\text{geo}}(s,t) = \sqrt{st}\,. \tag{5.4}$$

However, the arithmetic mean is not admissible, since $\theta_{\text{arith}}(s,0) \neq 0$ for $s > 0$.

Based on this averaging function, we can define the discrete optimal transport distance on $\mathscr{P}(\mathcal{X})$.

Definition 5.1.4 (Discrete Optimal Transport Distance)**.** The kinetic energy functional for measurable functions $\rho\colon [0,1] \times \mathbb{R}^{\mathcal{X}} \to \mathbb{R}_+$ and $m\colon [0,1] \times \mathbb{R}^{\mathcal{X}\times\mathcal{X}} \to \mathbb{R}$ is defined as

$$\mathcal{E}_{\text{trans}}(\rho, m) = \frac{1}{2}\int_0^1 \sum_{x,y\in\mathcal{X}} \Phi_{\mathrm{e}}\big(\rho(t,x), \rho(t,y), m(t,x,y)\big) Q(x,y)\pi(x) \,\mathrm{d}t$$

with $\Phi_{\mathrm{e}}\colon \mathbb{R}^3 \to \mathbb{R} \cup \{\infty\}$ given by

$$\Phi_{\mathrm{e}}(s,t,m) = \begin{cases} \dfrac{m^2}{\theta(s,t)} & \text{if } \theta(s,t) > 0\,, \\ 0 & \text{if } \theta(s,t) = 0 \text{ and } m = 0\,, \\ \infty & \text{else}\,. \end{cases}$$

The total path energy is then given by a sum of the kinetic energy and the indicator function of the set $\mathcal{CE}(\rho_A, \rho_B)$, *i.e.*,

$$\mathcal{E}(\rho, m) = \mathcal{E}_{\text{trans}}(\rho, m) + \mathcal{I}_{\mathcal{CE}(\rho_A,\rho_B)}(\rho, m)\,, \tag{5.5}$$

and the induced discrete transport distance is obtained by

$$\mathcal{W}_G(\rho_A,\rho_B) := \left(\inf_{(\rho,m)\in\mathcal{CE}(\rho_A,\rho_B)} \mathcal{E}_{\mathrm{trans}}(\rho,m) \right)^{\frac{1}{2}} = \left(\inf_{(\rho,m)\colon [0,1]\to\mathbb{R}^{\mathcal{X}}\times\mathbb{R}^{\mathcal{X}\times\mathcal{X}} \text{ measurable}} \mathcal{E}(\rho,m) \right)^{\frac{1}{2}} . \tag{5.6}$$

Under the condition that

$$C_\theta := \int_0^1 \frac{1}{\sqrt{\theta(1-r,1+r)}}\,\mathrm{d}r < \infty ,$$

in [Maa11, Theorem 3.8], it was verified that $\mathcal{W}_G$ defines a metric on $\mathscr{P}(\mathcal{X})$. Due to our assumptions on θ in Definition 5.1.2, we can always guarantee that $C_\theta < \infty$. Indeed, since $\theta(s,s) = s$ for $s \in \mathbb{R}_{\geqslant 0}$ and $s \mapsto \theta(s,t)$ is increasing on $\mathbb{R}_{\geqslant 0}$ for fixed $t \in \mathbb{R}_{\geqslant 0}$, it follows that $\theta(s,t) \geqslant \min\{s,t\}$ for $s, t \in \mathbb{R}_{\geqslant 0}$. Furthermore, in [EM12, Theorem 3.2], it was shown that the infimum in (5.6) is attained by an optimal pair (ρ,m), and the curve $(\rho_t)_{t\in[0,1]}$ is a constant speed geodesic for the distance $\mathcal{W}_G$, *i.e.*, it holds $\mathcal{W}_G(\rho_t,\rho_s) = |t-s|\mathcal{W}_G(\rho_A,\rho_B)$ for all $s,t \in [0,1]$. Finally, note that Φ_e is a convex and lower semi-continuous function and thus, finding an optimal transport geodesic minimizing (5.6) is a convex optimization problem.

Related Questions to the Discrete Optimal Transport Distance. The discrete optimal transport distance has been intensely investigated during the last few years, and certain properties have been proven. In [Maa11], it was shown that the optimal transport geodesic for a graph with two nodes does not coincide with the linear interpolation of the mass variable. Already for a graph with three nodes, the solution is unknown. We observe that the discrete optimal transport distance behaves effectively diffuse. Indeed, it turns out that on a complete graph with three nodes for an optimal transport geodesic between two nodes, the mass is not necessarily transported along the shortest path connecting these nodes. Instead, a small amount of mass is transported via the third node (see Figure 5.1), which is in sharp contrast to the displacement interpolation on continuous domains, where mass travels only along geodesics. Furthermore, we obtain that the momentum variable does not necessarily have the same sign during the transport, which also reflects the diffuse behavior of the discrete optimal transport distance (see Figure 5.1). We also refer to Section 5.6.2 for more numerical results on simple graphs.

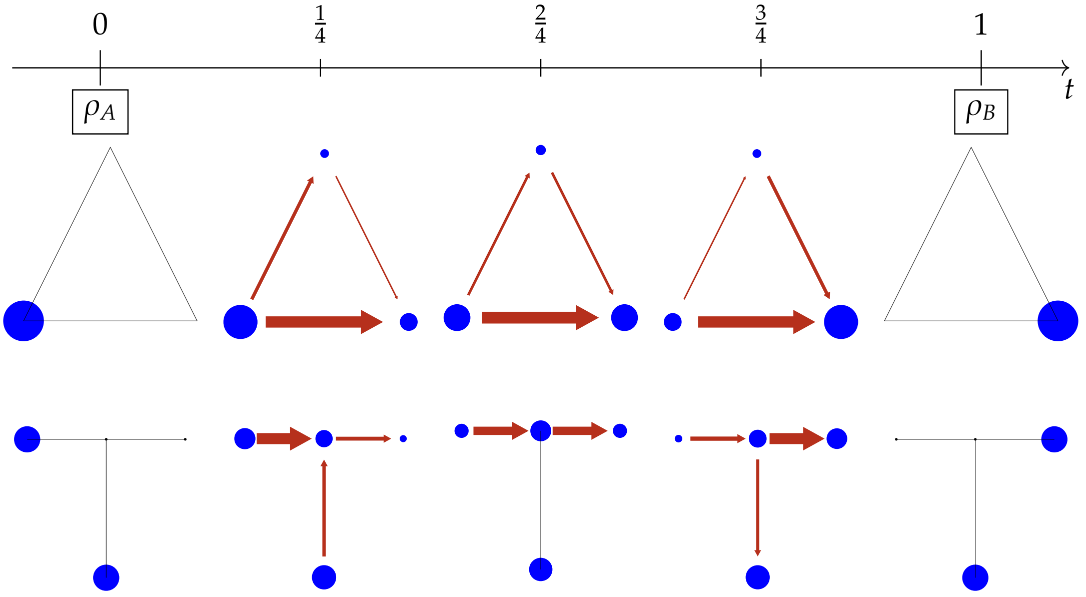


Figure 5.1: Two examples of optimal transport geodesic on graphs. Top: On a graph with three nodes, a small amount of mass (in blue) is transported along the longer way. Bottom: On a graph with four nodes, the momentum variable (depicted by red arrows) changes its sign during the transport.

Considering a sequence of graphs with appropriate Markov kernels approximating a domain in the Euclidean space, in [GM13, GKM18], it was shown that the distance $\mathcal{W}_G$ on a specific class of regular meshes converges in the Gromov–Hausdorff metric to the classical L^2-Wasserstein distance. We verify this convergence experimentally in Section 5.6.4. Moreover, in [Maa11, EM14], certain solutions to partial differential equations on graphs were identified as gradient flow trajectories w.r.t. suitable entropy functionals. In Section 5.7, we discuss solutions to the heat and porous medium equation on graphs.

5.2 A priori Bounds for Mass and Momentum

Now, we derive a priori bounds on energy minimizing curves of measures, which are useful for the Γ-convergence result in Section 5.4. Here, we essentially make use of the discrete structure in space. More precisely, the divergence operator on the graph does not reduce the regularity of the momentum variable and thus allows a higher regularity estimate on the mass variable, which does not hold for the classical optimal transport distance.

Lemma 5.2.1 (A priori Bounds for Mass and Momentum)**.** *Let $(\rho, m)\colon [0,1] \to \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ be measurable with bounded path energy,* i.e., *there is a constant $\bar{E} < \infty$ s.t. $\mathcal{E}(\rho, m) \leqslant \bar{E}$. Then, m and ρ are bounded in $L^2((0,1), \mathbb{R}^{\mathcal{X}\times\mathcal{X}})$ and $W^{1,2}((0,1), \mathbb{R}^{\mathcal{X}}) \cap C^{0,\frac{1}{2}}([0,1], \mathbb{R}^{\mathcal{X}})$, respectively, with bounds solely depending on $\mathcal{X}$ and $\bar{E}$.*

Proof. Since $\mathcal{E}(\rho, m) < \infty$, we have that $(\rho, m) \in \mathcal{CE}(\rho_A, \rho_B)$, and thus, for a.e. $t \in (0,1)$ the mass is preserved, *i.e.*,

$$\sum_{x\in\mathcal{X}} \rho(t,x)\pi(x) = \sum_{x\in\mathcal{X}} \rho_A(x)\pi(x) = 1\,.$$

In addition, $\rho(t,x)$ is nonnegative for all $x \in \mathcal{X}$ and a.e. $t \in (0,1)$. By symmetry and concavity of θ and since $\theta(s,s) = s$, we can estimate

$$\begin{aligned}\theta(\rho(t,x), \rho(t,y)) &= \frac{1}{2}\theta(\rho(t,x), \rho(t,y)) + \frac{1}{2}\theta(\rho(t,y), \rho(t,x))\\ &\leqslant \theta\left(\frac{\rho(t,x)+\rho(t,y)}{2}, \frac{\rho(t,x)+\rho(t,y)}{2}\right) = \frac{\rho(t,x)+\rho(t,y)}{2}\end{aligned}$$

and get

$$\begin{aligned}&\sum_{x,y\in\mathcal{X}} \theta(\rho(t,x),\rho(t,y))Q(x,y)\pi(x) \leqslant \frac{1}{2}\sum_{x,y\in\mathcal{X}} (\rho(t,x)Q(x,y)\pi(x) + \rho(t,y)Q(y,x)\pi(y))\\ &= \frac{1}{2}\sum_{x,y\in\mathcal{X}} (\rho(t,x)Q(y,x)\pi(y) + \rho(t,y)Q(x,y)\pi(x)) \leqslant C^* \sum_{x\in\mathcal{X}} \rho(t,x)\pi(x) = C^*\,.\end{aligned}$$

Thus, using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\Big(\sum_{x,y\in\mathcal{X}} |m(t,x,y)|Q(x,y)\pi(x)\Big)^2 \leqslant &\Big(\sum_{x,y\in\mathcal{X}} \Phi_e(\rho(t,x),\rho(t,y),m(t,x,y))Q(x,y)\pi(x)\Big)\\ &\cdot\Big(\sum_{x,y\in\mathcal{X}} \theta(\rho(t,x),\rho(t,y))Q(x,y)\pi(x)\Big)\,.\end{aligned}$$

Integrating in time leads to

$$\int_0^1 \|m(t,\cdot,\cdot)\|_Q^2 \,\mathrm{d}t = \int_0^1 \sum_{x,y\in\mathcal{X}} m(t,x,y)^2 Q(x,y)\pi(x)\,\mathrm{d}t \leqslant \frac{C^*}{C_*}\,\bar{E}\,.$$

Finally, using the continuity equation (5.3) and from above that $m \in L^2((0,1), \mathbb{R}^{\mathcal{X}\times\mathcal{X}})$, we obtain that

$$\int_0^1 \|\partial_t\rho\|_\pi^2 \,\mathrm{d}t \leqslant \int_0^1 \sum_{x\in\mathcal{X}} \Big|\sum_{y\in\mathcal{X}} m(t,x,y)Q(x,y)\Big|^2 \pi(x)\,\mathrm{d}t \leqslant C^* \int_0^1 \sum_{x,y\in\mathcal{X}} m(t,x,y)^2 Q(x,y)\pi(x)\,\mathrm{d}t\,.$$

Hence, $\rho \in W^{1,2}((0,1), \mathbb{R}^{\mathcal{X}})$ and the Sobolev embedding (see Theorem 2.2.2) implies $\rho \in C^{0,\frac{1}{2}}((0,1), \mathbb{R}^{\mathcal{X}})$. □

Furthermore, we note that a priori for a fixed time $t \in [0,1]$ and an edge $(x,y) \in \mathcal{S}$, the momentum variable $m(t,x,y)$ is not related to the variable $m(t,y,x)$ in the opposite direction, but for an optimizing path, we observe the following antisymmetry.

Lemma 5.2.2 (Antisymmetry of Optimal Momentum)**.** *Let* $\rho\colon [0,1] \to \mathbb{R}^{\mathcal{X}}$ *and* $m\colon [0,1] \to \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ *be an optimal path for* (5.6) *with* $\mathcal{E}(\rho,m) < \infty$*. Then for a.e.* $t \in [0,1]$ *and for all* $(x,y) \in \mathcal{S}$ *we have*

$$m(t,x,y) = -m(t,y,x)\,.$$

Proof. We define a momentum variable

$$\hat{m}(t,x,y) := -m(t,y,x)\,.$$

Then, we can verify that $\operatorname{div}_{\mathcal{X}}\hat{m} = \operatorname{div}_{\mathcal{X}} m$ and thus $(\rho,\hat{m}) \in \mathcal{CE}(\rho_A,\rho_B)$ as well. Because of the detailed balance condition $Q(x,y)\pi(x) = Q(y,x)\,\pi(y)$ and since $\Phi_{\mathrm{e}}(s,t,m) = \Phi_{\mathrm{e}}(t,s,-m)$, we find that $\mathcal{E}_{\mathrm{trans}}(\rho,\hat{m}) = \mathcal{E}_{\mathrm{trans}}(\rho,m)$. Now, we define a momentum variable

$$\bar{m}(t,x,y) := \frac{1}{2}\left(m(t,x,y) + \hat{m}(t,x,y)\right)\,,$$

which is antisymmetric in x and y. By convexity of $\mathcal{CE}(\rho_A,\rho_B)$ we also have that $(\rho,\bar{m}) \in \mathcal{CE}(\rho_A,\rho_B)$. Moreover, by convexity of $\mathcal{E}_{\mathrm{trans}}$ we get

$$\mathcal{E}_{\mathrm{trans}}(\rho,\bar{m}) \leqslant \frac{1}{2}\left(\mathcal{E}_{\mathrm{trans}}(\rho,m) + \mathcal{E}_{\mathrm{trans}}(\rho,\hat{m})\right) = \mathcal{E}_{\mathrm{trans}}(\rho,m)\,.$$

By definition, values of $m(t,x,y)$ for $(x,y) \notin \mathcal{S}$ have no impact on the kinetic energy $\mathcal{E}_{\mathrm{trans}}$. Assume that $\theta(\rho(t,x),\rho(t,y)) = 0$ for a.e. $t \in [0,1]$ and $(x,y) \in \mathcal{S}$. Since $\mathcal{E}_{\mathrm{trans}}(\rho,m) < \infty$ this would imply $m(t,x,y) = 0$ for a.e. $t \in [0,1]$ and $(x,y) \in \mathcal{S}$. Now, the function $z \mapsto \Phi_{\mathrm{e}}(s,t,z)$ for $z \in \mathbb{R}$ is even strictly convex for fixed $(s,t) \in (\mathbb{R}_+)^2$. Hence, we observe that $\mathcal{E}_{\mathrm{trans}}(\rho,\bar{m}) < \mathcal{E}_{\mathrm{trans}}(\rho,m)$ unless $\bar{m}$ already coincides with m for a.e. $t \in [0,1]$ and all $(x,y) \in \mathcal{S}$. □

Remark 5.2.3 (Bounded Energy of Optimal Path). In Corollary 5.4.2, we verify that an optimal path (ρ,m) for (5.6) always fulfills $\mathcal{E}(\rho,m) < \infty$.

5.3 Finite Element Discretization

In the following, we provide a fully numerical discretization of the path energy (5.5). Because the domain $\mathcal{X}$ is already discrete, we only need to define a discretization in time. Here, we choose a Galerkin discretization by dividing the time interval $[0,1]$ into N subintervals $I_i = [t_i,t_{i+1})$ for $i = 0,\dots,N-1$ with a uniform step size $h = \frac{1}{N}$ and $t_i = ih$. Then we define the finite element spaces

$$\begin{aligned}
V^1_{n,h} &= \{\psi_h \in C^0([0,1],\mathbb{R}^{\mathcal{X}}) \ : \ \psi_h(\cdot)|_{I_i} \text{ is affine } \forall i = 0,\dots,N-1\}\,,\\
V^0_{n,h} &= \{\psi_h\colon [0,1] \to \mathbb{R}^{\mathcal{X}} \ : \ \psi_h(\cdot)|_{I_i} \text{ is constant } \forall i = 0,\dots,N-1\}\,,\\
V^0_{e,h} &= \{\psi_h\colon [0,1] \to \mathbb{R}^{\mathcal{X}\times\mathcal{X}} \ : \ \psi_h(\cdot)|_{I_i} \text{ is constant } \forall i = 0,\dots,N-1\}\,.
\end{aligned}$$

Note that for a function $\psi_h \in V^1_{n,h}$ the time-derivative can be interpreted as a map

$$\partial_t\colon V^1_{n,h} \to V^0_{n,h}\,, \qquad (\partial_t\psi_h)(t_i) = \frac{1}{h}(\psi_h(t_{i+1}) - \psi_h(t_i)) \text{ for } i = 0,\dots,N-1\,.$$

Since a function $\psi_h \in V^0_{n,h}$ or $V^0_{e,h}$ is constant on time intervals $I_i = [t_i,t_{i+1})$, we often write $\psi_h(t_i)$ to refer to its value on the interval.

Now, we choose the discretized mass variable $\rho_h \in V^1_{n,h}$ in the space of continuous and piecewise affine functions and the momentum variable $m_h \in V^0_{e,h}$ as piecewise constant. Then, discrete solutions to the continuity equation are defined in analogy to Definition 5.1.1.

Definition 5.3.1 (Time Discrete Continuity Equation)**.** The set of solutions to the discretized continuity equation for given boundary values $\rho_A, \rho_B \in \mathbb{R}^{\mathcal{X}}$ is given by

$$\mathcal{CE}_h(\rho_A, \rho_B) = \Big\{(\rho_h, m_h) \in V^1_{n,h} \times V^0_{e,h} \ : h \sum_{i=0}^{N-1} \langle \partial_t \rho_h(t_i, \cdot) + \mathrm{div}_{\mathcal{X}} m_h(t_i, \cdot), \varphi_h(t_i, \cdot)\rangle_\pi = 0 \ \forall \varphi_h \in V^0_{n,h} \,,$$
$$\rho_h(t_0, x) = \rho_A(x) \,, \rho_h(t_N, x) = \rho_B(x)\Big\} \,.$$

Here, the choice of these different function spaces for ρ_h and m_h is motivated by the two expressions $\partial_t \rho_h$ and $\mathrm{div}_{\mathcal{X}} m_h$ appearing in the continuity equation, which then both lie in the space $V^0_{n,h}$. Thus, for $(\rho_h, m_h) \in \mathcal{CE}_h(\rho_A, \rho_B)$, we have that $\partial_t \rho_h + \mathrm{div}_{\mathcal{X}} m_h = 0$ for a.e. $t \in [0, 1]$. Consequently, the set of time discrete solutions $\mathcal{CE}_h(\rho_A, \rho_B) = \mathcal{CE}(\rho_A, \rho_B) \cap (V^1_{n,h} \times V^0_{e,h})$ is a subset of all time continuous solutions to the continuity equation.

Furthermore, we define a fully discrete path energy functional in analogy to Definition 5.6 and a discrete version of the transport metric $\mathcal{W}_G$.

Definition 5.3.2 (Time Discrete Optimal Transport Distance)**.** The averaging operator avg_h takes a vectorial Radon measure $\psi \in \mathscr{M}([0, 1], \mathbb{R}^{\mathcal{X}})$ to its average values on time intervals I_i, *i.e.*, it is given by

$$\mathrm{avg}_h \colon \mathscr{M}([0, 1], \mathbb{R}^{\mathcal{X}}) \to V^0_{n,h} \,, \qquad (\mathrm{avg}_h \, \psi)(t_i) = \psi(I_i) \ \text{ for } i = 0, \ldots, N-1 \,.$$

Analogously, we declare the averaging operator avg_h for $\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$-valued measures. Note that for $\psi_h \in V^1_{n,h}$ we find $(\mathrm{avg}_h \, \psi_h)(t_i) = \frac{1}{2}(\psi_h(t_i) + \psi_h(t_{i+1}))$. For $(\rho, m) \in \mathscr{M}([0, 1], \mathbb{R}^{\mathcal{X}}) \times \mathscr{M}([0, 1], \mathbb{R}^{\mathcal{X}\times\mathcal{X}})$ the discrete approximation for the kinetic energy functional is given by

$$\begin{aligned}\mathcal{E}_{\mathrm{trans},h}(\rho, m) &= \mathcal{E}_{\mathrm{trans}}(\mathrm{avg}_h \, \rho, \mathrm{avg}_h \, m) \\ &= \frac{h}{2} \sum_{i=0}^{N-1} \sum_{x,y \in \mathcal{X}} \Phi_{\mathrm{e}} \left(\mathrm{avg}_h \, \rho(t_i, x), \mathrm{avg}_h \, \rho(t_i, y), \mathrm{avg}_h \, m(t_i, x, y)\right) Q(x, y) \pi(x) \,.\end{aligned}$$

Finally, we obtain the time discrete energy functional by

$$\mathcal{E}_h(\rho, m) = \mathcal{E}_{\mathrm{trans},h}(\rho, m) + \mathcal{I}_{\mathcal{CE}_h(\rho_A,\rho_B)}(\rho, m) \,,$$

and for the associated time discrete approximation of the optimal transport distance we define

$$\mathcal{W}_{G,h}(\rho_A, \rho_B) := \left(\inf_{(\rho,m)\in\mathcal{CE}_h(\rho_A,\rho_B)} \mathcal{E}_{\mathrm{trans},h}(\rho, m)\right)^{\frac{1}{2}} = \left(\inf_{(\rho,m)\colon [0,1]\to\mathbb{R}^{\mathcal{X}}\times\mathbb{R}^{\mathcal{X}\times\mathcal{X}} \text{ measurable}} \mathcal{E}_h(\rho, m)\right)^{\frac{1}{2}} . \tag{5.7}$$

Finally, we remark that the degrees of freedom of the momentum variable $m_h \in V^0_{e,h}$ are restricted to the edges $(x, y) \in \mathcal{S}$. Thus, in the implementation, if the Markov kernel Q is sparse, *i.e.*, if $\mathcal{S}$ is only a small subset of $\mathcal{X} \times \mathcal{X}$, this implies a considerable reduction of computational complexity.

5.4 Γ-Convergence of Finite Element Discretization

Now, we show that the fully discrete distance $\mathcal{W}_{G,h}$ in (5.7) is a suitable approximation of $\mathcal{W}_G$ by proving a Γ-convergence result. For a basic introduction to Γ-convergence, we refer the reader to Section 2.3. We observe that the Γ-liminf inequality is a direct consequence of our conforming finite element discretization in the sense that $\mathcal{CE}_h(\rho_A, \rho_B) \subset \mathcal{CE}(\rho_A, \rho_B)$. However, the proof of the Γ-limsup inequality is more elaborated. Therefore, we first sketch the main ideas:

1. Basically, the recovery sequence is constructed by averages on time intervals according to the Galerkin discretization.
2. For positive mass $\rho(t, x) > 0$ for all $t \in [0, 1]$ and for all $x \in \mathcal{X}$, Jensen's inequality would directly provide the required Γ-limsup inequality, since $\Phi_{\mathrm{e}} \colon \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}_+$ is convex.

3. However, in the case that $\theta(\rho(t,x),\rho(t,y)) = 0$ and $m(t,x,y) \neq 0$, Jensen's inequality cannot be applied, since $\Phi_e(\rho(t,x),\rho(t,y),m(t,x,y)) = \infty$.

4. We first show in Proposition 5.4.1 that we can construct a trajectory between an arbitrary probability distribution on $\mathcal{X}$ and the uniform probability density $\mathbb{1} \in \mathscr{P}(\mathcal{X})$ given by $\mathbb{1}(x) = 1$ for all $x \in \mathcal{X}$.

5. To deal with the case that $\rho(t,x) = 0$ for some (t,x), we make use of this trajectory. More precisely, we modify the trajectory to be recovered by letting a small amount of mass ε move on a small time interval δ to the uniform distribution $\mathbb{1}$ (see Figure 5.2 for a sketch).

6. Finally, we have to choose ε, δ, and h in a suitable way.

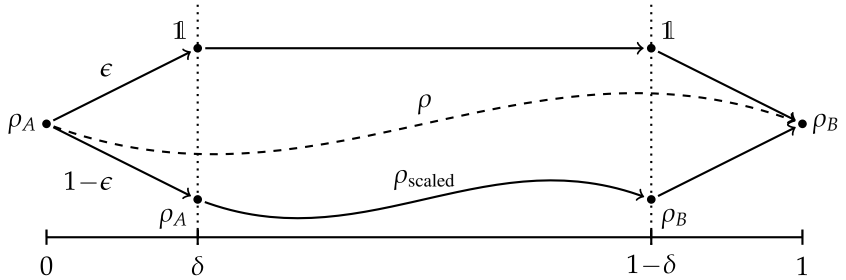


Figure 5.2: Sketch of the construction of the recovery sequence. A given ρ (dashed line) is regularized to a curve consisting of a weighted sum of two curves (with weights ϵ and $1-\epsilon$).

First, we explicitly construct a trajectory between ρ_A and $\mathbb{1}$ with uniformly bounded path energy. Furthermore, the time interpolation of this trajectory admits the same upper bound for the corresponding discrete path energy, where for the approximation of the mass variable, we define the Lagrange interpolation operator $\mathcal{I}_h \colon C^0([0,1],\mathbb{R}^{\mathcal{X}}) \to V^1_{n,h}$ by

$$(\mathcal{I}_h\rho)\,(t_i,x) := \rho(t_i,x) \quad \forall i = 0,\ldots,N\,.$$

Proposition 5.4.1 (Trajectory to Uniform Distribution)**.** *There exists a constant $C(\mathcal{X}) < \infty$ s.t. for any $\rho_A \in \mathscr{P}(\mathcal{X})$ there is a trajectory $(\rho,m) \in \mathcal{CE}(\rho_A,\mathbb{1})$ with $\mathcal{E}_{trans}(\rho,m) \leqslant C(\mathcal{X})$ and $(\mathcal{I}_h\rho,\operatorname{avg}_h m) \in \mathcal{CE}_h(\rho_A,\mathbb{1})$ with $\mathcal{E}_{trans,h}(\mathcal{I}_h\rho,\operatorname{avg}_h m)) \leqslant C(\mathcal{X})$ for every $h = 1/N$.*

Proof. For $x \in \mathcal{X}$ we define $\rho_A^x \in \mathscr{P}(\mathcal{X})$ to be the probability density on $\mathcal{X}$ with all mass concentrated on x, *i.e.*, $\rho_A^x = \frac{1}{\pi(x)}\delta_x$, where δ_x denotes the Kronecker symbol with $\delta_x(y) = 1$ if $x = y$ and 0 else.

Step 1: Construction of elementary flows. For $(x,y) \in \mathcal{X}\times\mathcal{X}$, $x \neq y$, with $Q(x,y) > 0$ we define $L(x,y) \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ as

$$L(x,y)(a,b) = \begin{cases} \dfrac{1}{Q(x,y)\pi(x)} & \text{if } (a,b) = (x,y), \\ \dfrac{-1}{Q(x,y)\pi(x)} & \text{if } (a,b) = (y,x), \\ 0 & \text{else.} \end{cases}$$

Then $L(x,y)$ satisfies $\operatorname{div}_{\mathcal{X}} L(x,y) = \rho_A^y - \rho_A^x$. Now, for any $(x,y) \in \mathcal{X}\times\mathcal{X}$, $x \neq y$, there exists a path $(x = x_0, x_1, \ldots, x_K = y)$ with $K < |\mathcal{X}|$ and $Q(x_k,x_{k+1}) > 0$ for $k = 0,\ldots,K-1$. We can add the corresponding $L(x_k,x_{k+1})$ along these edges to construct a flow $M(x,y)$ with $\operatorname{div}_{\mathcal{X}} M(x,y) = \rho_A^y - \rho_A^x$. All entries of all $M(x,y)$ are bounded by $\widetilde{C}(\mathcal{X}) := \frac{|\mathcal{X}|}{C_*}$, where C_* is defined in (5.2). For $x = y$ we can simply set $M(x,x) = 0$.

Now assume $\rho_A = \rho_A^x$ for some $x \in \mathcal{X}$. Let $m_0 = \sum_{y\in\mathcal{X}} M(x,y)\,\pi(y)$. We obtain

$$\operatorname{div}_{\mathcal{X}} m_0 = \sum_{y\in\mathcal{X}} \left(\tfrac{1}{\pi(y)}\delta_y - \tfrac{1}{\pi(x)}\delta_x\right)\pi(y) = \mathbb{1} - \rho_A^x\,.$$

Again, every entry of m_0 is bounded in absolute value by $\tilde{C}(\mathcal{X})$. Now, let $m(t) = 2\,m_0\,t$, and $\rho(t) = \rho_A^x + (\operatorname{div}_{\mathcal{X}} m_0)\,t^2 = (1-t^2)\cdot\rho_A^x + t^2\cdot\mathbb{1}$. Then $(\rho,m)\in\mathcal{CE}(\rho_A^x,\mathbb{1})$ is bounded by $|m(t,x,y)|\leqslant t\cdot 2\tilde{C}(\mathcal{X})$ and $\rho(t,x)\geqslant t^2$. By the monotonicity of Φ_{e}, we get for the kinetic energy

$$\mathcal{E}_{\mathrm{trans}}(\rho,m)\leqslant\frac{1}{2}\int_0^1\sum_{x,y\in\mathcal{X}}\frac{(t\cdot 2\tilde{C}(\mathcal{X}))^2}{t^2}Q(x,y)\pi(x)\,\mathrm{d}t = 2\tilde{C}(\mathcal{X})^2C^*\,.$$

Step 2: Construction of discrete counterparts. For fixed $h=\frac{1}{N}$ let $\rho_h=\mathcal{I}_h\rho$ and $m_h=\operatorname{avg}_h m$. By construction $(\rho_h,m_h)\in\mathcal{CE}_h(\rho_A^x,\mathbb{1})$. Then, we find $m_h(t_i,x,y)\leqslant(i+\frac{1}{2})\,h\,2\tilde{C}(\mathcal{X})$, $\rho_h(t_i,x)\geqslant i^2h^2$, $(\operatorname{avg}_h\rho_h)(t_i,x)\geqslant(i^2+i+\frac{1}{2})\,h^2$, and thus

$$\mathcal{E}_{\mathrm{trans},h}(\rho_h,m_h)=\mathcal{E}_{\mathrm{trans}}(\operatorname{avg}_h\rho_h,m_h)\leqslant\frac{1}{2}\sum_{i=0}^{N-1}h\frac{h^2\,4\tilde{C}(\mathcal{X})^2\,(i+\frac{1}{2})^2}{h^2(i^2+i+\frac{1}{2})}\sum_{x,y\in\mathcal{X}}Q(x,y)\pi(x)\leqslant 2\tilde{C}(\mathcal{X})^2C^*\,.$$

Step 3: Extension to arbitrary initial data. For given $x\in\mathcal{X}$ let (ρ^x,m^x) be the (continuous) trajectory between ρ_A^x and $\mathbb{1}$ as constructed above. Now, we can represent ρ_A as superposition of various ρ_A^x by

$$\rho_A=\sum_{x\in\mathcal{X}}\rho_A(x)\,\delta_x=\sum_{x\in\mathcal{X}}\rho_A(x)\pi(x)\,\rho_A^x\,.$$

By linearity of the continuity equation, the trajectory $(\rho,m)=\sum_{x\in\mathcal{X}}\rho_A(x)\,\pi(x)\cdot(\rho^x,m^x)$ then lies in $\mathcal{CE}(\rho_A,\mathbb{1})$. Since $\mathcal{E}_{\mathrm{trans}}$ is convex and 1-homogeneous, it is subadditive. Therefore, we can estimate the kinetic energy by

$$\mathcal{E}_{\mathrm{trans}}(\rho,m)\leqslant\sum_{x\in\mathcal{X}}\rho_A(x)\,\pi(x)\,\mathcal{E}_{\mathrm{trans}}(\rho^x,m^x)\leqslant\sum_{x\in\mathcal{X}}\rho_A(x)\,\pi(x)\,2\,\tilde{C}(\mathcal{X})^2C^*=2\,\tilde{C}(\mathcal{X})^2C^*\,.$$

In analogy, the same estimate holds for the discrete trajectory. Thus, the claim follows with $C(\mathcal{X})=2\,\tilde{C}(\mathcal{X})^2C^*$. □

Corollary 5.4.2 (Uniform Bound of Discrete Optimal Transport Distance). *Let $\rho_A,\rho_B\in\mathscr{P}(\mathcal{X})$ be fixed temporal boundary conditions. Then $\mathcal{W}_G$ and $\mathcal{W}_{G,h}$ are uniformly bounded.*

Proof. Proposition 5.4.1 allows constructing trajectories between arbitrary ρ_A, ρ_B via $\mathbb{1}$ as intermediate state. □

Theorem 5.4.3 (Γ-Convergence of Time Discrete Energies). *Let $\rho_A,\rho_B\in\mathscr{P}(\mathcal{X})$ be fixed temporal boundary conditions. Then, the sequence of functionals $(\mathcal{E}_h)_h$ Γ-converges for $h\to 0$ to the functional $\mathcal{E}$ with respect to the weak-* topology in $\mathscr{M}([0,1],\mathbb{R}^{\mathcal{X}}\times\mathbb{R}^{\mathcal{X}\times\mathcal{X}})$.*

Proof. We have to prove the Γ-liminf inequality and Γ-limsup inequality (see Definition 2.3.1).

Part I: Γ-liminf inequality.
For the Γ-liminf property, we have to demonstrate that the inequality

$$\mathcal{E}_{\mathrm{trans}}(\rho,m)+\mathcal{I}_{\mathcal{CE}(\rho_A,\rho_B)}(\rho,m)\leqslant\liminf_{h\to 0}\;\mathcal{E}_{\mathrm{trans},h}(\rho_h,m_h)+\mathcal{I}_{\mathcal{CE}_h(\rho_A,\rho_B)}(\rho_h,m_h)\tag{5.8}$$

holds for all sequences $(\rho_h,m_h)\overset{*}{\rightharpoonup}(\rho,m)$ in $\mathscr{M}([0,1],\mathbb{R}^{\mathcal{X}}\times\mathbb{R}^{\mathcal{X}\times\mathcal{X}})$. The statement is trivial if there is no subsequence with $(\rho_h,m_h)\in\mathcal{CE}_h(\rho_A,\rho_B)$. Thus, we may assume that (ρ_h,m_h) fulfills the discrete continuity equation for all h, in particular, ρ is nonnegative for every h and $\mathcal{E}_{\mathrm{trans},h}(\rho_h,m_h)=\mathcal{E}_{\mathrm{trans}}(\operatorname{avg}_h\rho_h,m_h)$. Moreover, since $\mathcal{CE}(\rho_A,\rho_B)$ is weak-* closed and $\mathcal{CE}_h(\rho_A,\rho_B)\subset\mathcal{CE}(\rho_A,\rho_B)$, we also have that (ρ,m) fulfills the continuous continuity equation. Now, the convergence $\rho_h\overset{*}{\rightharpoonup}\rho$ for $h\to 0$ implies that $\operatorname{avg}_h\rho_h\overset{*}{\rightharpoonup}\rho$ for $h\to 0$. Since Φ_{e} is jointly convex and lower semi-continuous in ρ and m, the kinetic energy functional $\mathcal{E}_{\mathrm{trans}}$ is weak-$*$ lower semi-continuous and the Γ-liminf inequality (5.8) holds.

Part II: Γ-limsup inequality.
To verify the Γ-limsup property we need to show that for any $(\rho, m) \in \mathscr{M}([0,1], \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}\times\mathcal{X}})$ there exists a recovery sequence $(\rho_h, m_h) \overset{*}{\rightharpoonup} (\rho, m)$ with

$$\limsup_{h\to 0} \mathcal{E}_{\mathrm{trans},h}(\rho_h, m_h) + \mathcal{I}_{\mathcal{CE}_h(\rho_A,\rho_B)}(\rho_h, m_h) \leqslant \mathcal{E}_{\mathrm{trans}}(\rho, m) + \mathcal{I}_{\mathcal{CE}(\rho_A,\rho_B)}(\rho, m)\,.$$

First, we can restrict to the case $\mathcal{E}_{\mathrm{trans}}(\rho, m) < \infty$ and $(\rho, m) \in \mathcal{CE}(\rho_A, \rho_B)$.

Step 1: Regularizing the Continuous Trajectory (ρ, m)**.** Let $(\rho_{A,\mathbb{1}}, m_{A,\mathbb{1}}) \in \mathcal{CE}(\rho_A, \mathbb{1})$ be the trajectory from ρ_A to $\mathbb{1}$ as constructed in Proposition 5.4.1. Analogously, let $(\rho_{\mathbb{1},B}, m_{\mathbb{1},B}) \in \mathcal{CE}(\mathbb{1}, \rho_B)$ be the corresponding trajectory from $\mathbb{1}$ to ρ_B with $(\rho_{\mathbb{1},B}, m_{\mathbb{1},B})(t,\cdot) := (\rho_{B,\mathbb{1}}, -m_{B,\mathbb{1}})(1-t,\cdot)$. Then, for $\delta \in (0, \frac{1}{2})$ and $\epsilon = \delta^2$, we define (as sketched in Figure 5.2)

$$\rho_\delta(t) = \begin{cases} (1-\epsilon)\,\rho_A(t) + \epsilon\,\rho_{A,\mathbb{1}}\left(\frac{t}{\delta}\right) & \text{for } t \in [0,\delta)\,, \\ (1-\epsilon)\,\rho\left(\frac{t-\delta}{1-2\delta}\right) + \epsilon\,\mathbb{1} & \text{for } t \in [\delta, 1-\delta)\,, \\ (1-\epsilon)\,\rho_B(t) + \epsilon\,\rho_{\mathbb{1},B}\left(\frac{t-(1-\delta)}{\delta}\right) & \text{for } t \in [1-\delta, 1]\,, \end{cases}$$

and

$$m_\delta(t) = \begin{cases} \frac{\epsilon}{\delta}\, m_{A,\mathbb{1}}\left(\frac{t}{\delta}\right) & \text{for } t \in [0,\delta)\,, \\ \frac{(1-\epsilon)}{1-2\delta}\, m\left(\frac{t-\delta}{1-2\delta}\right) & \text{for } t \in [\delta, 1-\delta)\,, \\ \frac{\epsilon}{\delta}\, m_{\mathbb{1},B}\left(\frac{t-(1-\delta)}{\delta}\right) & \text{for } t \in [1-\delta, 1]\,. \end{cases}$$

We observe that $(\rho_\delta, m_\delta) \in \mathcal{CE}(\rho_A, \rho_B)$. To evaluate the kinetic energy of (ρ_δ, m_δ), we define the kinetic energy in space $\mathcal{D}_{\mathrm{trans}} \colon \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}\times\mathcal{X}} \to \mathbb{R} \cup \{\infty\}$ by

$$\mathcal{D}_{\mathrm{trans}}(\rho, m) = \frac{1}{2} \sum_{x,y\in\mathcal{X}} \Phi_{\mathrm{e}}(\rho(x), \rho(y), m(x,y)) Q(x,y)\pi(x)\,.$$

Furthermore, we decompose the energy into the contributions on the time intervals $I_l = [0,\delta]$, $I_m = [\delta, 1-\delta]$ and $I_r = [1-\delta, 1]$. More precisely, for $\chi \in \{l, m, r\}$, we define the time-space kinetic energy on the specific interval as

$$\mathcal{E}^{\chi}_{\mathrm{trans}} = \int_{I_\chi} \mathcal{D}_{\mathrm{trans}}(\rho_\delta(t), m_\delta(t))\,\mathrm{d}t\,,$$

s.t. $\mathcal{E}_{\mathrm{trans}}(\rho_\delta, m_\delta) = \mathcal{E}^{l}_{\mathrm{trans}} + \mathcal{E}^{m}_{\mathrm{trans}} + \mathcal{E}^{r}_{\mathrm{trans}}$. Now, $\mathcal{D}_{\mathrm{trans}}$ is jointly convex and 1-homogeneous and therefore subadditive. Moreover, it is 2-homogeneous in the second argument. Therefore, we obtain

$$\begin{aligned} \mathcal{E}^{m}_{\mathrm{trans}} &\leqslant \frac{1-\epsilon}{(1-2\delta)^2} \int_{I_m} \mathcal{D}_{\mathrm{trans}}\left(\rho\left(\frac{t-\delta}{1-2\delta}\right), m\left(\frac{t-\delta}{1-2\delta}\right)\right)\,\mathrm{d}t \\ &= \frac{1-\epsilon}{(1-2\delta)} \int_0^1 \mathcal{D}_{\mathrm{trans}}(\rho(t), m(t))\,\mathrm{d}t = \frac{1-\epsilon}{(1-2\delta)} \mathcal{E}_{\mathrm{trans}}(\rho, m)\,. \end{aligned}$$

Furthermore, using Proposition 5.4.1, we obtain $\mathcal{E}^{l}_{\mathrm{trans}} + \mathcal{E}^{r}_{\mathrm{trans}} \leqslant 2\,C(\mathcal{X})\,\delta$.

Step 2: Construction of Recovery Sequence by Local Averages of the Regularized Trajectory. Now, we construct the recovery sequence by a discretization in time of the regularized continuous trajectory. As before, we set $\epsilon = \delta^2$. First, for a fixed h, we have to choose δ appropriately. Since we can restrict to the case $(\rho, m) \in \mathcal{CE}(\rho_A, \rho_B)$, the a priori bound in Lemma 5.2.1 implies that $\rho \in C^{0,\frac{1}{2}}([0,1], \mathbb{R}^{\mathcal{X}})$. Thus, there is a function of type $g(s) := C s^{\frac{1}{2}}$ for some constant $C \in \mathbb{R}_+$ s.t. $|\rho(t,x) - \rho(t',x)| \leqslant g(|t-t'|)$ for all $x \in \mathcal{X}$. Now, we set $\Delta := g(2h)$ and define a regularization parameter

$$\delta := \min\left\{ ih \;:\; i \in \mathbb{N},\, ih \geqslant \Delta^{\frac{1}{4}} \right\}\,.$$

Then, in the limit $h \to 0$, we have convergence $\Delta \to 0$. Hence, $\delta \to 0$ and consequently $\epsilon = \delta^2 \to 0$. In particular, for h sufficiently small $2 \geqslant \frac{1}{1-2\delta}$ and thus $\Delta = g(2h) \geqslant g(\frac{h}{1-2\delta})$. Therefore, Δ is a uniform upper bound for the variation of ρ_δ on any interval of the size h. We now define the recovery sequence by

$$\rho_h := \mathcal{I}_h \rho_\delta \,, \quad m_h := \operatorname{avg}_h m_\delta \,.$$

Obviously, it holds that $(\rho_h, m_h) \in \mathcal{CE}_h(\rho_A, \rho_B)$. By construction of the recovery sequence, we have in the limit for $h \to 0$, convergence $(\rho_\delta - \rho_h, m_\delta - m_h) \overset{*}{\rightharpoonup} 0$. Furthermore, since in the limit $\delta \to 0$, we get convergence $(\rho_\delta, m_\delta) \overset{*}{\rightharpoonup} (\rho, m)$, which implies that $(\rho_h, m_h) \overset{*}{\rightharpoonup} (\rho, m)$ for $h \to 0$.

Step 3: Energy Estimate for the Recovery Sequence. Note that δ is chosen to be an integer multiple of h. Thus, the division of $[0,1]$ into the three intervals $[0,\delta]$, $[\delta, 1-\delta]$ and $[1-\delta, 1]$ in the construction of (ρ_δ, m_δ) is compatible with the grid discretization of step size h. Therefore, we can decompose the discrete kinetic energy as above into the three contributions

$$\mathcal{E}_{\text{trans},h}(\rho_h, m_h) = \mathcal{E}^l_{\text{trans},h} + \mathcal{E}^m_{\text{trans},h} + \mathcal{E}^r_{\text{trans},h} \,,$$

which we can estimate in analogy by

$$\mathcal{E}^l_{\text{trans},h} \leqslant \frac{\epsilon}{\delta} \mathcal{E}_{\text{trans},h}(\mathcal{I}_h \rho_{A,\mathbb{1}}, \operatorname{avg}_h m_{A,\mathbb{1}}), \quad \mathcal{E}^r_{\text{trans},h} \leqslant \frac{\epsilon}{\delta} \mathcal{E}_{\text{trans},h}(\mathcal{I}_h \rho_{\mathbb{1},B}, \operatorname{avg}_h m_{\mathbb{1},B}) \,,$$

and consequently by using Proposition 5.4.1 we observe that

$$\mathcal{E}^l_{\text{trans},h} + \mathcal{E}^r_{\text{trans},h} \leqslant 2\, C(\mathcal{X})\, \delta \,.$$

For the interior part we first define the set of corresponding intervals by $S_m := \{\, i \in \{0, \dots, N-1\} \,:\, I_i \subset I_m\}$, s.t. we can write the kinetic energy $\mathcal{E}^m_{\text{trans}}$ and its discrete counterpart $\mathcal{E}^m_{\text{trans},h}$ as

$$\mathcal{E}^m_{\text{trans}} = \frac{1}{2} \sum_{i \in S_m} \sum_{x,y \in \mathcal{X}} \int_{I_i} \Phi_{\text{e}}\left(\rho_\delta(t,x), \rho_\delta(t,y), m_\delta(t,x,y)\right) \,\mathrm{d}t\; Q(x,y)\pi(x) \,,$$

$$\mathcal{E}^m_{\text{trans},h} = \frac{1}{2} \sum_{i \in S_m} \sum_{x,y \in \mathcal{X}} h\, \Phi_{\text{e}}\left((\operatorname{avg}_h \mathcal{I}_h \rho_\delta)(t_i,x), (\operatorname{avg}_h \mathcal{I}_h \rho_\delta)(t_i,y), (\operatorname{avg}_h m_\delta)(t_i,x,y)\right)\; Q(x,y)\pi(x) \,.$$

Because the integrand $\Phi_{\text{e}} \colon \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}_+$ is convex, for every interval I_i with $i \in S_m$, we can apply Jensen's inequality, which gives

$$\int_{I_i} \Phi_{\text{e}}\left(\rho_\delta(t,x), \rho_\delta(t,y), m(t,x,y)\right) \,\mathrm{d}t \geqslant h\, \Phi_{\text{e}}\left((\operatorname{avg}_h \rho_\delta)(t_i,x), (\operatorname{avg}_h \rho_\delta)(t_i,y), (\operatorname{avg}_h m_\delta)(t_i,x,y)\right) \,.$$

By construction of ρ_δ and by definition of Δ, we have for any $i \in S_m$ and $z \in \mathcal{X}$ that

$$(\operatorname{avg}_h \rho_\delta)(t_i,z) \leqslant (\operatorname{avg}_h \mathcal{I}_h \rho_\delta)(t_i,z) + \Delta \,, \qquad \text{and} \qquad (\operatorname{avg}_h \mathcal{I}_h \rho_\delta)(t_i,z) \geqslant \epsilon \,.$$

Since the function $s \to \frac{s}{s+\Delta}$ is monotone, we obtain

$$\frac{(\operatorname{avg}_h \mathcal{I}_h \rho_\delta)(t_i,z)}{(\operatorname{avg}_h \rho_\delta)(t_i,z)} \geqslant \frac{(\operatorname{avg}_h \mathcal{I}_h \rho_\delta)(t_i,z)}{(\operatorname{avg}_h \mathcal{I}_h \rho_\delta)(t_i,z) + \Delta} \geqslant \frac{\epsilon}{\epsilon + \Delta}.$$

Now, by the joint 1-homogeneity of θ and the monotonicity of θ in each single component, we get for all $x, y \in \mathcal{X}$ that

$$\frac{\theta\left((\operatorname{avg}_h \mathcal{I}_h \rho_\delta)(t_i,x), (\operatorname{avg}_h \mathcal{I}_h \rho_\delta)(t_i,y)\right)}{\theta\left((\operatorname{avg}_h \rho_\delta)(t_i,x), (\operatorname{avg}_h \rho_\delta)(t_i,y)\right)} \geqslant \frac{\epsilon}{\epsilon + \Delta} = \frac{1}{1 + \frac{\Delta}{\epsilon}} \,.$$

Hence,

$$\begin{aligned}\mathcal{E}^m_{\mathrm{trans},h} &= \frac{1}{2}\sum_{i\in S_m}\sum_{x,y\in\mathcal{X}} h\,\frac{(\mathrm{avg}_h\, m_\delta)^2(t_i,x,y)}{\theta((\mathrm{avg}_h\,\mathcal{I}_h\rho_\delta)(t_i,x),(\mathrm{avg}_h\,\mathcal{I}_h\rho_\delta)(t_i,y))}Q(x,y)\pi(x)\\ &\leqslant \frac{1+\Delta}{2\epsilon}\sum_{i\in S_m}\sum_{x,y\in\mathcal{X}} h\,\frac{(\mathrm{avg}_h\, m_\delta)^2(t_i,x,y)}{\theta((\mathrm{avg}_h\,\rho_\delta)(t_i,x),(\mathrm{avg}_h\,\rho_\delta)(t_i,y))}Q(x,y)\pi(x) = \left(1+\tfrac{\Delta}{\epsilon}\right)\mathcal{E}^m_{\mathrm{trans}}\,.\end{aligned}$$

By definition, we have $\epsilon = \delta^2 \geqslant \Delta^{\frac{1}{2}}$, and thus, $\frac{\Delta}{\epsilon} \leqslant \epsilon$. Altogether, we obtain for h sufficiently small

$$\mathcal{E}_{\mathrm{trans},h}(\rho_h,m_h) = \mathcal{E}^l_{\mathrm{trans},h} + \mathcal{E}^m_{\mathrm{trans},h} + \mathcal{E}^r_{\mathrm{trans},h} \leqslant 2\,C(\mathcal{X})\,\delta + (1+\epsilon)\,\frac{1-\epsilon}{1-2\delta}\,\mathcal{E}_{\mathrm{trans}}(\rho,m)\,,$$

which converges to $\mathcal{E}_{\mathrm{trans}}(\rho,m)$ for $h\to 0$. □

Now, following Theorem 2.3.3, the Γ-convergence result provides convergence of discrete geodesics to continuous geodesics. First, we show in analogy to Lemma 5.2.1 that also the discrete momenta are uniformly bounded in L^2.

Lemma 5.4.4 (L^2-Bound for the Discrete Momentum)**.** *Let* $(\rho_h,m_h)\in V^1_{n,h}\times V^0_{e,h}$ *with finite discrete energy* $\mathcal{E}_h(\rho_h,m_h)\leqslant\bar{E}<\infty$. *Then,* m_h *is bounded in* $L^2([0,1],\mathbb{R}^{\mathcal{X}\times\mathcal{X}})$ *with a bound only depending on* $(\mathcal{X},Q,\pi)$.

Proof. The proof works in analogy to Lemma 5.2.1. First, we can estimate

$$\begin{aligned}\Big(\sum_{x,y\in\mathcal{X}}|m_h(t_i,x,y)|Q(x,y)\pi(x)\Big)^2 &\leqslant \Big(\sum_{x,y\in\mathcal{X}}\Phi_{\mathrm{e}}(\mathrm{avg}_h\,\rho_h(t_i,x),\mathrm{avg}_h\,\rho_h(t_i,y),m_h(t_i,x,y))Q(x,y)\pi(x)\Big)\\ &\quad\cdot\Big(\sum_{x,y\in\mathcal{X}}\theta(\mathrm{avg}_h\,\rho_h(t_i,x),\mathrm{avg}_h\,\rho_h(t_i,y))Q(x,y)\pi(x)\Big)\,.\end{aligned}$$

Furthermore, we have a bound

$$\sum_{x,y\in\mathcal{X}}\theta(\mathrm{avg}_h\,\rho_h(t_i,x),\mathrm{avg}_h\,\rho_h(t_i,y))Q(x,y)\pi(x)\leqslant C^*\,,$$

where C^* is defined in (5.1). Here, we have used that $(\rho_h,m_h)\in\mathcal{CE}(\rho_A,\rho_B)$, and thus, the mass is preserved, *i.e.*, $\sum_{x\in\mathcal{X}}\mathrm{avg}_h\,\rho_h(t_i,x)\pi(x) = \sum_{x\in\mathcal{X}}\rho_h(t_i+\frac{h}{2},x)\pi(x) = \sum_{x\in\mathcal{X}}\rho_A(x)\pi(x) = 1$ for all $i=0,\ldots,N-1$. Moreover, since $\mathcal{E}_{\mathrm{trans},h}(\rho_h,m_h)<\infty$, we have that $\mathrm{avg}_h\,\rho_h\geqslant 0$. Finally, using that $\mathcal{X}$ is finite and summing up in time, we establish the bound. □

Theorem 5.4.5 (Convergence of Discrete Geodesics)**.** *Let* $\rho_A,\rho_B\in\mathscr{P}(\mathcal{X})$ *be fixed temporal boundary conditions and let* (ρ_h,m_h) *be a sequence of minimizers of the discrete energy functionals* $\mathcal{E}_h$. *Then* (ρ_h,m_h) *is uniformly bounded in* $C^{0,\frac{1}{2}}([0,1],\mathbb{R}^{\mathcal{X}})\times L^2((0,1),\mathbb{R}^{\mathcal{X}\times\mathcal{X}})$, *and there exists a subsequence (here again indexed by h), s.t.* $\rho_h\to\rho$ *strongly in* $C^{0,\alpha}([0,1],\mathbb{R}^{\mathcal{X}})$ *for any* $\alpha\in[0,\frac{1}{2})$ *and* $m_h\to m$ *weakly in* L^2, *where* (ρ,m) *is a minimizer of the energy functional* $\mathcal{E}$.

Proof. For a sequence of minimizers $(\rho_h,m_h)_h$, the discrete energy $\mathcal{E}_h(\rho_h,m_h)$ is uniformly bounded by Corollary 5.4.2. Since $(\rho_h,m_h)\in\mathcal{CE}(\rho_A,\rho_B)$, the total variation of $(\rho_h)_h$ is uniformly bounded. Furthermore, by Lemma 5.4.4, the L^2-norm of $(m_h)_h$ is uniformly bounded. Hence, the sequence $(\rho_h,m_h)_h$ has a weakly-* convergent subsequence, which, by Theorem 5.4.3 and the fundamental theorem of Γ-convergence 2.3.3, converges weakly-* to a minimizer (ρ,m) of $\mathcal{E}$.

We can improve the convergence by taking into account the regularity for solutions to the continuity equation. Indeed, since m_h is uniformly bounded in $L^2((0,1),\mathbb{R}^{\mathcal{X}\times\mathcal{X}})$, the continuity equation $\partial_t\rho_h = -\mathrm{div}_{\mathcal{X}}m_h$ implies that ρ_h is uniformly bounded in $W^{1,2}((0,1),\mathbb{R}^{\mathcal{X}})$. Thus, $(\rho_h)_h$ is uniformly bounded in $C^{0,\frac{1}{2}}((0,1),\mathbb{R}^{\mathcal{X}})$ and compact in $C^{0,\alpha}((0,1),\mathbb{R}^{\mathcal{X}})$ for all $\alpha\in[0,\frac{1}{2})$ by the Sobolev embedding theorem 2.2.2. □

5.5 Optimization with Proximal Splitting

For the numerical optimization scheme, to compute a minimizer of the fully discrete energy (5.7), we apply proximal splitting methods as discussed in Section 3.2.3. We recall that for the classical L^2-Wasserstein distance, the splitting can be performed in a way s.t. the proximal mappings are obtained by solving a linear system coupled in space and time and a pointwise projection onto a convex set. Now, for the discrete transport distance, it turns out that the convex dual of the kinetic energy functional $\mathcal{E}_{\text{trans}}$ cannot be computed pointwise, since ρ and m are coupled spatially over the whole graph according to the transition kernel Q. Furthermore, because of our finite element discretization, there is a coupling in time via the averaging operator avg_h. Thus, computing the proximal operator of $\mathcal{E}_{\text{trans},h}$ would require to solve a nonlinear minimization problem fully coupled in space and time. Therefore, to simplify the numerical scheme, we propose the use of auxiliary variables to decouple the optimization problem. This requires to solve a minimization problem on a higher-dimensional space by taking into account additional proximal operators, but each turns out to be much simpler to compute.

5.5.1 Relaxation via Slack Variables

Here, we introduce several slack variables to decouple the fully discrete optimization problem (5.7). Since in the following, we only investigate discrete spaces, we often neglect the time discretization parameter h to indicate corresponding functions.

Edge-Based Kinetic Energy. First, for the classical L^2-Wasserstein distance, we recall from (3.13) that the integrand of the kinetic energy for a pair $(\vartheta, m) \in \mathbb{R} \times \mathbb{R}$ is given by

$$\Phi(\vartheta, m) = \begin{cases} \dfrac{m^2}{\vartheta} & \text{if } \vartheta > 0\,, \\ 0 & \text{if } (\vartheta, m) = (0,0)\,, \\ \infty & \text{else}\,. \end{cases}$$

Now, for the discrete optimal transport distance, the integrand of kinetic energy in Definition 5.1.4 is given on edges via $\Phi_{\text{e}}(s,t,m) = \Phi(\theta(s,t),m)$, where $\theta(s,t)$ is a suitable average of the adjacent nodes satisfying the assumptions in Definition 5.1.2. Thus, Φ_{e} couples the momentum variable m on this edge with the mass variable ρ on the adjacent nodes x and y. Therefore, we introduce a variable ϑ representing the mass on the edges, s.t. the kinetic energy functional can be decoupled on edge values. We show that the corresponding relaxation does not change the minimizer.

Lemma 5.5.1 (Edge-Based Kinetic Energy)**.** *The set*

$$\mathcal{K}_{pre} := \left\{ (\bar{\rho}, \vartheta) \in V^0_{n,h} \times V^0_{e,h} \;:\; 0 \leqslant \vartheta(t_i, x, y) \leqslant \theta(\bar{\rho}(t_i,x), \bar{\rho}(t_i,y)) \;\forall i = 0, \ldots, N-1,\; \forall x, y \in \mathcal{X} \right\}\,.$$

is convex.
We define the edge-based kinetic energy $\mathcal{E}_{trans,e}\colon V^0_{n,h} \times V^0_{e,h} \to \mathbb{R} \cup \{\infty\}$ by

$$\mathcal{E}_{trans,e}(\vartheta, m) := \frac{1}{2} \int_0^1 \sum_{x,y \in \mathcal{X}} \Phi(\vartheta(t,x,y), m(t,x,y)) Q(x,y) \pi(x)\,\mathrm{d}t\,.$$

Then, for $(\rho, m) \in V^1_{n,h} \times V^0_{e,h}$ we can compute the kinetic energy functional $\mathcal{E}_{trans,h}$ by

$$\mathcal{E}_{trans,h}(\rho, m) = \inf \left\{ \mathcal{E}_{trans,e}(\vartheta, m) + \mathcal{I}_{\mathcal{K}_{pre}}(\text{avg}_h\,\rho, \vartheta) \;:\; \vartheta \in V^0_{e,h} \right\}\,.$$

Proof. The convexity of $\mathcal{K}_{\text{pre}}$ follows since θ is a concave function.

Now, for any $\vartheta \in V^0_{e,h}$ with $(\bar{\rho}, \vartheta) \in \mathcal{K}_{\text{pre}}$, we have that $\vartheta(t_i,x,y) \leqslant \theta(\bar{\rho}(t_i,x), \bar{\rho}(t_i,y))$. By monotonicity of Φ in its first argument, this implies

$$\Phi(\vartheta(t_i,x,y), m(t_i,x,y)) \geqslant \Phi(\theta(\bar{\rho}(t_i,x), \bar{\rho}(t_i,y)), m(t_i,x,y)) = \Phi_{\text{e}}(\bar{\rho}(t_i,x), \bar{\rho}(t_i,y), m(t_i,x,y))\,.$$

Since for $(\rho, m) \in V^1_{n,h} \times V^0_{e,h}$ we have that $\bar\rho := \operatorname{avg}_h \rho \in V^0_{n,h}$ and $\operatorname{avg}_h m = m$, this implies

$$\begin{aligned}\mathcal{E}_{\text{trans}}(\bar\rho, m) &= \mathcal{E}_{\text{trans}}(\operatorname{avg}_h \rho, \operatorname{avg}_h m) = \mathcal{E}_{\text{trans}}(\operatorname{avg}_h \rho, m)\\ &\leqslant \inf\left\{\mathcal{E}_{\text{trans,e}}(\vartheta, m) + \mathcal{I}_{\mathcal{K}_{\text{pre}}}(\bar\rho, \vartheta) \;:\; \vartheta \in V^0_{e,h}\right\}.\end{aligned}$$

To show equality, we observe that $\vartheta(t_i, x, y) := \theta(\bar\rho(t_i, x), \bar\rho(t_i, y))$ obviously satisfies $(\bar\rho, \vartheta) \in \mathcal{K}_{\text{pre}}$ and has energy $\mathcal{E}_{\text{trans,e}}(\vartheta, m) = \mathcal{E}_{\text{trans}}(\bar\rho, m)$. □

Auxiliary Variables. Now, we explicitly introduce an auxiliary variable for the average value $\operatorname{avg}_h \rho$. So far, the coupling according to the graph structure is transferred to the set $\mathcal{K}_{\text{pre}}$, where $\operatorname{avg}_h \rho$ is defined on nodes, and $\bar\rho$ is defined on edges. Therefore, we introduce auxiliary variables $\rho^-, \rho^+ \in V^0_{e,h}$ to represent the mass on a directed edge according to the adjacent node.

Lemma 5.5.2 (Decoupling in Time). *For $(\rho, \vartheta) \in V^1_{n,h} \times V^0_{e,h}$ we have that*

$$\begin{aligned}\mathcal{I}_{\mathcal{K}_{pre}}(\operatorname{avg}_h \rho, \vartheta) = \inf\Big\{&\mathcal{I}_{\mathcal{J}_{avg}}(\rho, \bar\rho) + \mathcal{I}_{\mathcal{J}_=}(\bar\rho, q) + \mathcal{I}_{\mathcal{J}_\pm}(q, \rho^-, \rho^+) + \mathcal{I}_{\mathcal{K}}(\rho^-, \rho^+, \vartheta) \;:\\ &(\bar\rho, q, \rho^-, \rho^+) \in (V^0_{n,h})^2 \times (V^0_{e,h})^2\Big\},\end{aligned}$$

where we define the following sets

$$\begin{aligned}\mathcal{J}_{avg} &:= \left\{(\rho, \bar\rho) \in V^1_{n,h} \times V^0_{n,h} \;:\; \bar\rho = \operatorname{avg}_h \rho\right\},\\ \mathcal{J}_= &:= \left\{(\bar\rho, q) \in (V^0_{n,h})^2 \;:\; \bar\rho = q\right\},\\ \mathcal{J}_\pm &:= \left\{(q, \rho^-, \rho^+) \in V^0_{n,h} \times (V^0_{e,h})^2 \;:\; q(t_i, x) = \rho^-(t_i, x, y),\, q(t_i, y) = \rho^+(t_i, x, y)\right\}, && (5.9)\\ \mathcal{K} &:= \left\{(\rho^-, \rho^+, \vartheta) \in (V^0_{e,h})^3 \;:\; (\rho^-(t_i, x, y), \rho^+(t_i, x, y), \vartheta(t_i, x, y)) \in K\right\}, && (5.10)\end{aligned}$$

with

$$K := \{(\rho^-, \rho^+, \vartheta) \in \mathbb{R}^3 \;:\; 0 \leqslant \vartheta \leqslant \theta(\rho^-, \rho^+)\}.$$

Proof. For fixed $\rho \in V^1_{n,h}$ there is precisely one tuple $(\bar\rho, q, \rho^-, \rho^+)$ s.t.

$$(\rho, \bar\rho) \in \mathcal{J}_{avg}, \quad (\bar\rho, q) \in \mathcal{J}_=, \text{ and } \quad (q, \rho^-, \rho^+) \in \mathcal{J}_\pm,$$

i.e., the tuple is given by $\bar\rho = \operatorname{avg}_h \rho$, $q = \bar\rho$, $\rho^-(t_i, x, y) = q(t_i, x)$, and $\rho^+(t_i, x, y) = q(t_i, y)$. For this (ρ^-, ρ^+) we find $(\rho^-, \rho^+, \vartheta) \in \mathcal{K}$ if and only if $(\operatorname{avg}_h \rho, \vartheta) \in \mathcal{K}_{\text{pre}}$. □

Later, the additional set $\mathcal{J}_=$ simplifies the partition of the final optimization problem into a primal and a dual component. Indeed, the sets $\mathcal{J}_{avg}$, $\mathcal{J}_=$, $\mathcal{J}_\pm$ and $\mathcal{K}$ are products of simpler low-dimensional sets, implying more straightforward computations of the relevant proximal mappings and projections.

Splitting of the Relaxed Energy Functional. Finally, we arrive at an equivalent formulation for the discrete minimization problem (5.7):

$$\begin{aligned}\mathcal{W}_{G,h}(\rho_A, \rho_B)^2 = \inf\Big\{&(\mathcal{F} + \mathcal{G})(\rho, m, \vartheta, \rho^-, \rho^+, \bar\rho, q) \;:\\ &(\rho, m, \vartheta, \rho^-, \rho^+, \bar\rho, q) \in V^1_{n,h} \times (V^0_{e,h})^4 \times (V^0_{n,h})^2\Big\}\end{aligned} \tag{5.11}$$

with functionals

$$\begin{aligned}\mathcal{F}(\rho, m, \vartheta, \rho^-, \rho^+, \bar\rho, q) &:= \mathcal{E}_{\text{trans,e}}(\vartheta, m) + \mathcal{I}_{\mathcal{J}_\pm}(q, \rho^-, \rho^+) + \mathcal{I}_{\mathcal{J}_{avg}}(\rho, \bar\rho),\\ \mathcal{G}(\rho, m, \vartheta, \rho^-, \rho^+, \bar\rho, q) &:= \mathcal{I}_{\mathcal{CE}_h(\rho_A, \rho_B)}(\rho, m) + \mathcal{I}_{\mathcal{K}}(\rho^-, \rho^+, \vartheta) + \mathcal{I}_{\mathcal{J}_=}(\bar\rho, q).\end{aligned}$$

Here, the splitting into $\mathcal{F}$ and $\mathcal{G}$ already fits into the requirements for a proximal splitting algorithm as presented in Section 3.2.3. For our numerical scheme, we use the Chambolle–Pock algorithm (3.11), where we consider the Hilbert space $H = V^1_{n,h} \times (V^0_{e,h})^4 \times (V^0_{n,h})^2$ with the scalar product

$$\begin{aligned}
&\left\langle \left(\rho_1, m_1, \vartheta_1, \rho_1^-, \rho_1^+, \bar\rho_1, q_1\right), \left(\rho_2, m_2, \vartheta_2, \rho_2^-, \rho_2^+, \bar\rho_2, q_2\right)\right\rangle_H \\
&\quad := h\sum_{i=0}^{N} \langle \rho_1(t_i,\cdot), \rho_2(t_i,\cdot)\rangle_\pi + h\sum_{i=0}^{N-1} \langle \bar\rho_1(t_i,\cdot), \bar\rho_2(t_i,\cdot)\rangle_\pi + \langle q_1(t_i,\cdot), q_2(t_i,\cdot)\rangle_\pi \\
&\qquad + h\sum_{i=0}^{N-1} \langle m_1(t_i,\cdot), m_2(t_i,\cdot)\rangle_Q + \langle \vartheta_1(t_i,\cdot), \vartheta_2(t_i,\cdot)\rangle_Q \\
&\qquad + h\sum_{i=0}^{N-1} \langle \rho_1^-(t_i,\cdot), \rho_2^-(t_i,\cdot)\rangle_Q + \langle \rho_1^+(t_i,\cdot), \rho_2^+(t_i,\cdot)\rangle_Q \,.
\end{aligned} \tag{5.12}$$

and the induced norm denoted by $\|\cdot\|_H$, which is used for the proximal mappings. Note that by Moreau's decomposition (see Theorem 3.2.9), the proximal map of the Fenchel dual can be computed by the proximal map of the primal and vice-versa. Furthermore, the choice of our slack variables allows to compute the proximal maps of the involved six operators for $\mathcal{F}$ and $\mathcal{G}$ separately. In Figure 5.3, we summarize the proximal splitting algorithm, including the following observations concerning the computational methods of the particular operators.

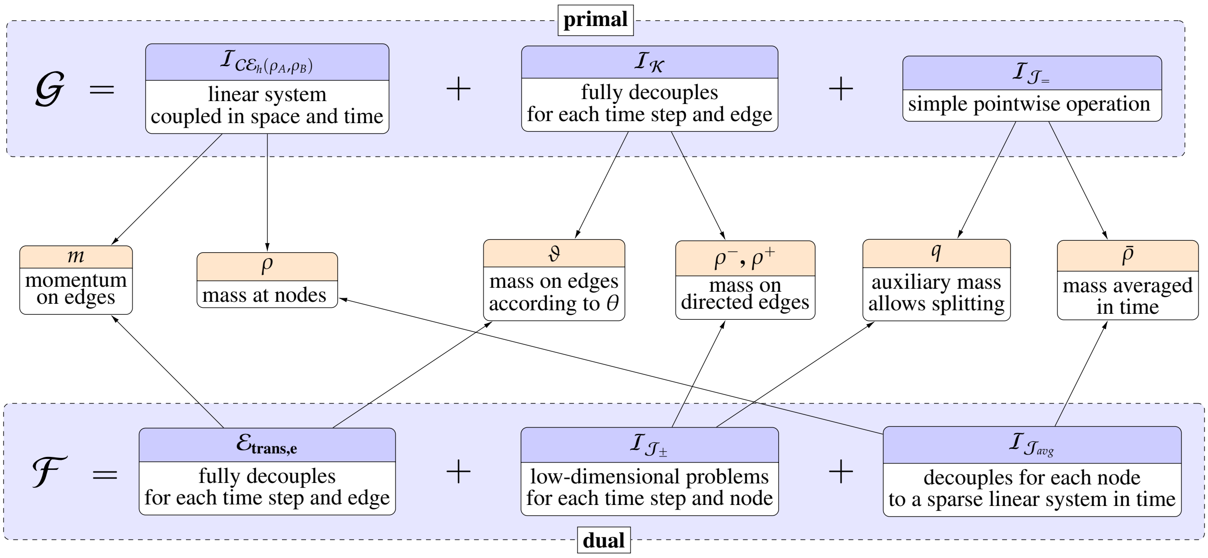

primal
$\mathcal{G}$ =
$\mathcal{I}_{\mathcal{CE}_h(\rho_A,\rho_B)}$
linear system coupled in space and time
+
$\mathcal{I}_{\mathcal{K}}$
fully decouples for each time step and edge
+
$\mathcal{I}_{\mathcal{J}_=}$
simple pointwise operation
m
momentum on edges
ρ
mass at nodes
ϑ
mass on edges according to θ
ρ^-, ρ^+
mass on directed edges
q
auxiliary mass allows splitting
$\bar\rho$
mass averaged in time
$\mathcal{F}$ =
$\mathcal{E}_{\mathbf{trans,e}}$
fully decouples for each time step and edge
+
$\mathcal{I}_{\mathcal{J}_\pm}$
low-dimensional problems for each time step and node
+
$\mathcal{I}_{\mathcal{J}_{avg}}$
decouples for each node to a sparse linear system in time
dual

Figure 5.3: Sketch of proximal splitting algorithm.

5.5.2 Projection onto $\mathcal{CE}_h(\rho_A, \rho_B)$

In analogy to the classical optimal transport distance, we show that projecting onto $\mathcal{CE}_h(\rho_A, \rho_B)$ requires solving an elliptic problem on the time-space domain (*cf*. Lemma 3.2.14).

Proposition 5.5.3 (Projection onto $\mathcal{CE}_h(\rho_A, \rho_B)$)**.** *Given $(\rho, m) \in V^1_{n,h} \times V^0_{e,h}$, the solution (ρ^{pr}, m^{pr}) to the projection problem*

$$\mathrm{proj}_{\mathcal{CE}_h(\rho_A,\rho_B)}(\rho, m) = \operatorname*{arg\,min}_{(\rho^{pr}, m^{pr}) \in \mathcal{CE}_h(\rho_A,\rho_B)} \frac{h}{2}\sum_{i=0}^{N} \|\rho^{pr}(t_i,\cdot) - \rho(t_i,\cdot)\|_\pi^2 + \frac{h}{2}\sum_{i=0}^{N-1} \|m^{pr}(t_i,\cdot) - m(t_i,\cdot)\|_Q^2$$

is given by

$$\rho^{pr}(t_i,x) = \rho(t_i,x) + \frac{\varphi(t_i,x)-\varphi(t_{i-1},x)}{h} \qquad \forall i = 1,\dots,N-1\,, \tag{5.13a}$$

$$\rho^{pr}(t_0,x) = \rho_A(x)\,,\ \rho^{pr}(t_N,x) = \rho_B(x)\,, \tag{5.13b}$$

$$m^{pr}(t_i,x,y) = m(t_i,x,y) + \nabla_{\mathcal{X}}\varphi(t_i,x,y) \qquad \forall i = 0,\dots,N-1\,, \tag{5.13c}$$

where φ solves the elliptic equation on the time-space domain

$$\begin{aligned}
&\pi(x)\frac{\varphi(t_1,x)-\varphi(t_0,x)}{h^2} + \pi(x)\,\Delta_{\mathcal{X}}\varphi(t_0,x)\\
&\qquad\qquad = -\pi(x)\left(\frac{\rho(t_1,x)-\rho_A(x)}{h} + \mathrm{div}_{\mathcal{X}} m(t_0,x)\right)\,,\\
&\pi(x)\frac{-\varphi(t_{N-1},x)+\varphi(t_{N-2},x)}{h^2} + \pi(x)\Delta_{\mathcal{X}}\varphi(t_{N-1},x)\\
&\qquad\qquad = -\pi(x)\left(\frac{\rho_B(x)-\rho(t_{N-1},x)}{h} + \mathrm{div}_{\mathcal{X}} m(t_{N-1},x)\right)\,,\\
&\pi(x)\frac{\varphi(t_{i+1},x)-2\varphi(t_i,x)+\varphi(t_{i-1},x)}{h^2} + \pi(x)\,\Delta_{\mathcal{X}}\varphi(t_i,x)\\
&\qquad\qquad = -\pi(x)\left(\frac{\rho(t_{i+1},x)-\rho(t_i,x)}{h} + \mathrm{div}_{\mathcal{X}} m(t_i,x)\right)\,,
\end{aligned} \tag{5.14}$$

for $i = 1,\dots,N-2$ and $x \in \mathcal{X}$.

The factors $\pi(x)$ in (5.14) could be canceled but they simplify further analysis.

Proof. We define the Lagrangian corresponding to the projection problem as

$$\begin{aligned}
\mathcal{L}(\rho^{\mathrm{pr}},m^{\mathrm{pr}},\varphi,\lambda_A,\lambda_B) =&\frac{h}{2}\sum_{i=0}^{N}\|\rho^{\mathrm{pr}}(t_i,\cdot)-\rho(t_i,\cdot)\|_\pi^2 + \frac{h}{2}\sum_{i=0}^{N-1}\|m^{\mathrm{pr}}(t_i,\cdot)-m(t_i,\cdot)\|_Q^2\\
&+h\sum_{i=0}^{N-1}\sum_{x\in\mathcal{X}}\varphi(t_i,x)\left(\frac{\rho^{\mathrm{pr}}(t_{i+1},x)-\rho^{\mathrm{pr}}(t_i,x)}{h} + \mathrm{div}_{\mathcal{X}} m^{\mathrm{pr}}(t_i,x)\right)\pi(x)\\
&+\sum_{x\in\mathcal{X}}\left(\lambda_B(x)(\rho^{\mathrm{pr}}(t_N,x)-\rho_B(x)) + \lambda_A(x)(\rho^{\mathrm{pr}}(t_0,x)-\rho_A(x))\right)\pi(x)\,,
\end{aligned}$$

where λ_A,λ_B are the Lagrange multipliers for the boundary conditions $\rho^{\mathrm{pr}}(t_0,\cdot) = \rho_A$ and $\rho^{\mathrm{pr}}(t_N,\cdot) = \rho_B$. The optimality conditions in ρ^{pr} and m^{pr} imply (5.13a) and (5.13c). Furthermore, (5.13b) reflects the boundary conditions, which are to be ensured in $\mathcal{CE}_h(\rho_A,\rho_B)$. Inserting these relations into the continuity equation $\partial_t\rho^{\mathrm{pr}} + \mathrm{div}_{\mathcal{X}} m^{\mathrm{pr}} = 0$ leads to the system of equations (5.14). □

Now, the system (5.14) can be written as a linear system $SZ = F$ for a coordinate vector

$$Z = (\varphi(t_i,x))_{i=0,\dots N-1,\,x\in X}$$

representing a function $\varphi \in V_{n,h}^0$ in the canonical basis

$$(\varphi^{i,x})_{i=0,\dots,N-1,\,x\in\mathcal{X}}\,,\quad \text{where}\quad (\varphi^{i,x})(t_j,y) = \delta_{i,j}\cdot\delta_{x,y}\,,$$

and the standard Euclidean inner product with respect to this basis. Furthermore, $F \in \mathbb{R}^{N|\mathcal{X}|}$ is a vector and $S \in \mathbb{R}^{(N|\mathcal{X}|)\times(N|\mathcal{X}|)}$ is a matrix, which is symmetric since $\pi(x)Q(x,y) = \pi(y)Q(y,x)$ and sparse if Q is sparse. However, the matrix S is not invertible. Thus, to solve the linear system, we first compute the kernel of S.

Lemma 5.5.4 (Kernel of System Matrix)**.** *The kernel of S is spanned by functions that are constant in space and time.*

Proof. Assume that there is Z is in the kernel of S, which is not constant. We denote by ϕ_h the associated function in $V^0_{n,h}$. Now, let $I_+(\mu) := \{(i,x) \in \{0,\dots,N-1\} \times \mathcal{X} \,:\, \phi_h(i,x) > \mu\}$ for $\mu = \min \phi_h(i,x)$ and define $\varphi \in V^0_{n,h}$ via $\varphi(t_i,x) = 1$ if $(i,x) \in I_+(\mu)$ and $\varphi(t_i,x) = 0$ else. Let W be the associated nodal vector to φ. By assumption on Z the set $I_+(\mu)$ is nonempty and thus it is easy to see that $W^\top SZ < 0$ and thus Z cannot be in the kernel of S, which proves the claim. □

Thus, we introduce another Lagrange multiplier λ to impose the constraint $\sum_{i=0}^{N-1}\sum_{x\in\mathcal{X}} \varphi(t_i,x) = 0$. Considering the vector $w \in \mathbb{R}^{N|\mathcal{X}|}$ with entries $w^{i,x} = 1$ this constraint can be written as $w^\top \varphi = 0$. Then the linear system

$$\begin{pmatrix} S & w \\ w^T & 0 \end{pmatrix} \begin{pmatrix} Z \\ \lambda \end{pmatrix} = \begin{pmatrix} F \\ 0 \end{pmatrix}$$

is uniquely solvable, and the solution implies $\lambda = 0$ if $F \perp w$ (in the Euclidean sense), which is true because ρ_A and ρ_B are assumed to be of equal mass. Note that any $\tilde{Z} = Z + W$ with W in the kernel of S would not change the updates (5.13a) and (5.13c), since the contributions $\frac{W(t_i,x)-W(t_{i-1},x)}{h}$ and $\nabla_{\mathcal{X}} W(t_i,x,y)$ are zero.

5.5.3 Proximal Mapping of $\mathcal{E}_{\text{trans,e}}$

We recall from Section 3.2.3 that the computation of the proximal mapping of $\mathcal{E}_{\text{trans,e}}$ (see Proposition 3.2.12) is also required for the numerical solution scheme for the classical optimal transport distance. Now, for the discrete transport distance, it is crucial that we have decoupled the variables in such a way that the computation of $\mathcal{E}_{\text{trans,e}}$ can be performed pointwise. This is satisfied, since for $(\vartheta, m) \in (V^0_{e,h})^2$ we have

$$\mathcal{E}_{\text{trans,e}}(\vartheta,m) = \frac{h}{2}\sum_{i=0}^{N-1}\sum_{x,y\in\mathcal{X}} \Phi(\vartheta(t_i,x,y), m(t_i,x,y))Q(x,y)\pi(x)\,,$$

and thus, for $(p,q) \in (V^0_{e,h})^2$, we obtain for the dual

$$\begin{aligned}\mathcal{E}^*_{\text{trans,e}}(p,q) = \sup_{(\vartheta,m)\in(V^0_{e,h})^2} h \sum_{i=0}^{N-1} \Big[&\langle p(t_i,\cdot,\cdot), \vartheta(t_i,\cdot,\cdot)\rangle_Q + \langle q(t_i,\cdot,\cdot), m(t_i,\cdot,\cdot)\rangle_Q \\ &- \frac{1}{2}\sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \Phi(\vartheta(t_i,x,y), m(t_i,x,y))Q(x,y)\pi(x)\Big] \\ = \frac{h}{2}\sum_{i=0}^{N-1}\sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} &\Phi^*(p(t_i,x,y),q(t_i,x,y))Q(x,y)\pi(x) = \sum_{i=0}^{N-1}\sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \mathcal{I}_{\mathcal{B}}(p(t_i,x,y),q(t_i,x,y))\,,\end{aligned}$$

where $\Phi^* = \mathcal{I}_{\mathcal{B}}$ with the convex set

$$\mathcal{B} = \left\{ (p,q) \in \mathbb{R}^2 \,:\, p + \frac{q^2}{4} \leqslant 0 \right\}.$$

Therefore, the proximal mapping separates into two-dimensional problems for each time interval and edge, and for $\sigma > 0$, $(p^{\text{pr}}, q^{\text{pr}}) = \text{prox}_{\sigma\mathcal{E}^*_{\text{trans,e}}}(p,q)$, it is given by

$$(p^{\text{pr}}(t_i,x,y), q^{\text{pr}}(t_i,x,y)) = \text{proj}_{\mathcal{B}}(p(t_i,x,y), q(t_i,x,y))\,.$$

In Lemma 3.2.13, we have described the computational solution scheme of this projection problem with a Newton method.

5.5.4 Projection onto the Edge-Based Set $\mathcal{K}$

In the following, we want to compute for given $(\rho^-, \rho^+, \vartheta) \in (V^0_{e,h})^3$ the projection

$$(\rho^{-^{\mathrm{pr}}}, \rho^{+^{\mathrm{pr}}}, \vartheta^{\mathrm{pr}}) = \mathrm{proj}_{\mathcal{K}}(\rho^-, \rho^+, \vartheta)$$

onto the set $\mathcal{K}$, which is given by

$$\underset{(\rho^{-^{\mathrm{pr}}}, \rho^{+^{\mathrm{pr}}}, \vartheta^{\mathrm{pr}}) \in \mathcal{K}}{\arg\min} \; \frac{h}{2} \sum_{i=0}^{N-1} \left(\|\rho^{-^{\mathrm{pr}}}(t_i,\cdot,\cdot) - \rho^-(t_i,\cdot,\cdot)\|_Q^2 + \|\rho^{+^{\mathrm{pr}}}(t_i,\cdot,\cdot) - \rho^+(t_i,\cdot,\cdot)\|_Q^2 + \|\vartheta^{\mathrm{pr}}(t_i,\cdot,\cdot) - \vartheta(t_i,\cdot,\cdot)\|_Q^2 \right).$$

Recall from (5.10) that $\mathcal{K}$ is a product of the three-dimensional closed convex set K. Thus, the projection problem decouples into the edgewise projection

$$(\rho^{-^{\mathrm{pr}}}(t_i,x,y), \rho^{+^{\mathrm{pr}}}(t_i,x,y), \vartheta^{\mathrm{pr}}(t_i,x,y)) = \mathrm{proj}_K(\rho^-(t_i,x,y), \rho^+(t_i,x,y), \vartheta(t_i,x,y)),$$

for each time step t_i and each edge $(x,y) \in \mathcal{S}$. To compute the projection onto K, we make use of its special structure given as the graph of a concave function θ. In analogy to the subdifferential for convex functions, we denote the superdifferential of θ at a point $(s,t) \in \mathbb{R}^2$ by $\partial^+\theta(s,t) := -\partial(-\theta)(s,t)$, where $\partial(-\theta)(s,t)$ is the subdifferential of the convex function $-\theta$ at (s,t). Now, we recall from Lemma 3.2.11 that the projection $p^{\mathrm{pr}} = \mathrm{proj}_K(p)$ of $p \in \mathbb{R}^3$ is characterized by

$$p - p^{\mathrm{pr}} \in N_K(p^{\mathrm{pr}}) := \left\{ z \in \mathbb{R}^3 \;:\; \langle z, q - p^{\mathrm{pr}} \rangle \leqslant 0 \;\forall q \in K \right\},$$

where $N_K(p^{\mathrm{pr}})$ is the normal cone of K at p^{pr}. First, we observe that the computation of N_K can be distinguished into several cases.

Lemma 5.5.5 (Characterization of the Normal Cone)**.** *Let $\theta\colon \mathbb{R}^2 \to \mathbb{R}$ be an averaging function fulfilling the assumptions listed as in Definition 5.1.2 and let $K := \{p \in \mathbb{R}^3 \;:\; 0 \leqslant p_3 \leqslant \theta(p_1,p_2)\}$. Then the normal cone $N_K(p^{pr})$ at $p^{pr} \in K$ can be characterized in the following way.*

1. ***Interior Points.*** *If $p^{pr} \in \{(p_1,p_2,p_3) \in \mathbb{R}^3 : 0 < p_3 < \theta(p_1,p_2)\}$, then $N_K(p^{pr}) = \{(0,0,0)\}$.*

2. ***Bottom Facet.*** *If $p^{pr} \in \mathbb{R}_+ \times \mathbb{R}_+ \times \{0\}$, then $N_K(p^{pr}) = \{0\} \times \{0\} \times \mathbb{R}_{\leqslant 0}$.*

3. ***Coordinate Axis.*** *If $p^{pr} = (p_1^{pr}, 0, 0)$ with $p_1^{pr} \in \mathbb{R}_+$, then*

$$N_K(p^{pr}) = \{0\} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}_{\leqslant 0} \;\cup\; \left\{ (0,q_2,q_3) \in \{0\} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}_+ \;:\; \left(0, -\tfrac{q_2}{q_3}\right) \in \partial^+\theta(p_1^{pr},0) \right\}.$$

   Moreover, we have that

$$\begin{aligned} (0,q) \in \partial^+\theta(p_1^{pr},0) \quad &\Leftrightarrow \quad q \geqslant \lim_{z \searrow 0} \partial_2\theta(p_1^{pr}, z), \\ \partial^+\theta(p_1^{pr},0) = \varnothing \quad &\Leftrightarrow \quad \lim_{z \searrow 0} \partial_2\theta(p_1^{pr}, z) = \infty. \end{aligned}$$

   In analogy, a similar statement holds for $p^{pr} = (0, p_2^{pr}, 0)$ with $p_2^{pr} \in \mathbb{R}_+$.

4. ***Origin.*** *If $p^{pr} = (0,0,0)$, then*

$$N_K(p^{pr}) = (\mathbb{R}_{\leqslant 0})^3 \;\cup\; \left\{ (q_1,q_2,q_3) \in \mathbb{R}_{\leqslant 0} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}_+ \;:\; \left(\tfrac{q_1}{q_3}, \tfrac{q_2}{q_3}\right) \in -\partial^+\theta(0) \right\}.$$

5. ***Upper Surface.*** *If $p^{pr} = \left(p_1^{pr}, p_2^{pr}, \theta(p_1^{pr}, p_2^{pr})\right)$ for $(p_1^{pr}, p_2^{pr}) \in \mathbb{R}^2_+$, then*

$$N_K(p^{pr}) = \left\{ \lambda \left(-\partial_1\theta(p_1^{pr}, p_2^{pr}), -\partial_2\theta(p_1^{pr}, p_2^{pr}), 1\right) \;:\; \lambda \in \mathbb{R}_+ \right\}.$$

Proof. We distinguish the particular cases.

Interior Points and Bottom Surface. These two cases trivially hold true.

Coordinate Axis. Let $p^{\mathrm{pr}} = (p_1^{\mathrm{pr}}, 0, 0)$ with $p_1^{\mathrm{pr}} > 0$. First, there exists $\varepsilon > 0$ s.t. the points $(p_1^{\mathrm{pr}} + \varepsilon, 0, 0)$, $(p_1^{\mathrm{pr}} - \varepsilon, 0, 0)$, and $(p_1^{\mathrm{pr}}, \varepsilon, 0)$ are in the set K, hence $N_K(p^{\mathrm{pr}}) \subset \{0\} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}$. Now, the planes $\mathbb{R} \times \{0\} \times \mathbb{R}$ and $\mathbb{R} \times \mathbb{R} \times \{0\}$ contain the point p^{pr}, but do not intersect the interior of K, thus $\{0\} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}_{\leqslant 0} \subset N_K(p^{\mathrm{pr}})$. Next, we investigate the set $\{0\} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}_{\leqslant 0}$. By definition of the set K, a point $(0, -z_2, 1) \in N_K(p^{\mathrm{pr}})$ is determined by the condition $z = (0, z_2) \in \partial^+ \theta(p_1^{\mathrm{pr}}, 0)$. We define a function $f\colon t \mapsto \theta(p_1^{\mathrm{pr}}, t)$ s.t. $z_2 \in \partial^+ f(0)$. Conversely, if $z_2 \notin \partial^+ f(0)$ then it follows that $(0, -z_2, 1) \notin N_K(p^{\mathrm{pr}})$. Thus, the normal cone is given by

$$N_K(p^{\mathrm{pr}}) = \{0\} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}_{\leqslant 0} \cup \{(0, -\lambda \cdot z, \lambda) \ : \ z \in \partial^+ f(0) \, , \, \lambda \in \mathbb{R}_+\} \, .$$

Because the auxiliary function f is concave and by monotonicity of the superdifferential, we observe that

$$\partial^+ f(0) = [\lim_{z \searrow 0} \partial_2 \theta(p_1^{\mathrm{pr}}, z), \infty) \, .$$

Origin. Let $p^{\mathrm{pr}} = (0, 0, 0)$. First, we observe that $(\mathbb{R}_{\leqslant 0})^3 \subset N_K(0) \subset \mathbb{R}_{\leqslant 0} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}$. To investigate the set $\mathbb{R}_{\leqslant 0} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}_+$, we consider the superdifferential of θ. Indeed, for every $z = (z_1, z_2) \in \partial^+ \theta(0)$, we have that $(-z_1, -z_2, 1) \in N_K(0)$. Conversely, $z = (z_1, z_2) \notin \partial^+ \theta(0)$ implies $(-z_1, -z_2, 1) \notin N_K(0)$.

Upper Surface. Let $p^{\mathrm{pr}} = (p_1^{\mathrm{pr}}, p_2^{\mathrm{pr}}, \theta(p_1^{\mathrm{pr}}, p_2^{\mathrm{pr}}))$ with $(p_1^{\mathrm{pr}}, p_2^{\mathrm{pr}}) \in \mathbb{R}^2_+$. Then there exists a neighborhood of p^{pr} s.t. K is the subgraph of a concave and differentiable function. Hence, the normal cone is spanned by the single outer normal vector $(-\partial_1 \theta(p_1^{\mathrm{pr}}, p_2^{\mathrm{pr}}), -\partial_2 \theta(p_1^{\mathrm{pr}}, p_2^{\mathrm{pr}}), 1)$. □

Now, from Lemma 5.5.5 we can extract the following algorithm.

Algorithm 5.1 Projection onto the cone K

```
function ProjectK($p_1$,$p_2$,$p_3$)
    if $0 \leqslant p_3 \leqslant \theta(p_1, p_2)$ return $(p_1, p_2, p_3)$
    if $p_3 \leqslant 0$ return $(\max\{p_1, 0\}, \max\{p_2, 0\}, 0)$
    if $(p_1 > 0) \wedge (p_2 \leqslant 0)$ then
        if $-p_2/p_3 \geqslant \lim_{z \searrow 0} \partial_2 \theta(p_1, z)$ return $(p_1, 0, 0)$
    end if
    if $(p_1 \leqslant 0) \wedge (p_2 > 0)$ then
        if $-p_1/p_3 \geqslant \lim_{z \searrow 0} \partial_1 \theta(z, p_2)$ return $(0, p_2, 0)$
    end if
    if $(p_1 \leqslant 0) \wedge (p_2 \leqslant 0)$ then
        if $(-p_1/p_3, -p_2/p_3) \in \partial^+ \theta(0)$ return $(0, 0, 0)$
    end if
    return ProjectKTop($p_1$,$p_2$,$p_3$)
end function
```

Thus, for a fully explicit solution scheme for a specific choice of θ, we still have to compute

1. the limits $\lim_{z \searrow 0} \partial_2 \theta(p_1, z)$ and $\lim_{z \searrow 0} \partial_1 \theta(z, p_2)$,
2. the superdifferential $\partial^+ \theta(0)$ at the origin, and
3. the function ProjectKTop(p_1,p_2,p_3) for the projection onto the upper surface corresponding to the case 5 in Lemma 5.5.5.

Next, we describe a general procedure to reduce the projection problem onto the upper surface to a one-dimensional optimization problem. Here, we essentially make use of the 1-homogeneity of θ. More precisely, it is sufficient to consider a curve $c\colon \mathbb{R}_+ \to \mathbb{R}^2$ of type

$$c(q) = (q^{-1/2}, q^{1/2}) \, .$$

Then all points on the upper surface can be expressed in terms of $\theta(c(q))$. A similar dimension reduction allows characterizing the superdifferential of θ at the origin by taking into account the curve c.

Lemma 5.5.6 (Projection onto the Upper Surface of K)**.** *Let $p \in \mathbb{R}^3$ s.t. the projection onto K is given by $p^{pr} = (p_1^{pr}, p_2^{pr}, \theta(p_1^{pr}, p_2^{pr}))$ with $(p_1^{pr}, p_2^{pr}) \in \mathbb{R}^2_+$. Consider a parametrized curve $w(q) = (q^{1/2}, q^{-1/2}, \theta(q^{1/2}, q^{-1/2}))$ on the upper surface of K with corresponding normal $n(q) = (-\partial_1\theta(q^{1/2}, q^{-1/2}), -\partial_2\theta(q^{1/2}, q^{-1/2}), 1)$. Then there exists a unique $(q, \tau) \in \mathbb{R}^2_+$ s.t. $p^{pr} = \tau w(q)$. More precisely, q is given as the unique root of the function $f(q) := \langle p, w(q) \times n(q)\rangle$ and $\tau = \langle p, \frac{w(q)}{\|w(q)\|^2}\rangle$.*

Proof. Let $p^{\mathrm{pr}} = (p_1^{\mathrm{pr}}, p_2^{\mathrm{pr}}, \theta(p_1^{\mathrm{pr}}, p_2^{\mathrm{pr}}))$ with $(p_1^{\mathrm{pr}}, p_2^{\mathrm{pr}}) \in \mathbb{R}^2_+$. By 1-homogeneity of θ, there exists a unique $q \in \mathbb{R}_+$ and $\tau \in \mathbb{R}_+$ s.t. $p^{\mathrm{pr}} = \tau \cdot w(q)$, where

$$q = \frac{p_1^{\mathrm{pr}}}{p_2^{\mathrm{pr}}}, \quad \text{and} \quad \tau = \left(p_1^{\mathrm{pr}} p_2^{\mathrm{pr}}\right)^{\frac{1}{2}}.$$

By definition, p lies in the plane spanned by $w(q)$ and $n(q)$, *i.e.*, $f(q) = \langle p, w(q) \times n(q)\rangle = 0$. Since the projection p^{pr} is unique, q must be the unique root of f. Then τ is given as the unique solution to the one-dimensional projection $\tau \mapsto \frac{1}{2}\|p - \tau \cdot w(q)\|^2$ onto the corresponding ray. □

Lemma 5.5.7 (Characterization of the Superdifferential of θ)**.** *The superdifferential of θ at the origin is given by*

$$\partial^+\theta(0) = \overline{\{\nabla\theta(q^{-1/2}, q^{1/2}) \ : \ q \in \mathbb{R}_+\}} + (\mathbb{R}_{\geqslant 0})^2 .$$

Proof. We consider the curve $c(q) = (q^{-1/2}, q^{1/2})$. First, we have to verify that any $r \in \{\nabla\theta(c(q)) \ : \ q \in \mathbb{R}_+\}$ is contained in $\partial^+\theta(0)$, *i.e.*, for every $p \in \mathbb{R}^2_+$, we have to show that $\langle r, p\rangle \geqslant \theta(p)$. Let $q \in \mathbb{R}_+$ s.t. $r = \nabla\theta(c(q))$. By concavity of θ, we get that $\langle\nabla\theta(c(q)), p - c(q)\rangle \geqslant \theta(p) - \theta(c(q))$. Then by 1-homogeneity of θ, for every $\lambda > 0$, we observe that $\nabla\theta(\lambda c(q)) = \nabla\theta(c(q))$. Thus, we obtain that $\langle\nabla\theta(c(q)), p - \lambda c(q)\rangle \geqslant \theta(p) - \theta(\lambda c(q))$. Passing to the limit $\lambda \to 0$ and using the continuity of θ on $\mathbb{R}^2_{\geqslant 0}$ leads to $\langle r, p\rangle \geqslant \theta(p)$. Because $\partial^+\theta(0)$ is a closed set (see, *e.g.*, [BC17, Proposition 16.4]), we conclude that

$$\overline{\{\nabla\theta(c(q)) \ : \ q \in \mathbb{R}_+\}} + (\mathbb{R}_{\geqslant 0})^2 \subset \partial^+\theta(0) .$$

In contrast, for every $w \in \mathbb{R}^2 \backslash (\mathbb{R}_{\geqslant 0})^2$ there exists $p \in \mathbb{R}^2_+$ s.t. $\theta(0) + \langle r + w, p\rangle < \theta(p)$, since $\langle w, p\rangle$ can be chosen arbitrarily small. □

Logarithmic Mean. Here, we consider the specific case of the logarithmic mean $\theta = \theta_{\log}$ (see (5.4)). Then for $s > 0$, we have that $\lim_{t\searrow 0} \partial_1\theta_{\log}(t, s) = \lim_{t\searrow 0} \partial_2\theta_{\log}(s, t) = \infty$. Therefore, we can explicitly describe the normal cone at $(s, 0, 0)$ by $N_K(s, 0, 0) = \{0\} \times \mathbb{R}_{\leqslant 0} \times \mathbb{R}_{\leqslant 0}$ In analogy, the normal cone at $(0, s, 0)$ is given by $N_K(0, s, 0) = \mathbb{R}_{\leqslant 0} \times \{0\} \times \mathbb{R}_{\leqslant 0}$. Consequently, the Algorithm 5.1 simplifies as follows.

Algorithm 5.2 Projection onto the cone K for $\theta_{\log}$

```
function ProjectK($p_1$,$p_2$,$p_3$)
    if $0 \leqslant p_3 \leqslant \theta_{\log}(p_1, p_2)$ return $(p_1, p_2, p_3)$
    if $p_3 \leqslant 0$ return $(\max\{p_1, 0\}, \max\{p_2, 0\}, 0)$
    if $(p_1 \leqslant 0) \wedge (p_2 \leqslant 0) \wedge (-p_1/p_3, -p_2/p_3) \in \partial^+\theta_{\log}(0)$ return $(0, 0, 0)$
    return ProjectKTop($p_1$,$p_2$,$p_3$)
end function
```

Now, we characterize the superdifferential $\partial^+\theta_{\log}(0)$.

Lemma 5.5.8 (Superdifferential of $\theta_{\log}$)**.** *Let $z = (z_1, z_2) \in \mathbb{R}^2$. If $\min\{z_1, z_2\} \leqslant 0$, then $z \notin \partial^+\theta_{log}(0)$. Otherwise, there is a unique $q_1 \in \mathbb{R}_+$ s.t. $\partial_1\theta_{log}(q_1^{-1/2}, q_1^{1/2}) = z_1$ and in this case $z \in \partial^+\theta_{log}(0)$ if and only if $z_2 \geqslant \partial_2\theta_{log}(q^{-1/2}, q^{1/2})$.*

Proof. For the logarithmic mean, we have that $\partial^+\theta_{\log}(0) \subset \mathbb{R}^2_+$, and thus, $z \notin \partial^+\theta_{\log}(0)$ if $\min\{z_1, z_2\} \leqslant 0$. We observe that the partial derivative

$$\partial_1\theta_{\log}(q^{-1/2}, q^{1/2}) = \frac{q - 1 - \log(q)}{\log^2(q)}$$

is monotone increasing in q with $\partial_1\theta_{\log}(q^{-1/2}, q^{1/2}) \to 0$ as $q \to 0$ and $\partial_1\theta_{\log}(q^{-1/2}, q^{1/2}) \to \infty$ as $q \to \infty$. Indeed, for $\beta(q) = \partial_1\theta_{\log}(q^{-1/2}, q^{1/2})$ with $\beta(1) = \frac{1}{2}$, we obtain a continuous extension on $\mathbb{R}_+$. Furthermore, we consider $\beta'(q) = \frac{2(1-q)+\log(q)(1+q)}{q\log^3(q)}$ with continuous extension $\frac{1}{6}$ for $q = 1$. We verify that $2(1-q) + \log(q)(1+q)$ is negative for $q < 1$ and positive for $q > 1$. This implies that $\beta'(q) > 0$. Furthermore, by symmetry, we obtain that the partial derivative $\partial_2\theta_{\log}(q^{-1/2}, q^{1/2})$ is monotone decreasing in q with $\partial_2\theta_{\log}(q^{-1/2}, q^{1/2}) \to \infty$ as $q \to 0$ and $\partial_2\theta_{\log}(q^{-1/2}, q^{1/2}) \to 0$ as $q \to \infty$. By the general characterization result in Lemma 5.5.7, we get that

$$\partial^+\theta_{\log}(0) = \{\nabla\theta_{\log}(q^{-1/2}, q^{1/2}) \;:\; q \in \mathbb{R}_+\} + (\mathbb{R}_{\geqslant 0})^2 \,.$$

Thus, for every $z \in \mathbb{R}^2_+$, there is a unique $q_1 \in \mathbb{R}_+$ s.t. $\partial_1\theta_{\log}(q_1^{-1/2}, q_1^{1/2}) = z_1$ with the property that $z_1 \geqslant \partial_1\theta_{\log}(q^{-1/2}, q^{1/2})$ if and only if $q \leqslant q_1$. Furthermore, there is a unique $q_2 \in \mathbb{R}_+$ s.t. $\partial_2\theta_{\log}(q_2^{-1/2}, q_2^{1/2}) = z_2$ with the property that $z_2 \geqslant \partial_2\theta_{\log}(q^{-1/2}, q^{1/2})$ if and only if $q \geqslant q_2$. Hence, $z \in \partial^+\theta_{\log}(0)$ if and only if $q_2 \leqslant q_1$, which is equivalent to $z_2 \geqslant \partial_2\theta_{\log}(q_1^{-1/2}, q_1^{1/2})$. □

Remark 5.5.9 (Numerical Implementation). To determine q in Lemma 5.5.8, we can implement a one-dimensional Newton iteration. Note that the function $q \mapsto \partial_1\theta_{\log}(q^{-1/2}, q^{1/2})$ becomes increasingly steep as $q \to 0$, which leads to increasingly unstable Newton iterations as $z_1 \to 0$, whereas for $q \in [1, \infty)$ the function is rather flat and easy to invert. To avoid the numerical instability for $q \to 0$, note that the roles of z_1 and z_2 in Lemma 5.5.8 can be swapped using the transformation $q \leftrightarrow q^{-1}$. Moreover, for $\max\{z_1, z_2\} < \frac{1}{2}$, we have $z \notin \partial^+\theta_{\log}(0)$. Thus, by swapping the values of z_1 and z_2 if $z_1 < z_2$ we can always remain in the regime $q \in [1, \infty)$. Additionally, we recommend to replace the function $\theta_{\log}(s, t)$ and its derivatives by a local Taylor expansion near the diagonal.

Geometric Mean. Now, we consider the specific case of the geometric mean $\theta = \theta_{\mathrm{geo}}$ (see (5.4)). For $s > 0$ we again find $\lim_{t\searrow 0}\partial_1\theta_{\mathrm{geo}}(t, s) = \lim_{t\searrow 0}\partial_2\theta_{\mathrm{geo}}(s, t) = \infty$ and consequently the same simplification of the algorithm applies as in the case of the logarithmic mean. For the test of the inclusion $z = (z_1, z_2) \in \partial^+\theta_{\mathrm{geo}}(0)$, we argue as in the proof of Lemma 5.5.8. The functions $\partial_1\theta_{\mathrm{geo}}(q^{-1/2}, q^{1/2}) = \frac{1}{2}q^{\frac{1}{2}}$ and $\partial_2\theta_{\mathrm{geo}}(q^{-1/2}, q^{1/2}) = \frac{1}{2}q^{-\frac{1}{2}}$ have the same monotonicity properties as for the logarithmic mean. Therefore, if $\min\{z_1, z_2\} \leqslant 0$ then $z \notin \partial^+\theta_{\mathrm{geo}}(0)$. Otherwise, $q_1 = 4z_1^2$ and thus the condition $\partial_2\theta_{\mathrm{geo}}(q_1^{-1/2}, q_1^{1/2}) \leqslant z_2$ is equivalent to $z_1 z_2 \geqslant \frac{1}{4}$. To summarize, we have obtained

$$\partial^+\theta_{\mathrm{geo}}(0) = \left\{z \in \mathbb{R}^2 \;:\; z_1 z_2 \geqslant \tfrac{1}{4} \wedge \min\{z_1, z_2\} > 0\right\} \,.$$

5.5.5 Proximal Mappings of Auxiliary Operators

Proximal Mapping of $\mathcal{I}_{\mathcal{J}_\pm}$. Given a point $(q, \rho^-, \rho^+) \in V^0_{n,h} \times (V^0_{e,h})^2$, the proximal map of $\mathcal{I}_{\mathcal{J}_\pm}$ is given by the projection $\mathrm{proj}_{\mathcal{J}_\pm}(q, \rho^-, \rho^+)$. Thus, we have to find the minimizer $(q^{\mathrm{pr}}, \rho^{-\mathrm{pr}}, \rho^{+\mathrm{pr}}) \in \mathcal{J}_\pm$ of

$$\sum_{i=0}^{N-1} \|q^{\mathrm{pr}}(t_i, \cdot) - q(t_i, \cdot)\|_\pi^2 + \|\rho^{-\mathrm{pr}}(t_i, \cdot, \cdot) - \rho^-(t_i, \cdot, \cdot)\|_Q^2 + \|\rho^{+\mathrm{pr}}(t_i, \cdot, \cdot) - \rho^+(t_i, \cdot, \cdot)\|_Q^2 \,.$$

Recall that for any $q^{\mathrm{pr}} \in V^0_{n,h}$ there is precisely one pair $(\rho^{-\mathrm{pr}}, \rho^{+\mathrm{pr}}) \in (V^0_{e,h})^2$ s.t. $(q^{\mathrm{pr}}, \rho^{-\mathrm{pr}}, \rho^{+\mathrm{pr}}) \in \mathcal{J}_\pm$, see (5.9). Therefore, we have to find $q^{\mathrm{pr}} \in V^0_{n,h}$ that minimizes

$$\begin{aligned}\sum_{i=0}^{N-1} \Big(\sum_{x\in\mathcal{X}} |q^{\mathrm{pr}}(t_i, x) - q(t_i, x)|^2\pi(x) + \frac{1}{2}\sum_{(x,y)\in\mathcal{X}^2} |q^{\mathrm{pr}}(t_i, x) - \rho^-(t_i, x, y)|^2 Q(x, y)\pi(x) \\ + \frac{1}{2}\sum_{(x,y)\in\mathcal{X}^2} |q^{\mathrm{pr}}(t_i, y) - \rho^+(t_i, x, y)|^2 Q(x, y)\pi(x)\Big) \,.\end{aligned}$$

Taking into account the detailed balance condition $Q(x,y)\pi(x) = Q(y,x)\pi(y)$, the optimality condition in q^{pr} for $i = 0,\dots,N-1$, $x \in \mathcal{X}$ is given by

$$q^{\mathrm{pr}}(t_i,x) = \frac{1}{1+\sum_{y\in\mathcal{X}} Q(x,y)}\left(q(t_i,x) + \frac{1}{2}\sum_{y\in\mathcal{X}}(\rho^-(t_i,x,y) + \rho^+(t_i,y,x))Q(x,y)\right).$$

By definition of the set $\mathcal{J}_\pm$ we obtain $\rho^{-\mathrm{pr}}(t_i,x,y) = q^{\mathrm{pr}}(t_i,x)$ and $\rho^{+\mathrm{pr}}(t_i,x,y) = q^{\mathrm{pr}}(t_i,y)$ for $(x,y) \in \mathcal{X}\times\mathcal{X}$.

Proximal Mapping of $\mathcal{I}_{\mathcal{J}_{avg}}$. Note that the original problem (5.11) does not change if we add the constraint $\rho(t_0,\cdot) = \rho_A$ and $\rho(t_N,\cdot) = \rho_B$ to the set $\mathcal{J}_{avg}$. That is, we consider the projection onto the set

$$\hat{\mathcal{J}}_{avg} = \left\{(\rho,\bar\rho) \in \mathcal{J}_{avg} \;:\; \rho(t_0,\cdot) = \rho_A,\, \rho(t_N,\cdot) = \rho_B\right\}.$$

To compute the projection we have to solve

$$\operatorname*{arg\,min}_{(\rho^{\mathrm{pr}},\bar\rho^{\mathrm{pr}})\in\hat{\mathcal{J}}_{avg}} \frac{1}{2}\sum_{i=0}^{N}\sum_{x\in\mathcal{X}} |\rho^{\mathrm{pr}}(t_i,x)-\rho(t_i,x)|^2\pi(x) + \frac{1}{2}\sum_{i=0}^{N-1}\sum_{x\in\mathcal{X}} |\bar\rho^{\mathrm{pr}}(t_i,x)-\bar\rho(t_i,x)|^2\pi(x).$$

Thus, we introduce a Lagrange multiplier $\lambda \in V^0_{n,h}$ and define the corresponding Lagrangian

$$\begin{aligned}\mathcal{L}(\rho^{\mathrm{pr}},\bar\rho^{\mathrm{pr}},\lambda) = {}& \frac{1}{2}\sum_{i=0}^{N}\sum_{x\in\mathcal{X}} |\rho^{\mathrm{pr}}(t,x)-\rho(t,x)|^2\pi(x) + \frac{1}{2}\sum_{i=0}^{N-1}\sum_{x\in\mathcal{X}} |\bar\rho^{\mathrm{pr}}(t,x)-\bar\rho(t,x)|^2\pi(x)\\ & - \sum_{i=0}^{N-1}\sum_{x\in\mathcal{X}} \lambda(t_i,x)\left(\operatorname{avg}_h \rho^{\mathrm{pr}}(t_i,x) - \bar\rho^{\mathrm{pr}}(t_i,x)\right)\pi(x).\end{aligned}$$

Because of the boundary constraints, we have for all $x \in \mathcal{X}$ that $\rho^{\mathrm{pr}}(t_0,x) = \rho_A(x)$ and $\rho^{\mathrm{pr}}(t_N,x) = \rho_B(x)$. The optimality condition in ρ^{pr} reads for all $x \in \mathcal{X}$ and for all interior time steps $i = 1,\dots,N-1$ as

$$\rho^{\mathrm{pr}}(t_i,x) = \rho(t_i,x) + \tfrac{1}{2}(\lambda(t_{i-1},x) + \lambda(t_i,x)). \tag{5.15}$$

Furthermore, the optimality condition in $\bar\rho^{\mathrm{pr}}$ implies that on each interval we have

$$\bar\rho^{\mathrm{pr}}(t_i,x) = \bar\rho(t_i,x) - \lambda(t_i,x). \tag{5.16}$$

Combining both with the constraint $\operatorname{avg}_h \rho^{\mathrm{pr}}(t_i,x) = \bar\rho^{\mathrm{pr}}(t_i,x)$, we obtain

$$\bar\rho(t_i,x) - \lambda(t_i,x) = \bar\rho^{\mathrm{pr}}(t_i,x) = \operatorname{avg}_h \rho^{\mathrm{pr}}(t_i,x) = \operatorname{avg}\rho(t_i,x) + \tfrac{1}{4}(\lambda(t_{i-1},x) + 2\lambda(t_i,x) + \lambda(t_{i+1},x))$$

for all interior elements ($i = 1,\dots,N-2$) and for all $x \in \mathcal{X}$. Analogously, using the boundary conditions, we get

$$\begin{aligned}\bar\rho(t_0,x) - \lambda(t_0,x) &= \tfrac{1}{2}(\rho_A(x) + \rho(t_1,x)) + \tfrac{1}{4}(\lambda(t_0,x) + \lambda(t_1,x)),\\ \bar\rho(t_{N-1},x) - \lambda(t_{N-1},x) &= \tfrac{1}{2}(\rho_B(x) + \rho(t_{N-1},x)) + \tfrac{1}{4}(\lambda(t_{N-2},x) + \lambda(t_{N-1},x)).\end{aligned}$$

Thus, for each $x \in \mathcal{X}$ the Lagrange multiplier λ satisfies the linear system of equations

$$\begin{aligned}&\tfrac{1}{4}(5\lambda(t_0,x) + \lambda(t_1,x)) = \bar\rho(t_0,x) - \tfrac{1}{2}(\rho_A(x) + \rho(t_1,x)),\\ &\tfrac{1}{4}(\lambda(t_{i-1},x) + 6\lambda(t_i,x) + \lambda(t_{i+1},x)) = \bar\rho(t_i,x) - \tfrac{1}{2}(\rho(t_{i+1},x) + \rho(t_i,x)) \quad \forall i = 1,\dots,N-2,\\ &\tfrac{1}{4}(\lambda(t_{N-2},x) + 5\lambda(t_{N-1},x)) = \bar\rho(t_{N-1},x) - \tfrac{1}{2}(\rho_B(x) + \rho(t_{N-1},x)).\end{aligned}$$

This system is solvable, since the corresponding matrix with diagonal $(5,6,\dots,6,5)$ and off-diagonal 1 is strictly diagonal dominant. Then, given the Lagrange multiplier λ, the solution to the projection problem is given by (5.15) and (5.16). Thus, to compute the proximal mapping of $\mathcal{I}_{\hat{\mathcal{J}}_{avg}}$ we must solve a sparse system in time for each node separately. Since the involved matrix is constant, it can be pre-factored.

Proximal Mapping of $\mathcal{I}_{\mathcal{J}_=}$. Finally, the proximal map of $\mathcal{I}_{\mathcal{J}_=}$ is simply given by the projection

$$\operatorname{proj}_{\mathcal{J}_=}(\bar{\rho}, q) = \underset{(\bar{\rho}^{\mathrm{pr}}, q^{\mathrm{pr}}) \in V^0_{n,h} \times V^0_{n,h} \,:\, \bar{\rho}^{\mathrm{pr}} = q^{\mathrm{pr}}}{\arg\min} \quad \frac{h}{2} \sum_{i=0}^{N-1} \sum_{x \in \mathcal{X}} \left(|\bar{\rho} - \bar{\rho}^{\mathrm{pr}}|^2 + |q - q^{\mathrm{pr}}|^2 \right) \pi(x) = \frac{1}{2}(\bar{\rho} + q, \bar{\rho} + q)\,.$$

5.6 Numerical Results for Optimal Transport Geodesics on Graphs

In the following, we show our numerical results obtained by the optimization scheme in Section 5.5. First, we consider in Section 5.6.1 a two-node graph where the exact solution is explicitly known by solving a first-order ordinary differential equation. Computing the ODE with an Euler scheme, we can compare our numerically computed discrete optimal transport geodesic with the exact one. Then we investigate simple graphs with a small number of nodes in Section 5.6.2. Note that even in the case of a graph with three nodes, so far, there is no explicit expression of the solution. Next, we verify in Section 5.6.4 the Gromov–Hausdorff convergence to the classical L^2-optimal transport distance. Finally, we apply in Section 5.6.5 our solution scheme to larger graphs. As the stopping criteria for the iterative algorithm in (3.11), we consider the L^2-error of the mass variable $\int_0^1 \|\rho^{k+1} - \rho^k\|_\pi^2 \, \mathrm{d}t < \varepsilon$ with threshold $\varepsilon = 10^{-10}$, where k denotes the iteration step.

5.6.1 Comparison with the Exact Solution for the Two-Node Graph

We consider a graph $\mathcal{X} = \{a, b\}$ with two nodes a, b, where for $p, q \in (0, 1]$ the Markov chain and stationary distribution are given by

$$Q = \begin{pmatrix} 0 & p \\ q & 0 \end{pmatrix}, \quad \pi = \begin{pmatrix} \frac{q}{p+q} \\ \frac{p}{p+q} 1 \end{pmatrix}.$$

In [Maa11], an explicit solution trajectory for the optimal transport problem was constructed for temporal boundary data

$$\rho_A = \left(\frac{p+q}{q}, 0 \right), \quad \text{and} \quad \rho_B = \left(0, \frac{p+q}{p} \right).$$

Note that every probability measure on $\mathcal{X}$ can be described by a single parameter $r \in [-1, 1]$ via

$$\rho(r) = (\rho_a(r), \rho_b(r)) := \left(\frac{p+q}{q} \frac{1-r}{2}, \frac{p+q}{p} \frac{1+r}{2} \right).$$

In particular, we have $\rho_A = \rho(-1)$ and $\rho_B = \rho(1)$. Using this representation, it was shown that for $-1 \leqslant \alpha \leqslant \beta \leqslant 1$ the optimal transport distance is given by

$$\mathcal{W}_G(\rho(\alpha), \rho(\beta)) = \frac{1}{2} \sqrt{\frac{1}{p} + \frac{1}{q}} \int_\alpha^\beta \frac{1}{\sqrt{\theta(\rho_a(r), \rho_b(r))}} \, \mathrm{d}r, \tag{5.17}$$

and the optimal transport geodesic from $\rho(\alpha)$ to $\rho(\beta)$ is given by $\rho(\gamma(t))$ for $t \in [0, 1]$, where γ satisfies the differential equation

$$\gamma'(t) = 2(\beta - \alpha) \mathcal{W}_G\left(\rho(\alpha), \rho(\beta)\right) \sqrt{\frac{pq}{p+q} \theta(\rho_a(\gamma(t)), \rho_b(\gamma(t)))}\,. \tag{5.18}$$

In the special case, where θ is the logarithmic mean $\theta_{\log}$ and $p = q$, we obtain that

$$\theta_{\log}(\rho_a(r), \rho_b(r)) = \frac{r}{\operatorname{arctanh}(r)},$$

and consequently, the discrete transport distance is given by

$$\mathcal{W}_G(\rho(\alpha), \rho(\beta)) = \frac{1}{\sqrt{2p}} \int_\alpha^\beta \sqrt{\frac{\operatorname{arctanh}(r)}{r}}\, \mathrm{d}r\,.$$

Furthermore, the optimal transport geodesic between $\rho(\alpha)$ and $\rho(\beta)$ is given by $\rho(\gamma(t))$ for $t \in [0, 1]$, where γ satisfies the differential equation

$$\gamma'(t) = \sqrt{2p}(\beta - \alpha)\mathcal{W}_G(\rho(\alpha), \rho(\beta)) \sqrt{\frac{\gamma(t)}{\operatorname{arctanh}(\gamma(t))}}\,.$$

For this two-node graph, we numerically compute the optimal transport geodesic. This allows evaluating the distance $\mathcal{W}_G$ directly, which we can compare with a numerical quadrature of (5.17). Using the approximation of $\mathcal{W}_G$, we use an explicit Euler scheme to compute the solution ρ_h^{ODE} of the ODE (5.18). In Figure 5.4, for the case $p = q = 1$, we compare our numerical solution to the approximation of the ODE with an implicit Euler scheme for $N = 2000$.

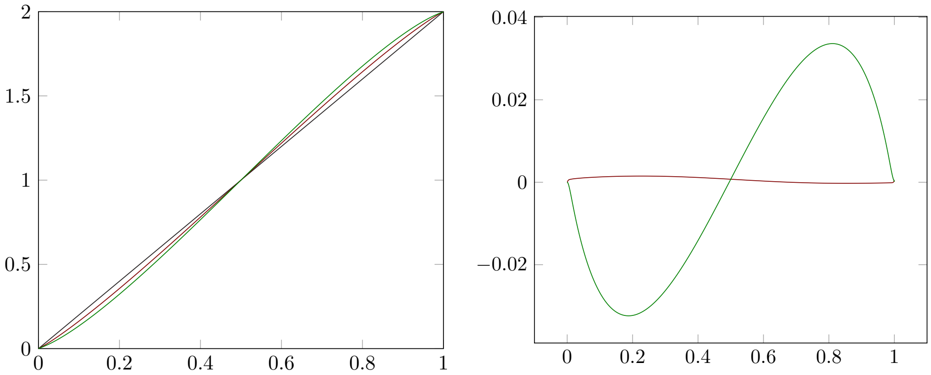


Figure 5.4: The mass distribution at b is plotted over the time interval $[0, 1]$. Left: Numerical solution for a 2-node graph $\mathcal{X} = \{a, b\}$ for the logarithmic (red) and geometric (green). The black line represents the diagonal, which is the solution in the case of the (inadmissible) arithmetic averaging. Right: Difference of the numerical solution for the logarithmic (red) and geometric (green) mean with the Euler scheme solution ρ_h^{ODE} for the logarithmic mean.

5.6.2 Exploring the Diffuse Behavior on Simple Graphs

In the following, we study the behavior of the discrete optimal transport distance on some simple graphs. Usually, we set the stationary distribution and the Markov kernel to

$$\pi(x) = \frac{d(x)}{|E|}\,, \quad Q(x, y) = \frac{1}{\pi(x)|E|}\,, \tag{5.19}$$

where for each node x we denote by $d(x)$ the number of outgoing edges. Here, we choose a time step size of $h = \frac{1}{100}$ and display the solution (ρ, m) at intermediate time steps indicated on the arrow in the first row. The mass variable $\rho(t, x)$ is represented by blue discs with an area proportional to $\rho(t, x)\pi(x)$. For the momentum variable $m(t, x, y)$, we use red arrows with a thickness proportional to $|m(t, x, y)|Q(x, y)\pi(x)$, where the direction of the arrow indicates the direction of the flow, *i.e.*, it points from x to y if $m(y, x) = -m(x, y) > 0$ (*cf.* Lemma 5.2.2).

Circles with Three and Four Nodes. First, we consider in Figure 5.5 numerically computed geodesic paths on circles with three and four nodes, where the initial mass ρ_A is supported on a single node x, and the mass ρ_B is supported on a single neighboring node of x. We observe that in the case of three nodes, a small amount of mass is also transport along the longer path, whereas for a circle with four nodes all mass is transported along the shortest connecting path.

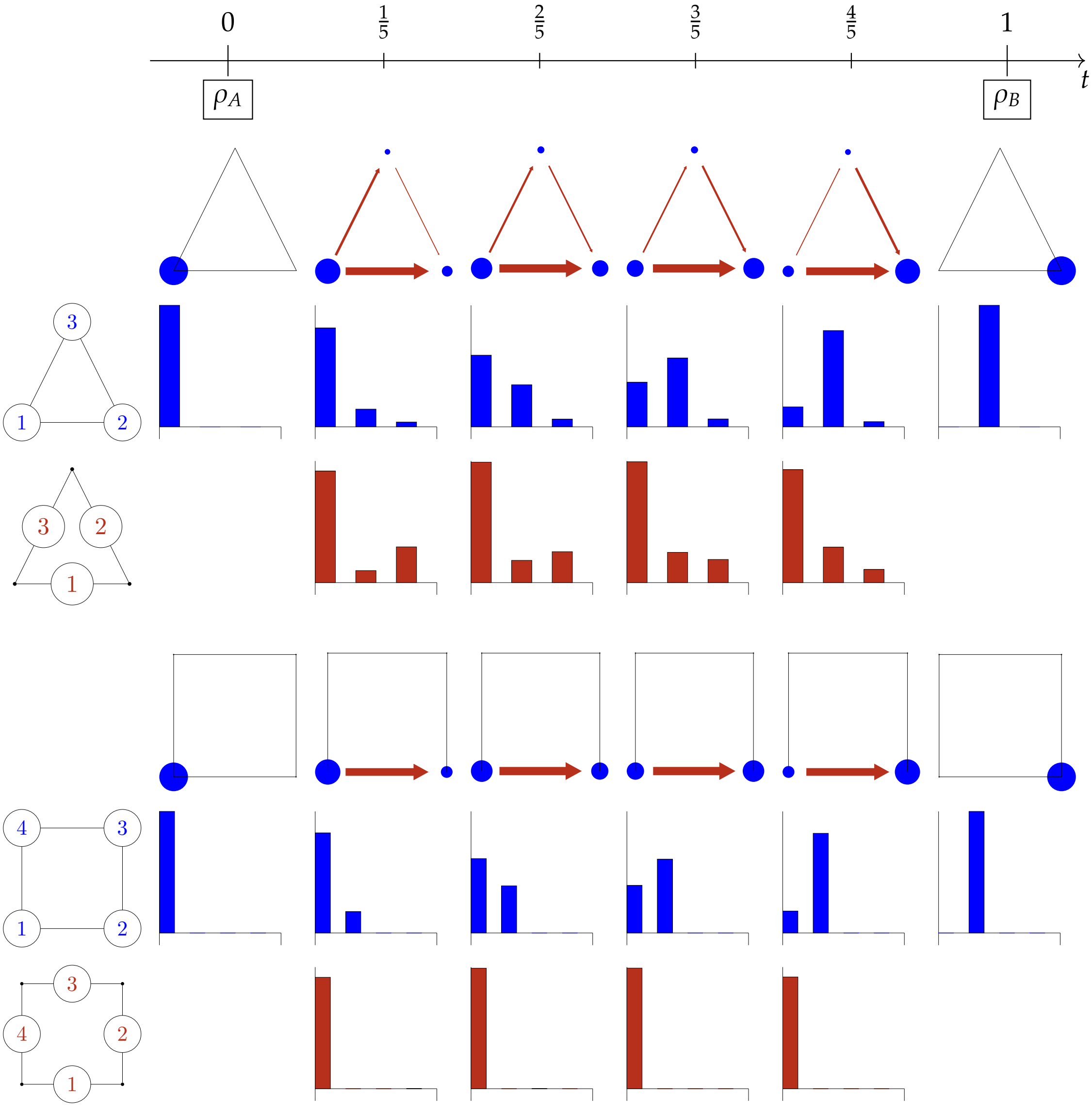


Figure 5.5: Numerically computed geodesics on circles with three and four nodes and corresponding histograms of the mass (blue) and the momentum (red) variable.

Lattice. Next, we investigate in Figure 5.6 the diffuse behavior on a 3 × 3-lattice by transporting mass from the middle left to the middle right. Since the 3 × 3-lattice consists of subgraphs given by circles with four nodes, we would expect that also on the lattice mass is only transported along the shortest path. Indeed, in [EMW19], it was established that a so-called retraction property on a subgraph is sufficient to guarantee that geodesics are supported on this subgraph. For $Q = 1$ and $\pi = 1$, the retraction property can be verified for the middle horizontal line, and thus, mass is only transported along this shortest path. However, for Q and π chosen as in (5.19) the retraction property does not hold and our numerical results show that an essential amount of mass is not transported along the middle horizontal line.

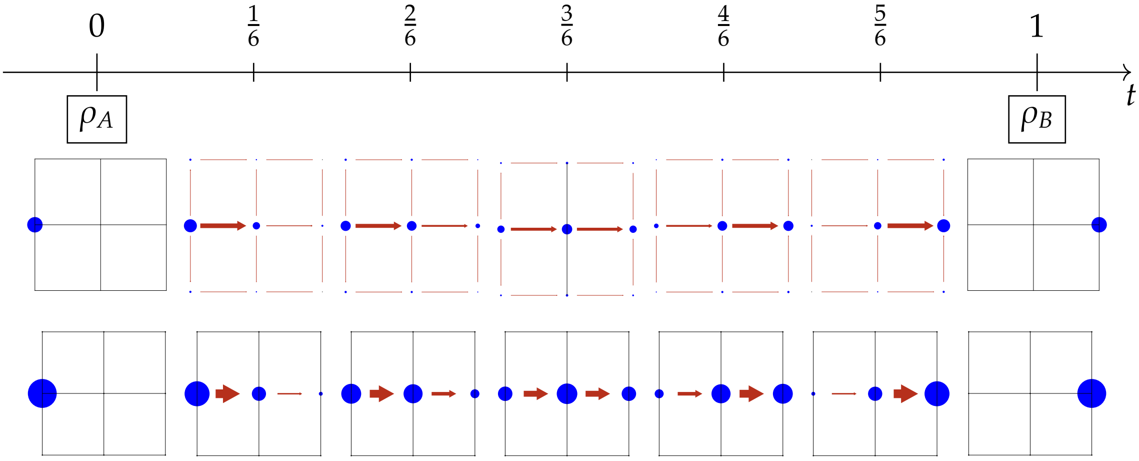


Figure 5.6: Numerically computed geodesics on a 3 × 3-lattice for different choices of Q and π. Top: Markov kernel Q and stationary distribution π given as in (5.19). Bottom: $Q(x, y) = 1$ for all edges (x, y) and $\pi(x) = 1$ for all nodes x.

Cube and Hypercube. In Figure 5.7, we consider the cube $\{0, 1\}^3$ and the hypercube $\{0, 1\}^4$. Note that the computed solutions are symmetric in the sense that $\rho(t, x) = \rho(1 - t, x)$ and $m(t, x, y) = m(1 - t, x, y)$ for all $t \in (0, 1)$. Furthermore, the distribution of mass is constant on all nodes at time $t = \frac{1}{2}$.

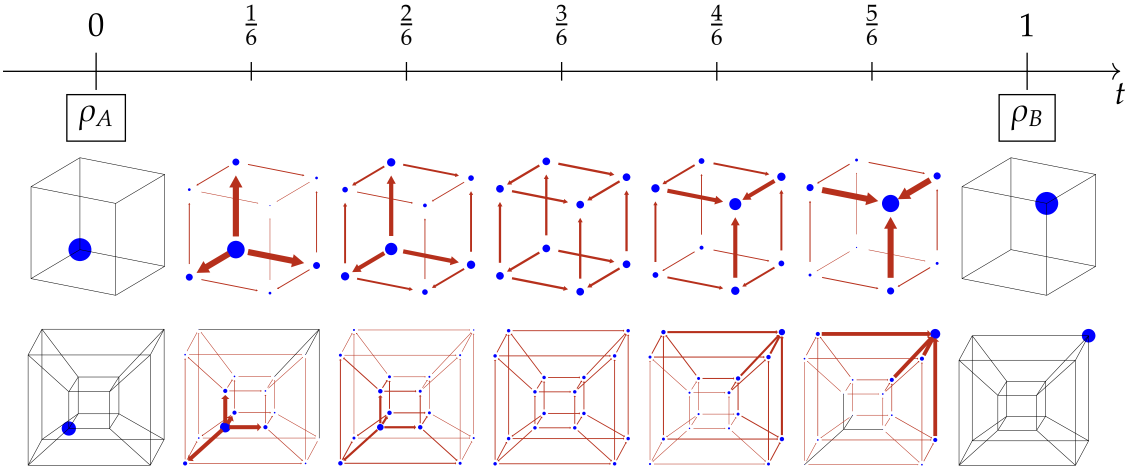


Figure 5.7: Numerically computed geodesic on the cube (top) and the hypercube (bottom). We observe an equidistribution of the mass at time $t = \frac{1}{2}$.

Convexity of the Entropy Functional. In Figure 5.8, for the cube, we verify that the entropy functional

$$\mathcal{H}(\rho) = \sum_{x \in \mathcal{X}} \rho(x) \log(\rho(x)) \pi(x)$$

is convex along discrete optimal transport geodesics. This result was proven in [Maa11, Proposition 2.12].

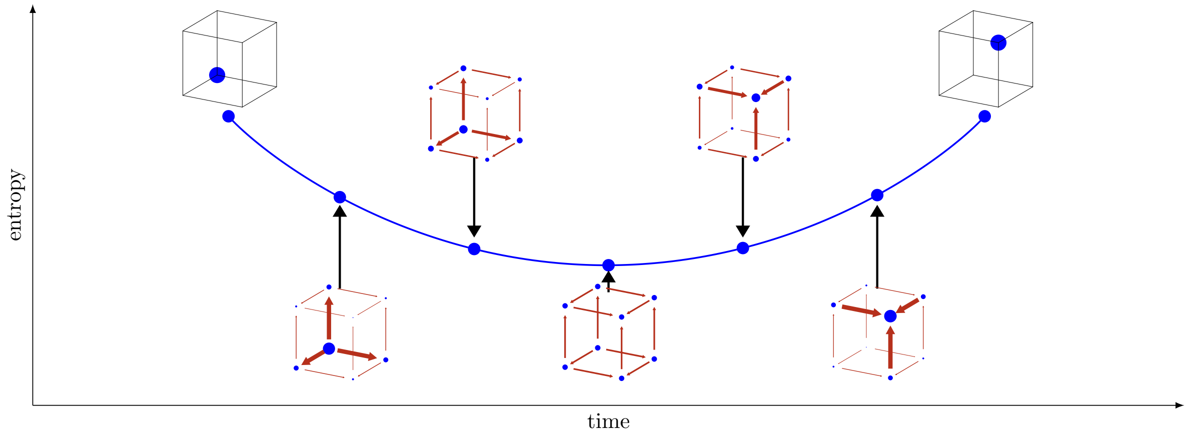


Figure 5.8: Entropy functional is convex along a discrete optimal transport geodesic.

Change of Sign for Momentum Variable. Finally, in Figure 5.9, we depict an example of a graph with four nodes, which shows that the sign of the momentum variable on a fixed edge may change along a geodesic path.

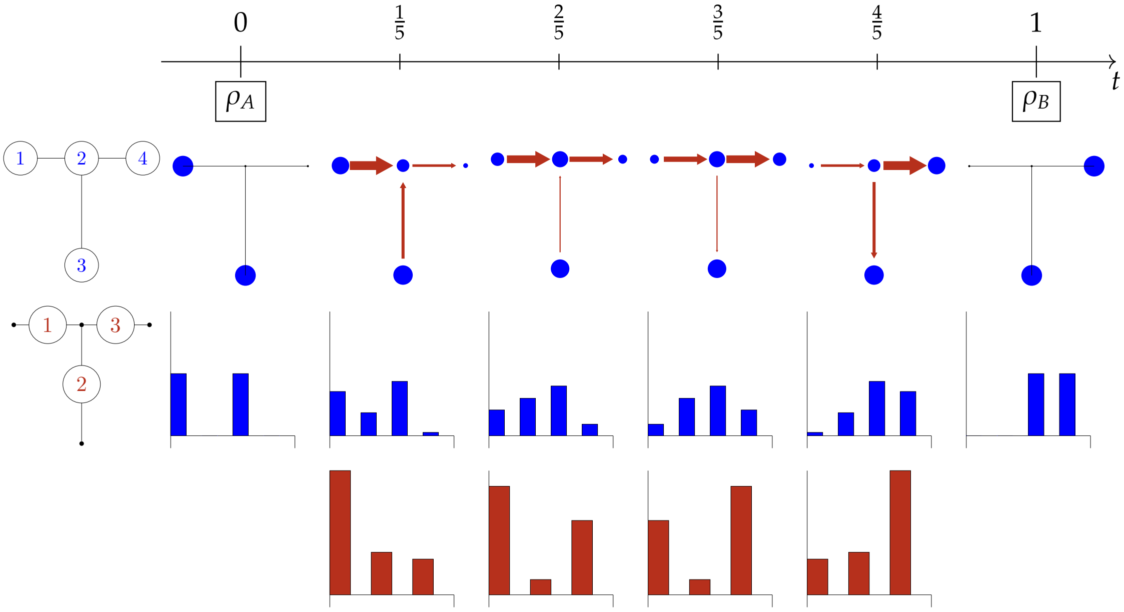


Figure 5.9: Numerically computed geodesic on a graph with four nodes. Note that the sign of the momentum variable m for the edge with index 2 changes ($cf.$ $t = \frac{1}{5}$ and $t = \frac{4}{5}$).

5.6.3 Experimental Convergence Rate in Time

In Theorem 5.4.5, we have established the convergence of minimizing paths of our fully discrete approximation for a time step size $h \to 0$. Here, we study this convergence numerically. We take into account a square lattice with 3×3 nodes and compute an optimal transport path between the mass concentrated at the midpoint of the square and uniform distribution, *i.e.*,

$$\rho_A = \delta_{\left(\frac{1}{2},\frac{1}{2}\right)}, \quad \text{and} \quad \rho_B = \mathbb{1}\,.$$

For the discretization in time we choose $N = 8, 16, 32, 64, 128, 256, 512, 1024$. Since the exact solution for this example is unknown, we consider our computational result $(\rho_{\text{approx}}, m_{\text{approx}})$ for the finest discretization $N = 1024$ as an approximation. Then we compare the solutions (ρ_h, m_h) with this approximation in the corresponding norms for which we have shown convergence, *i.e.*, we consider

$$\|\rho_{\text{approx}} - \rho_h\|_{W^{1,2}([0,1],\mathbb{R}^{\mathcal{X}})}, \quad \text{and} \quad \|m_{\text{approx}} - m_h\|_{L^2([0,1],\mathbb{R}^{\mathcal{X}})}\,.$$

In Figure 5.10, we plot these errors in a log-log scale and experimentally obtain linear convergence in h.

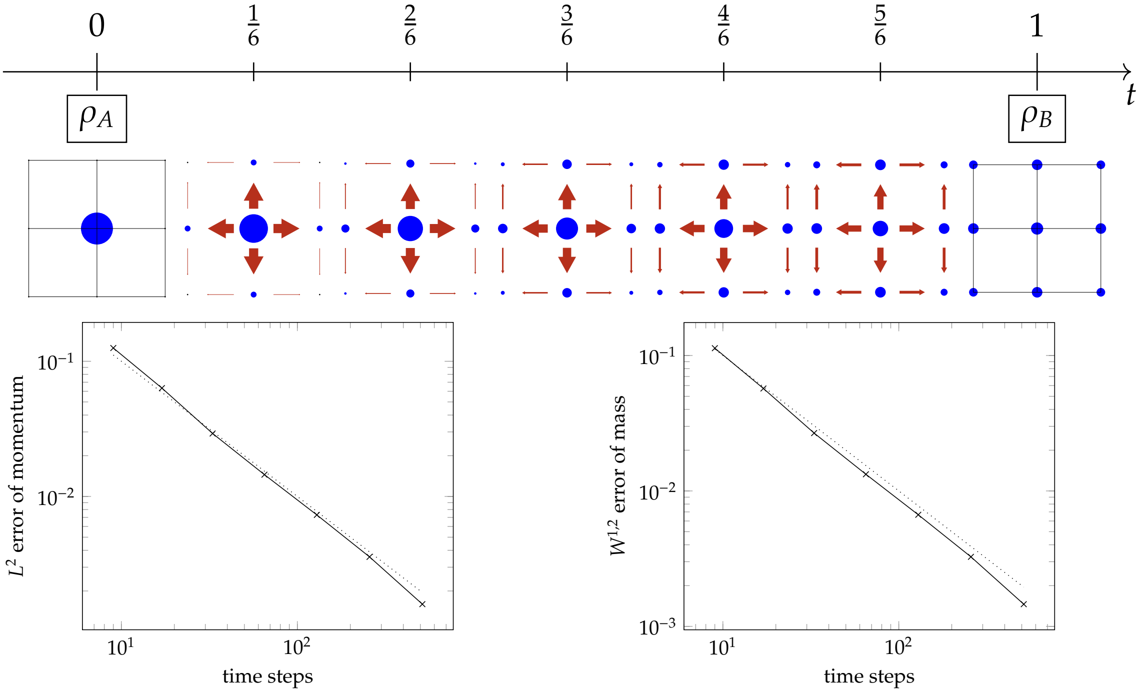


Figure 5.10: Numerical verification of the convergence in time on a square lattice with 3×3 nodes for a time discretization with $N = 8, 16, 32, 64, 128, 256, 512, 1024$. Below we plot the errors $\|m_{\text{approx}} - m_h\|_{L^2([0,1],\mathbb{R}^{\mathcal{X}\times\mathcal{X}})}$ and $\|\rho_{\text{approx}} - \rho_h\|_{W^{1,2}([0,1],\mathbb{R}^{\mathcal{X}})}$ in a log-log scale. The convergence order is linear in h (dotted lines).

5.6.4 Experimental Results Related to the Gromov–Hausdorff Convergence in Space

In [GM13], it was shown that for the d-dimensional torus $\boldsymbol{T}^d$ the discrete transport distance $\mathcal{W}_G$ on a discretized torus $\boldsymbol{T}^d_M$ with uniform mesh size $\frac{1}{M}$ converges in the Gromov–Hausdorff metric to the classical L^2-Wasserstein distance on $\boldsymbol{T}^d$. This result was extended in [GKM18] to a certain class of regular meshes via a finite volume scheme, but also counterexamples have been found.

Note that for the classical L^2-Wasserstein distance, the optimal transport geodesic connecting two point masses $\rho_A = \delta_0$ and $\rho_B = \delta_1$ is given by the transport of the Dirac measure with constant speed:

$$\rho(t,x) = \delta_t(x)\,.$$

Gromov–Hausdorff Convergence on a Line. First, for $d = 1$, we consider the unit interval $I = [0,1]$ and a sequence of space discretizations

$$\mathcal{X}_M = \{x_0, \dots, x_M\}$$

with uniform mesh size $\frac{1}{M}$ for $M \in \mathbb{N}_+$. The corresponding Markov kernel Q_M for $\mathcal{X}_M$ is defined by

$$\begin{cases} Q_M(x_i, x_{i+1}) = Q_M(x_i, x_{i-1}) = \dfrac{1}{2} \quad \text{for } i = 1, \dots, x_{M-1}\,, Q_M(x_0, x_1) = Q_M(x_M, x_{M-1}) = 1\,. \end{cases}$$

For this sequence of graphs, we compute discrete optimal transport geodesics between δ_0 and δ_1. In Figure 5.11, we plot the density distribution of the discrete optimal transport geodesic at time $t = \frac{1}{2}$ for different grid sizes $\frac{1}{M}$. According to the Gromov–Hausdorff convergence result, we expect an increasing mass concentration at $x = \frac{1}{2}$ for $M \to \infty$, which we can indeed observe.

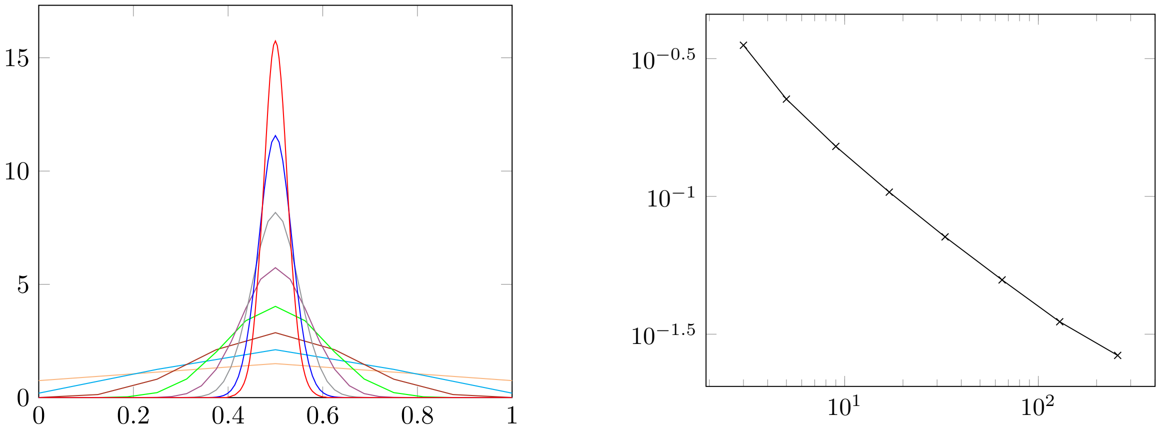


Figure 5.11: Left: Linearly interpolated densities for the $\mathcal{W}_G$ geodesic on a one-dimensional chain graph between a Dirac mass at the beginning and the end, at $t = 0.5$ with $M = 2$ (yellow), 4 (turquoise), 8 (brown), 16 (green), 32 (violet), 64 (gray), 128 (blue), and 256 (red). Right: Convergence of the L^2-Wasserstein distance to a Dirac measure at $x = \frac{1}{2}$ for $M \to \infty$.

To quantify the convergence rate experimentally, we compute the L^2-Wasserstein distance of the approximatively computed discrete geodesic at time $t = \frac{1}{2}$ to the Dirac measure $\delta_{\frac{1}{2}}$, which is explicitly given by

$$\mathcal{W}\left(\rho_M\left(\tfrac{1}{2}\right), \delta_{\frac{1}{2}}\right)^2 = \left(\sum_{m=0}^{M} \left|\tfrac{m}{M} - \tfrac{1}{2}\right|^2 \rho_M\left(\tfrac{1}{2}, \tfrac{m}{M}\right)\right)\,.$$

In Table 5.1, we compute the expected order of convergence. As an initialization of the variables in the proximal splitting algorithm we use an adaptive scheme in time, *i.e.*, we first compute a solution for $N = 32$, then prolongate this result to a finer discretization by doubling N and repeat until $N = 1024$. Remember that we have linear convergence of the mass and momentum variables in the appropriate norms for $N \to \infty$. However, for $M = 256$, the error in time seems to be quite large for $N = 1024$, s.t. the associated expected order of convergence is inaccurate. Moreover, it turns out that transporting a Dirac measure with the discrete optimal transport distance is quite singular, since we frequently have to deal with the unstable case in the projection of the logarithmic cone K at 0. For all other results up to a space discretization with $M = 128$, the difference of the discrete optimal transport distance for $N = 1024$ to $N = 512$ is small, s.t. we expect from our numerical results a convergence order $\frac{1}{2}$.

M	$\mathcal{W}_M := \mathcal{W}\left(\rho_M\left(\frac{1}{2}\right), \delta_{\frac{1}{2}}\right)$	EOC $= \frac{\log(\mathcal{W}_M/\mathcal{W}_{M/2})}{\log(1/2)}$
2	0.3535819979	
4	0.225432747	0.649347723
8	0.1518529497	0.5700221686
16	0.1036187107	0.5513903948
32	0.071299038	0.5393300146
64	0.0497631136	0.5188058612
128	0.0350786465	0.5044836728
256	0.0264826129	0.4055476142

Table 5.1: Expected order of convergence of the discrete geodesics from Figure 5.11 in the L^2-Wasserstein distance.

Gromov–Hausdorff Convergence on a Square. Next, for $d = 2$, we consider a square lattice of uniform grid size $\frac{1}{M}$ with $M \in \mathbb{N}$ and nodes $\mathcal{X}_M = \{(i/M, j/M) \ : \ i, j \in (0, \dots, M)\}$, where the weights of the Markov kernel Q are proportional to the number of adjacent edges. We compute optimal transport geodesic connecting the Dirac masses $\delta_{(0,0)}$ and $\delta_{(1,1)}$. An example for $M = 2$ is depicted in Figure 5.12.

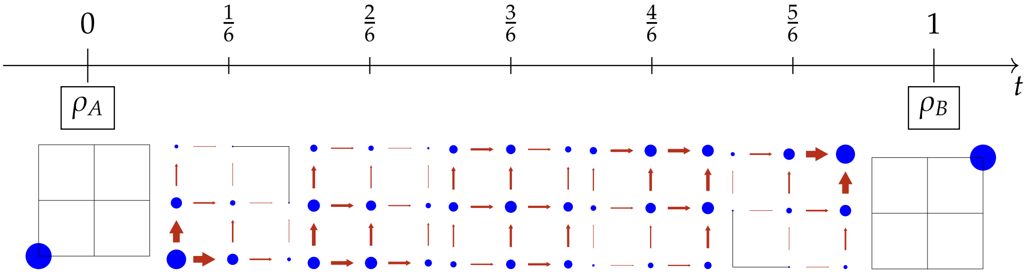


Figure 5.12: Numerically computed geodesics on a 3 × 3-lattice connecting $\rho_A = \delta_{(0,0)}$ and $\rho_B = \delta_{(1,1)}$.

For increasing M, we expect an increasing mass concentration on the space diagonal. In Figure 5.13, for decreasing mesh size $\frac{1}{M}$, we plot the in time accumulated density values along the diagonal and the off-diagonals of nodes. More precisely, we define the bands of nodes

$$B_M^k = \left\{(x_1, x_2) \in \mathcal{X}_M \times \mathcal{X}_M \ : \ x_2 = x_1 + \frac{k}{M}\right\}$$

for $k = -M, \dots, M$, where B_M^0 represents the diagonal. Then we compare the values $\int_0^1 \sum_{x \in B_M^k} \rho(t, x)\pi(x) \, \mathrm{d}t$.

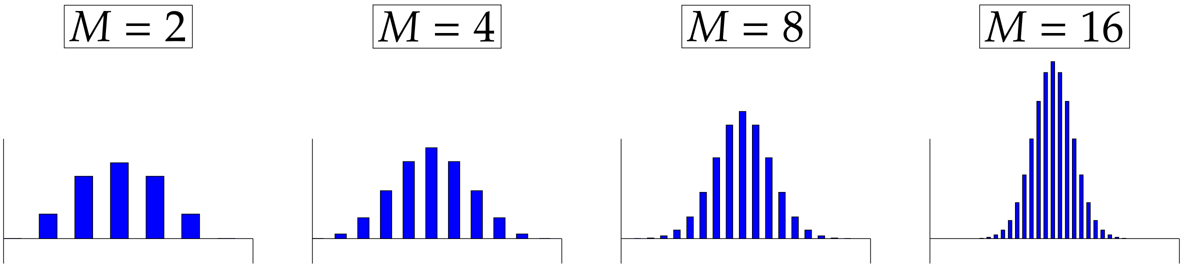


Figure 5.13: Geodesics in the distance $\mathcal{W}_G$ on a two-dimensional grid graph between Dirac masses at diagonally opposite ends. We show accumulated densities along the diagonal and the off-diagonals (see text for details). The width of the bars is scaled with the number of lines.

5.6.5 Discrete Geodesics on Triangular Meshes of a Human Hand

In Figure 5.14, we take into account a triangular mesh of a human hand and compute the discrete optimal transport geodesic between a mass ρ_A supported on the fingers and a mass ρ_B supported on the wrist. This example is not intended as a real application, but it demonstrates that our numerical algorithm can be performed on large and complex graphs.

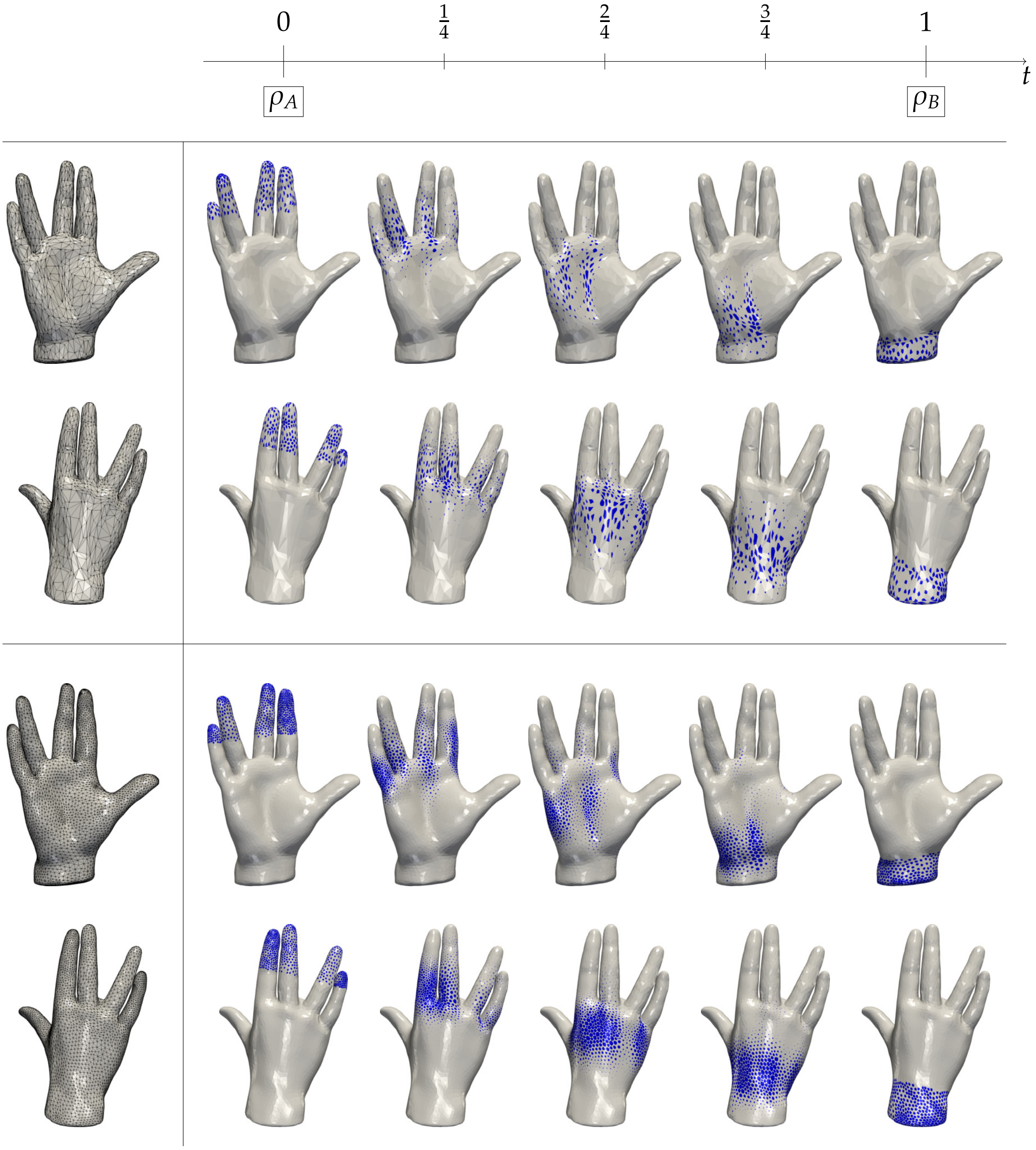


Figure 5.14: Extraction of a discrete optimal transport geodesics for two different triangular meshes (shown on the left) of a human hand. We depict all results from two different view positions. For each node, we represent the mass by a blue neighborhood with an area of a proportional size. Top: The mesh has 1828 nodes and the geodesic is computed for 257 time steps. We plot the result at the time steps $t = 0, 65, 129, 193, 257$. Bottom: The mesh has 6094 nodes and the geodesic is computed for 33 time steps. We plot the result at the time steps $t = 0, 9, 17, 25, 33$.

5.7 Simulation of the Gradient Flow of the Entropy

For the L^2-Wasserstein distance, in the seminal work [JKO98], it was shown that the heat equation can be interpreted as a gradient flow trajectory of the entropy functional (*cf*. Section 3.1.4). Now, for the discrete optimal transport distance, in [Maa11], an analogous result was provided. Indeed, for the logarithmic mean $\theta_{\log}$ as an averaging function, the heat flow is a gradient flow trajectory for the entropy w.r.t. the metric $\mathcal{W}_G$. Here, the entropy functional $\mathcal{H}\colon \mathscr{P}(\mathcal{X}) \to \mathbb{R}$ is given by

$$\mathcal{H}(\rho) = \sum_{x\in\mathcal{X}} \rho(x)\log(\rho(x))\pi(x)\,,$$

with the convention '$0\log 0 = 0$'. Moreover, in [EM14], a similar result was shown for the Renyi entropy

$$\mathcal{H}_s(\rho) = \frac{1}{s-1}\sum_{x\in\mathcal{X}} \rho(x)^s\pi(x)\,.$$

By choosing, *e.g.*, $s = \frac{1}{2}$ and taking into account the geometric mean θ_{geo} as averaging function, the gradient flow of $\mathcal{H}_s$ w.r.t. the metric $\mathcal{W}_G$ is given by the porous medium equation $\partial_t\rho = \Delta_{\mathcal{X}}\rho^s$.

5.7.1 Adaption of the Numerical Scheme

In the following, we verify numerically that gradient flow trajectories coincide with solutions to the corresponding partial differential equations. Therefore, we make use of the minimizing movement scheme (3.7). Given an initial density $\rho_0 \in \mathscr{P}(\mathcal{X})$ and a time step size $\tau > 0$, an implicit time discrete gradient flow scheme for $\mathcal{H}$ is defined by the iteration

$$\rho_{k+1} = \operatorname*{arg\,min}_{\rho_B\in\mathscr{P}(\mathcal{X})} \frac{1}{2}\mathcal{W}_{G,h}(\rho_k,\rho_B)^2 + \tau\,\mathcal{H}(\rho_B)\,. \tag{5.20}$$

Note that for our fully discrete optimal transport distance $\mathcal{W}_{G,h}$, the time step size h appears as an inner discretization parameter.

For a fully numerical scheme, to compute a solution of the minimizing movement step in (5.20), we essentially have to make two modifications compared to our method for fully discrete geodesic paths. First, we define a discrete continuity equation with a free endpoint. For initial value $\rho_A \in \mathscr{P}(\mathcal{X})$, let

$$\mathcal{CE}_h(\rho_A) = \left\{(\rho_h, m_h, \rho_B) \in V^1_{n,h} \times V^0_{e,h} \times \mathbb{R}^{\mathcal{X}} \;:\; (\rho_h, m_h) \in \mathcal{CE}_h(\rho_A,\rho_B)\right\}\,.$$

Furthermore, we take into account the entropy of this free endpoint, which additionally appears as a further variable. Then, analogously to (5.11), the minimization problem (5.20) can be written as

$$\begin{aligned}\min\,\{\mathcal{F}(\rho_h, m_h, \vartheta_h, \rho^-_h, \rho^+_h, \bar\rho_h, q_h, \rho_B) + \mathcal{G}(\rho_h, m_h, \vartheta_h, \rho^-_h, \rho^+_h, \bar\rho_h, q_h, \rho_B) \;:\;& \\ (\rho_h, m_h, \vartheta_h, \rho^-_h, \rho^+_h, \bar\rho_h, q_h, \rho_B) \in V^1_{n,h} \times (V^0_{e,h})^4 \times (V^0_{n,h})^2 \times \mathbb{R}^{\mathcal{X}}\}&\end{aligned}$$

with

$$\begin{aligned}\mathcal{F}(\rho_h, m_h, \vartheta_h, \rho^-_h, \rho^+_h, \bar\rho_h, q_h, \rho_B) :=& \mathcal{E}_{\text{trans,e}}(\vartheta_h, m_h) + \mathcal{I}_{\mathcal{J}_\pm}(q_h, \rho^-_h, \rho^+_h) + \mathcal{I}_{\mathcal{J}_{avg}}(\rho_h, \bar\rho_h) + 2\tau\cdot\mathcal{H}(\rho_B)\,,\\ \mathcal{G}(\rho_h, m_h, \vartheta_h, \rho^-_h, \rho^+_h, \bar\rho_h, q_h, \rho_B) :=& \mathcal{I}_{\mathcal{CE}_h(\rho_k)}(\rho_h, m_h, \rho_B) + \mathcal{I}_{\mathcal{K}}(\rho^-_h, \rho^+_h, \vartheta_h) + \mathcal{I}_{\mathcal{J}_=}(\bar\rho_h, q_h)\,.\end{aligned}$$

As for the geodesic interpolation problem, the splitting into $\mathcal{F}$ and $\mathcal{G}$ allows applying a proximal splitting algorithm. We extend the space H by a factor $\mathbb{R}^{\mathcal{X}}$, and adapt the scalar product in (5.12) by adding the term $h\langle\rho_{B,1}(\cdot), \rho_{B,2}(\cdot)\rangle_\pi$ for the additional variable ρ_B. The proximal step of $\mathcal{F}^*$ then requires to compute an additional proximal step of $(2\tau\mathcal{H})^*$. In the proximal step of $\mathcal{G}$, the projection onto $\mathcal{CE}_h(\rho_A,\rho_B)$ is replaced by a projection onto $\mathcal{CE}_h(\rho_k)$. Next, we describe these modifications in detail.

We recall that the proximal mappings of $\mathcal{H}^*$ and $\mathcal{H}$ are linked by Moreau's decomposition (see Theorem 3.2.9). Moreover, the computation of the proximal mapping of $\mathcal{H}$ can be performed on the space $\mathbb{R}^{\mathcal{X}}_{\geqslant 0}$, since the constraint $\rho_B \in \mathscr{P}(\mathcal{X})$ in the formulation of $\mathcal{H}$ is enforced via the mass-preserving condition in the discrete continuity equation. Then the computation of the proximal mapping of $\mathcal{H}$ decouples in space, and the resulting one-dimensional problem can be solved via Newton's method.

To project onto the set $\mathcal{CE}_h(\rho_A)$ of solutions to the continuity equation with free endpoint, in analogy to Proposition 5.5.3, a discrete elliptic equation on the time-space domain has to be solved.

Proposition 5.7.1 (Projection onto $\mathcal{CE}_h(\rho_A)$)**.** *The projection*

$$\begin{aligned}\operatorname{proj}_{\mathcal{CE}_h(\rho_A)}(\rho, m, \rho_B) = \operatorname*{arg\,min}_{(\rho^{pr}, m^{pr}, \rho_B^{pr}) \in \mathcal{CE}_h(\rho_A)} \; & \frac{h}{2} \sum_{i=0}^{N} \|\rho^{pr}(t_i, \cdot) - \rho_h(t_i, \cdot)\|_\pi^2 \\ & + \frac{h}{2} \sum_{i=0}^{N-1} \|m^{pr}(t_i, \cdot) - m_h(t_i, \cdot)\|_Q^2 + \frac{h}{2} \|\rho_B^{pr} - \rho_B\|_\pi^2\end{aligned} \tag{5.21}$$

onto the set $\mathcal{CE}_h(\rho_A)$ of solutions to the discrete continuity equation with initial data ρ_A can be computed by solving the following linear system in the Lagrange multiplier $\varphi_h \in V^0_{n,h}$:

$$\begin{aligned}&\frac{\varphi_h(t_1, x) - \varphi_h(t_0, x)}{h^2} + \Delta_{\mathcal{X}} \varphi_h(t_0, x) \\ &\quad = -\left(\frac{\rho_h(t_1, x) - \rho_A(x)}{h} + \operatorname{div}_{\mathcal{X}} m_h(t_0, x) \right) , \\ &\frac{-\frac{3}{2}\varphi_h(t_{N-1}, x) - \varphi_h(t_{N-2}, x)}{h^2} + \Delta_{\mathcal{X}} \varphi_h(t_{N-1}, x) \\ &\quad = -\left(\frac{\frac{1}{2}(\rho_B(x) + \rho_h(t_N, x)) - \rho_h(t_{N-1}, x)}{h} + \operatorname{div}_{\mathcal{X}} m_h(t_{N-1}, x) \right) , \\ &\frac{\varphi_h(t_{i+1}, x) - 2\varphi_h(t_i, x) + \varphi_h(t_{i-1}, x)}{h^2} + \Delta_{\mathcal{X}} \varphi_h(t_i, x) \\ &\quad = -\left(\frac{\rho_h(t_{i+1}, x) - \rho_h(t_i, x)}{h} + \operatorname{div}_{\mathcal{X}} m_h(t_i, x) \right)\end{aligned} \tag{5.22}$$

with $i = 1, \ldots, N-2$ and $x \in \mathcal{X}$. Then the solution $(\rho^{pr}, m^{pr}, \rho_B^{pr})$ to (5.21) *is given by*

$$\begin{aligned}\rho_B^{pr}(x) &= \frac{1}{2}\left(\rho_h(t_N, x) + \rho_B(x) - \frac{\varphi_h(t_{N-1}, x)}{h} \right) , \\ \rho^{pr}(t_i, x) &= \rho_h(t_i, x) + \frac{\varphi_h(t_i, x) - \varphi_h(t_{i-1}, x)}{h} , \\ \rho^{pr}(t_0, x) &= \rho_A(x) , \; \rho^{pr}(t_N, x) = \rho_B^{pr}(x) , \\ m^{pr}(t_i, x, y) &= m_h(t_i, x, y) + \nabla_{\mathcal{X}} \varphi_h(t_i, x, y)\end{aligned}$$

for all $i = 1, \ldots, N-2$ and $x, y \in \mathcal{X}$.

Proof. In analogy to the proof of Proposition 5.5.3. □

Remark 5.7.2 (Comparison to the Projection onto $\mathcal{CE}_h(\rho_A, \rho_B)$). Note that, in contrast to Lemma 5.5.4, the linear system (5.22) is no longer degenerated due to the additional freedom of ρ_B, and thus, no additional Lagrange multiplier is required.

5.7.2 Numerical Results for Gradient Flows

For our numerical computations, we choose a line of five nodes with

$$\begin{aligned}
&\text{stationary distribution} && \pi = \frac{1}{8}(1,2,2,2,1)\,, \\
&\text{Markov kernel} && Q(x,y) = \frac{1}{8\pi(x)} \text{ for } x,y \text{ adjacent, and} \\
&\text{initial mass} && \rho = \frac{1}{2}(1,1,5,1,1)\,.
\end{aligned}$$

In Figure 5.15, we compare the solution to the heat equation with our numerical result of the gradient flow of the entropy $\mathcal{H}$ and a logarithmic averaging operator. Furthermore, we compare the solution to the porous medium equation with our numerical result of the gradient flow of the entropy $\mathcal{H}_{\frac{1}{2}}$ and a geometric averaging operator. Here, the solution to the heat equation and the porous medium equation are approximated by a simple explicit Euler scheme. Note that the entropy functional $\mathcal{H}$ is minimized for equidistributed $\rho \in \mathscr{P}(\mathcal{X})$. Thus, in the example above, we experimentally obtain that the iterates ρ_k in (5.20) converge to the uniform distribution $\mathbb{1} = (1,1,1,1,1)$ for $k \to \infty$. In Figure 5.16, we plot for $3 \cdot 10^4$ minimizing movement steps the entropy functional $\mathcal{H}(\rho_k)$ and the difference to the uniform distribution $\|\rho_k - \mathbb{1}\|_2$. We observe, in both cases, an exponential decay rate.

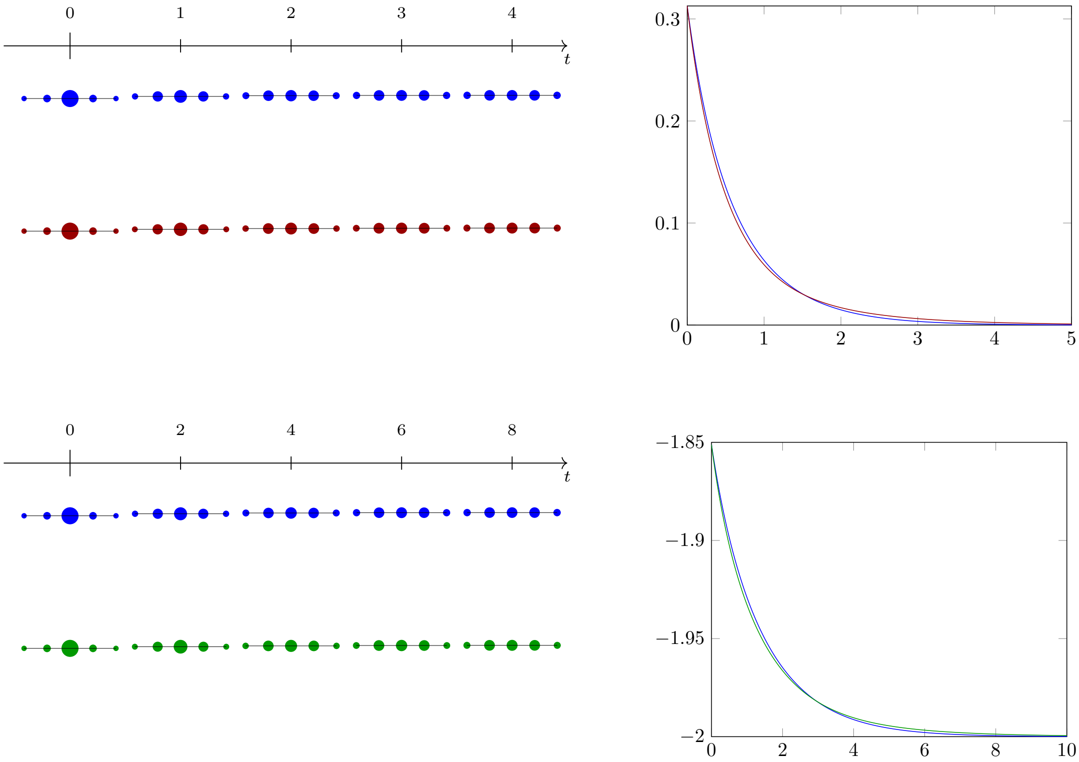


Figure 5.15: Numerical solution to the heat flow (top) and the porous medium equation (bottom) based on an explicit Euler scheme (blue) with time step size 10^{-3} and for the gradient flow of the associated entropy using the logarithmic mean (red) and the geometric mean (green), respectively, with $\tau = 10^{-3}$ and $h = 100$. Panels on the left show the mass distributions on the graph at different times, panels on the right show the values of the entropies over time.

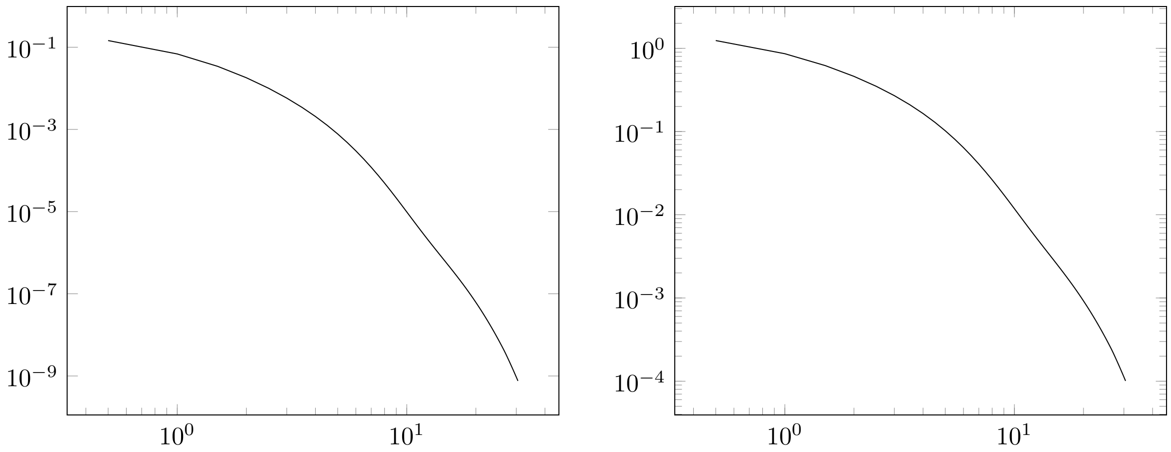


Figure 5.16: Convergence of gradient flow of the entropy using the logarithmic mean. We use a time step size $\tau = 10^{-3}$, $h = 100$ and $3 \cdot 10^4$ minimizing movement steps. Left: Entropy functional $\mathcal{H}(\rho_k)$. Right: Difference to the uniform distribution $\|\rho_k - \mathbb{1}\|_2$.

5.8 Conclusion and Outlook

We have arrived at a fully numerical scheme to approximate geodesics for discrete optimal transport introduced by Maas [Maa11]. Our finite element discretization in time has been chosen s.t. a Γ-convergence result can be established. Compared to the classical optimal transport distance, we have used a similar proximal splitting algorithm, where auxiliary slack variables have been necessary to decouple the nonlinearity given by the averaging operator in space and time, which then basically requires a projection onto a three-dimensional convex set. We have verified that our numerically computed solutions satisfy specific properties, which have been proven previously in the literature.

Concerning the Γ-convergence result, it has been essential for the Γ-liminf inequality that the set of solutions to the discretized continuity equation is contained in the set of solutions to the continuous continuity equation. This conforming approximation property is no longer valid for a discretization with both piecewise constant mass and momentum variables $(\rho, m) \in V^0_{n,h} \times V^0_{e,h}$. However, we have obtained similar results with such finite element spaces. Moreover, the Γ-limsup estimate could be obtained directly by Jensen's inequality for piecewise constant mass, since no additional Lagrange interpolation operator is required. Therefore, we have also implemented a discontinuous Galerkin discretization for piecewise constant mass and piecewise affine momentum variables $(\rho_h, m_h) \in V^0_{n,h} \times V^{1,-1}_{e,h}$ to combine the advantages for the Γ-liminf and Γ-limsup inequality. In the numerical results for the discontinuous Galerkin discretization, we have obtained oscillations of the momentum variable, which could be reduced by an additional L^2-regularizer.

Our implementation has taken into account the sparsity of the Markov kernel Q since we have considered the momentum variable $m(x, y)$ only on edges where $Q(x, y) > 0$. Moreover, in Lemma 5.2.2, we have proven that for an optimal path (ρ, m), the momentum variable is antisymmetric in the sense that $m(t, x, y) = -m(t, y, x)$ for all $t \in [0, 1]$. This additional information has not been incorporated in our discretization, but indeed, for our computational results, we have observed the antisymmetry of m. An alternative discretization was taken into account in [Sch18], where the degrees of freedom for the momentum variable were reduced by a factor two. Furthermore, instead of the variables (ρ, m), the variables $(\rho\pi, mQ\pi)$ were considered, which allows eliminating the stationary distribution π in the energy functional.

Finally, we remark that Erbar [Erb16] constructed a similar discrete transport distance, which allows identifying the spatially homogeneous Boltzmann equation as a gradient flow trajectory.

Part II

Numerical Methods for Elastic Shape Optimization

Chapter 6

Foundations in Elasticity and Shape Optimization

The second part of this thesis is concerned with several compliance shape optimization problems. In general, we consider forces acting on a reference domain of an elastic object leading to a deformed domain crucially depending on the material. Under certain mechanical assumptions, we derive partial differential equations to describe the corresponding equilibrium deformations. We start in Section 6.1 to give a short introduction to the theory of elastic bodies in $\mathbb{R}^3$. In Chapter 7, we consider a special class of elastic bodies, whose microstructure is given by a periodic cell. To describe the macroscopic behavior of such objects, we recall the theory of homogenization in Section 6.2. Later, in Chapter 8, we investigate a further class of elastic bodies, which can be described by a two-dimensional surface with a small thickness. Finally, in Section 6.3, we give an introduction to elastic shape optimization, where we ask for an optimal distribution of the material on the reference domain to obtain optimal stability w.r.t. the given forces.

6.1 Elastic Bodies

Here, we give an overview of the theory of elastic bodies in $\mathbb{R}^3$ by mainly following the famous book by Ciarlet [Cia88]. Let $\Omega_A \subset \mathbb{R}^3$ be the reference domain of an elastic body. We assume that Ω_A is a bounded, open, and connected set. A map $\Phi\colon \overline{\Omega_A} \to \mathbb{R}^3$ is called deformation if it is injective on Ω_A and orientation-preserving. We denote the corresponding deformed domain by $\Omega_B := \Phi(\Omega_A)$ and suppose Φ to be the identity on a fixed part of the boundary $\Gamma_A \subset \partial\Omega_A$. Furthermore, let a body force $F_B\colon \Omega_B \to \mathbb{R}^3$ and a surface force $G_B\colon \Gamma_B^N \to \mathbb{R}^3$ act on the deformed domain, where the free boundary is given by $\Gamma_B^N = \partial\Omega_B \backslash \Gamma_B = \partial\Omega_B \backslash \Gamma_A$. In Figure 6.1, we depict a 2D sketch of this configuration.

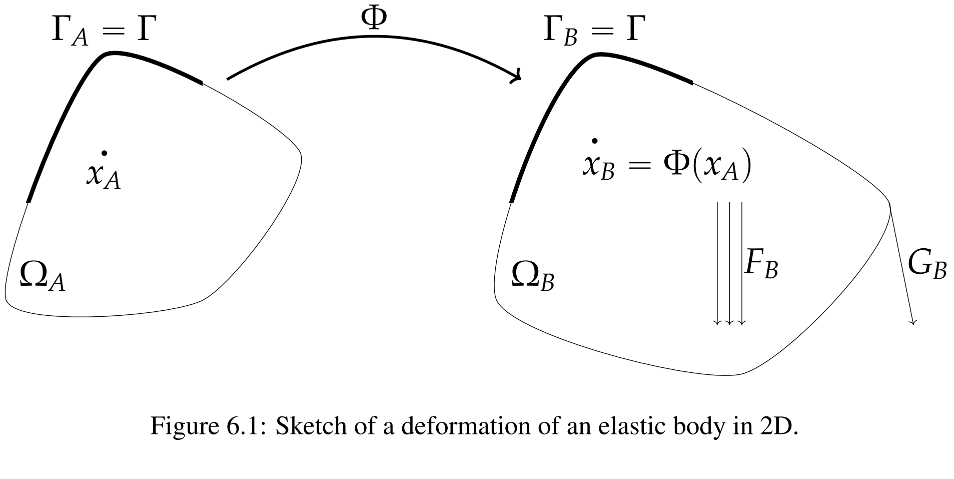


Figure 6.1: Sketch of a deformation of an elastic body in 2D.

We aim to derive equilibrium equations for the deformation Φ corresponding to the acting forces. In the following, we assume that the boundary $\partial\Omega_A$, the deformation Φ, and the forces F_B, G_B are sufficiently smooth, but we do not discuss the specifically required regularity in detail, therefore the following derivations are rather formally. We take into account the Cauchy–Euler stress principle. This fundamental axiom of continuum mechanics postulates the existence of a vector field t_B describing contact forces between two parts of the body.

Axiom 6.1.1 (Cauchy–Euler Stress Principle)**.** There exists a vector field $t_B \colon \overline{\Omega_B} \times \mathcal{S}^2 \to \mathbb{R}^3$, which satisfies

1. $t_B(x_B, n_B) = G_B(x_B)$ for all $x_B \in \Gamma_B^N$ where the outer unit normal n_B exists,
2. (force balance) for all $Y \subset \Omega_B$ we have $\int_Y F_B(x_B)\,\mathrm{d}x_B + \int_{\partial Y} t_B(x_B, \nu_Y)\,\mathrm{d}\mathscr{H}^2(x_B) = 0$, and
3. (momentum balance) for all $Y \subset \Omega_B$ we have $\int_Y x_B \times F_B(x_B)\,\mathrm{d}x_B + \int_{\partial Y} x_B \times t_B(x_B, \nu_Y)\,\mathrm{d}\mathscr{H}^2(x_B) = 0$.

Here, for a subset $Y \subset \Omega_B$, ν_Y denotes the outer unit normal along ∂Y.

Furthermore, under suitable regularity assumptions on t_B, the existence of a so-called Cauchy stress tensor $\mathbb{T}_B \colon \overline{\Omega_B} \to \mathbb{R}^{3\times 3}_{\mathrm{sym}}$ can be established, which relates t_B to a partial differential equation on the deformed domain.

Theorem 6.1.2 (Cauchy's Theorem)**.** *Assume that F_B is continuous and t_B is continuously differentiable in the first and continuous in the second argument. Then there exists a continuously differentiable tensor field $\mathbb{T}_B \colon \overline{\Omega_B} \to \mathbb{R}^{3\times 3}_{sym}$ s.t. $t_B(x_B, \nu) = \mathbb{T}_B(x_B)\nu$ for all $x_B \in \overline{\Omega_B}$, $\nu \in \mathcal{S}^2$ and*

$$\begin{cases} -\mathrm{div}\mathbb{T}_B(x_B) = F_B(x_B) & \forall x_B \in \Omega_B\,, \\ \mathbb{T}_B(x_B)n_B(x_B) = G_B(x_B) & \forall x_B \in \Gamma_B^N\,. \end{cases} \tag{6.1}$$

Proof. See [Cia88, Theorem 2.3-1]. □

To transform the PDE (6.1) to the undeformed domain, we introduce the first Piola–Kirchhoff stress tensor $\mathbb{T}_A$, which is defined by solving $\int_{\Omega_B} \mathrm{div}\mathbb{T}_B(x_B)\theta(x_B)\,\mathrm{d}x_B = \int_{\Omega_A} \mathrm{div}\mathbb{T}_A(x_A)\theta(\Phi(x_A))\,\mathrm{d}x_A$ for all deformations θ, and thus, is pointwise given by

$$\mathbb{T}_A(x_A) = \det(D\Phi(x_A))\mathbb{T}_B(\Phi(x_A))(D\Phi(x_A))^{-T}\,.$$

Since $\mathbb{T}_A$ is in general not symmetric, we usually consider the second Piola–Kirchhoff stress tensor

$$\Sigma_A(x_A) = D\Phi(x_A)^{-1}\mathbb{T}_A(x_A) = \det(D\Phi(x_A))D\Phi(x_A)^{-1}\mathbb{T}_B(\Phi(x_A))D\Phi(x_A)^{-T}\,,$$

which is symmetric. Then the PDE (6.1) transforms to

$$\begin{cases} -\mathrm{div}(D\Phi(x_A)\Sigma_A(x_A)) = F_A(x_A) := \det(D\Phi(x_A))F_B(\Phi(x_A)) & \forall x_A \in \Omega_A\,, \\ D\Phi(x_A)\Sigma_A(x_A)n_A(x_A) = G_A(x_A) := \det(D\Phi(x_A))|D\Phi(x_A)^{-T}n_A(x_A)|G_B(\Phi(x_A)) & \forall x_A \in \Gamma_A^N\,. \end{cases} \tag{6.2}$$

In the following, we restrict to elastic materials, which are defined by the property that the Cauchy stress tensor only depends on the gradient of the deformation.

Definition 6.1.3 (Elastic Material)**.** A material is called elastic if there exists a mapping $\mathbb{T}^{\mathrm{resp}} \colon \overline{\Omega_A} \times \mathbb{R}^{3\times 3}_+ \to \mathbb{R}^{3\times 3}_{\mathrm{sym}}$ called the response function for the Cauchy stress, s.t. for all deformations Φ we have the constitutive relation of the material

$$\mathbb{T}_B(\Phi(x_A)) = \mathbb{T}^{\mathrm{resp}}(x_A, D\Phi(x_A))\,.$$

The response functions for the first and second Piola–Kirchhoff stress tensor are defined by

$$\mathbb{T}_A^{\mathrm{resp}}(x_A, M) := \det(M)\,\mathbb{T}^{\mathrm{resp}}(x_A, M)\,M^{-T} \quad \text{and} \quad \Sigma_A^{\mathrm{resp}}(x_A, M) := \det(M)M^{-1}\,\mathbb{T}^{\mathrm{resp}}(x_A, M)\,M^{-T}\,,$$

s.t.

$$\mathbb{T}_A(x_A) = \mathbb{T}_A^{\mathrm{resp}}(x_A, D\Phi(x_A)) \quad \text{and} \quad \Sigma_A(x_A) = \Sigma_A^{\mathrm{resp}}(x_A, D\Phi(x_A))\,.$$

Furthermore, we assume invariance under change of the observer.

Axiom 6.1.4 (Material Frame-Indifference)**.** Let Φ_1, Φ_2 be deformations of Ω_A s.t. $\Phi_2 = Q\Phi_1$ with $Q \in SO(3)$ and denote by t_1, t_2 the corresponding vector fields in the Cauchy–Euler stress principle in Axiom 6.1.1. Then we have for all $x_A \in \Omega_A$ and for all $\nu \in \mathcal{S}^2$ that $t_2(\Phi_2(x_A), Q\nu) = Qt_1(\Phi_1(x_A), \nu)$.

As a direct consequence, we can express the response function of the second Piola–Kirchhoff stress tensor in terms of the symmetrized deformation gradient

$$\mathbb{C}(x_A) := D\Phi(x_A)^T D\Phi(x_A)\,, \tag{6.3}$$

which we call the (right) Cauchy–Green strain tensor.

Theorem 6.1.5 (Characterization of Material Frame-Indifference)**.** *The response function $\mathbb{T}^{resp}$ for the Cauchy stress satisfies the axiom of material frame-indifference if and only if for all $x_A \in \overline{\Omega_A}$ we have*

$$\mathbb{T}^{resp}(x_A, QM) = Q\mathbb{T}^{resp}(x_A, M)Q^T \quad \forall M \in \mathbb{R}^{3\times3}_+, Q \in SO(3)\,.$$

Furthermore, this is equivalent to the existence of a mapping $\Sigma_A^{resp,sym}\colon \overline{\Omega_A} \times \mathbb{R}^{3\times3}_{sym,+} \to \mathbb{R}^{3\times3}_{sym}$ s.t. $\Sigma_A^{resp}(x_A, M) = \Sigma_A^{resp,sym}(x_A, M^T M)$ for all $M \in \mathbb{R}^{3\times3}_+$.

Proof. See [Cia88, Theorem 3.3-1]. □

Moreover, we consider a special class of so-called isotropic materials.

Definition 6.1.6 (Isotropic Elastic Material)**.** An elastic material is isotropic at a point $x_A \in \overline{\Omega_A}$ if its response function for the Cauchy stress satisfies

$$\mathbb{T}^{\mathrm{resp}}(x_A, MQ) = \mathbb{T}^{\mathrm{resp}}(x_A, M) \quad \forall M \in \mathbb{R}^{3\times3}_+, Q \in SO(3)\,.$$

Now, isotropy implies that the response function for the Cauchy stress can be expressed in terms of the (left) Cauchy–Green strain tensor $D\Phi(x_A)D\Phi(x_A)^T$.

Theorem 6.1.7 (Characterization of Isotropy)**.** *An elastic material is isotropic at a point $x_A \in \overline{\Omega_A}$ if and only if there exists a mapping $\mathbb{T}_A^{resp,sym}(x_A, \cdot)\colon \mathbb{R}^{3\times3}_{sym,+} \to \mathbb{R}^{3\times3}_{sym}$ s.t.*

$$\mathbb{T}^{resp}(x_A, M) = \mathbb{T}_A^{resp,sym}(x_A, MM^T) \quad \forall M \in \mathbb{R}^{3\times3}_+\,.$$

Proof. See [Cia88, Theorem 3.4-1]. □

Then the Rivlin–Ericksen theorem allows a representation of the response function for the second Piola–Kirchhoff stress tensor in terms of the principle invariants $\iota(\mathbb{C}) = (\mathrm{tr}(\mathbb{C}), \mathrm{tr}(\mathrm{cof}(\mathbb{C})), \det(\mathbb{C}))$ of the Cauchy–Green strain tensor $\mathbb{C}$ as defined in (6.3), which can be computed w.r.t. the deformation by

$$\iota_1(\mathbb{C}) = \|D\Phi\|_F^2\,, \quad \iota_2(\mathbb{C}) = \|\mathrm{cof}(D\Phi)\|_F^2\,, \quad \iota_3(\mathbb{C}) = \det(D\Phi)^2\,.$$

Theorem 6.1.8 (Rivlin–Ericksen Representation Theorem)**.** *Let Φ be a deformation of Ω_A. For an elastic material whose response function is frame-indifferent and isotropic at $x_A \in \overline{\Omega_A}$, the Cauchy stress tensor is given by*

$$\mathbb{T}_B(\Phi(x_A)) = \mathbb{T}^{resp}(x_A, D\Phi(x_A)) = \mathbb{T}_A^{resp,sym}(x_A, D\Phi(x_A)D\Phi(x_A)^T)\,,$$

where the response function is of the form $\mathbb{T}_A^{resp,sym}(x_A, S) = \sum_{k=0}^2 \beta_k(x_A, \iota(S))S^k$ with real valued functions β_k. Furthermore, the second Piola–Kirchhoff stress tensor is given by

$$\Sigma_A(x_A) = \Sigma_A^{resp}(x_A, D\Phi(x_A)) = \Sigma_A^{resp,sym}(x_A, D\Phi(x_A)^T D\Phi(x_A))\,,$$

where the response function is of the form $\Sigma_A^{resp,sym}(x_A, S) = \sum_{k=0}^2 \gamma_k(x_A, \iota(S))S^k$ with real valued functions γ_k.

Proof. See [Cia88, Theorem 3.6-2]. □

Next, we approximate the response function near the identity.

Theorem 6.1.9 (Response Function Near Identity). *Let there be given an elastic material whose response function is frame-indifferent and isotropic at $x_A \in \overline{\Omega_A}$. Assume that the functions γ_k in Theorem 6.1.8 are differentiable for $S = \mathbb{1}_{3\times3}$. Then there exist $\pi(x_A), \lambda(x_A), \mu(x_A) \in \mathbb{R}$ s.t. for all $\mathbb{C} \in \mathbb{R}^{3\times3}_{sym,+}$ we have*

$$\Sigma_A^{resp,sym}(x_A, \mathbb{C}) = -\pi(x_A) + \lambda(x_A)\,\mathrm{tr}(\mathbb{E})\mathbb{1}_{3\times3} + 2\mu(x_A)\mathbb{E} + o(\mathbb{E}; x_A)\,,$$

where $\mathbb{E} := \frac{1}{2}(\mathbb{C} - \mathbb{1}_{3\times3})$. More precisely, we can specify $-\pi(x_A) = \mathbb{T}^{resp}(x_A, \mathbb{1}_{3\times3})$.

Proof. [Cia88, Theorem 3.7-1]. □

In general, for a deformation Φ of Ω_A, we define the Green–Saint-Venant strain tensor by

$$\mathbb{E}(x_A) := \frac{1}{2}(\mathbb{C}(x_A) - \mathbb{1}_{3\times3})\,. \tag{6.4}$$

Furthermore, we say that $\overline{\Omega_A}$ is a natural state if $\mathbb{T}^{\mathrm{resp}}(x_A, \mathbb{1}_{3\times3}) = 0$ for all $x_A \in \overline{\Omega_A}$. We call a material homogeneous if its response function does not depend on the position. Under these two additional assumptions we can write

$$\Sigma_A^{\mathrm{resp,sym}}(x_A, \mathbb{C}) = \Sigma_A^{\mathrm{resp,sym}}(\mathbb{C}) = \lambda\,\mathrm{tr}(\mathbb{E})\mathbb{1}_{3\times3} + 2\mu\mathbb{E} + o(\mathbb{E})\,. \tag{6.5}$$

Here, the values λ, μ are called Lamé–Navier parameters. It is often convenient to consider instead Young's modulus E and the Poisson ratio ν, which are given by

$$E = \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu}\,, \quad \nu = \frac{\lambda}{2(\lambda + \mu)}\,.$$

Then λ and μ are vice-versa determined by

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}\,, \quad \mu = \frac{E}{2(1+\nu)}\,. \tag{6.6}$$

By neglecting the higher-order terms in (6.5), a possible response function is given in the following definition.

Definition 6.1.10 (Saint-Venant–Kirchhoff Material). An elastic material is a Saint-Venant–Kirchhoff material if its response function is of the form

$$\Sigma_A^{\mathrm{resp,sym}}(x_A, \mathbb{C}) = \Sigma_A^{\mathrm{resp,sym}}(x_A, \mathbb{1}_{3\times3} + 2\mathbb{E}) = \lambda\,\mathrm{tr}(\mathbb{E})\mathbb{1}_{3\times3} + 2\mu\mathbb{E} \tag{6.7}$$

for all $\mathbb{C} = \mathbb{1}_{3\times3} + 2\mathbb{E} \in \mathbb{R}^{3\times3}_{\mathrm{sym},+}$.

6.1.1 Hyperelastic Materials

Next, we consider a special class of so-called hyperelastic materials, which allows solving the PDE (6.2) by variational methods, *i.e.*, finding a stationary point of an energy functional.

Definition 6.1.11 (Hyperelastic Material). An elastic material is hyperelastic if there exists a stored energy density function $\mathbb{W} : \overline{\Omega_A} \times \mathbb{R}^{3\times3}_{+} \to \mathbb{R}$, s.t. the response function for the first Piola–Kirchhoff stress tensor is given by

$$\mathbb{T}_A^{\mathrm{resp}}(x_A, M) = \partial_M \mathbb{W}(x_A, M) \quad \forall x_A \in \overline{\Omega_A}\ \forall M \in \mathbb{R}^{3\times3}_{+}\,.$$

As in Theorem 6.1.5, the Axiom 6.1.4 of material frame-indifference leads to the existence of a function $\mathbb{W}^{\mathrm{sym}}$ s.t. $\mathbb{W}(x_A, M) = \mathbb{W}^{\mathrm{sym}}(x_A, M^T M)$, which can be determined by $\Sigma_A^{\mathrm{resp,sym}}(x_A, \mathbb{C}) = 2\partial_{\mathbb{C}}\mathbb{W}^{\mathrm{sym}}(x_A, \mathbb{C})$ (see [Cia88, Theorem 4.2-1, 4.2-2]). Together with the isotropy constraint 6.1.6, it can be verified (see [Cia88, Theorem 4.4-1]) that similar to the Rivlin–Ericksen representation Theorem 6.1.8 the hyperelastic energy density function $\mathbb{W}^{\mathrm{sym}}$

can be expressed in terms of the principle invariants $\iota(\mathbb{C})$. Looking at the approximative behavior near the identity as in Theorem 6.1.9 for a homogeneous material in a natural state, we obtain (see [Cia88, Theorem 4.5-1])

$$\mathbb{W}^{\mathrm{sym}}(x_A, \mathbb{C}) = \mathbb{W}^{\mathrm{sym}}(\mathbb{C}) = \frac{\lambda}{2}(\operatorname{tr}\mathbb{E})^2 + \mu \operatorname{tr}(\mathbb{E}^2) + o(\|\mathbb{E}\|^2)\,.$$

Especially, a Saint-Venant–Kirchhoff material is a hyperelastic material with

$$\mathbb{W}^{\mathrm{sym}}(x_A, \mathbb{C}) = \mathbb{W}^{\mathrm{sym}}(\mathbb{C}) = \frac{\lambda}{2}(\operatorname{tr}\mathbb{E})^2 + \mu \operatorname{tr}(\mathbb{E}^2)\,.$$

A further example of a hyperelastic energy density function (see, *e.g.*, [Cia88]), which we take into account later in Chapter 8, is given by

$$\mathbb{W}(M) = \frac{\mu}{2}\|M\|_F^2 + \frac{\lambda}{4}\det(M)^2 - \left(\mu + \frac{\lambda}{2}\right)\log(\det(M)) - d\frac{\mu}{2} - \frac{\lambda}{4} \tag{6.8}$$

for $M \in \mathbb{R}_+^{d\times d}$ and in space dimension $d = 2, 3$.

Now, we define the stored elastic, the potential, and the free energy functionals by

$$\begin{aligned}
\mathcal{E}_{\mathrm{stored}}(\Phi) &= \int_{\Omega_A} \mathbb{W}(x_A, D\Phi(x_A))\,\mathrm{d}x_A\,,\\
\mathcal{E}_{\mathrm{pot}}(\Phi) &= \int_{\Omega_A} F_A \cdot \Phi\,\mathrm{d}x_A + \int_{\Gamma_A^N} G_A \cdot \Phi\,\mathrm{d}\mathcal{H}^2(x_A)\,,\\
\mathcal{E}_{\mathrm{free}}(\Phi) &= \mathcal{E}_{\mathrm{stored}}(\Phi) - \mathcal{E}_{\mathrm{pot}}(\Phi)\,.
\end{aligned} \tag{6.9}$$

Then we are interested in minimizing the free energy over an admissible set of deformations. By formally considering the Euler–Lagrange equations to this minimization problem, we recover the PDE (6.2). This reformulation as a variational problem is essential for our numerical solution scheme, since it can be solved, *e.g.*, by Newton's method.

Moreover, even for large strains, the variational problem allows proving the existence of deformations minimizing the free energy. In [Bal77], such an existence result was established for deformations in the Sobolev space $W^{1,\alpha}(\Omega_A, \mathbb{R}^3)$ for $\alpha \geqslant 2$, provided that the stored energy function $\mathbb{W}$ is polyconvex, has certain growth conditions in the principle invariants, and converges to infinity as $\det(M)$ tends to zero. Note that the density function of a Saint-Venant–Kirchoff material is not polyconvex, s.t. for an existence result quasiconvexification is required.

6.1.2 Linear Elasticity

For a linearization, we introduce the displacement $U = \Phi - \mathrm{id}$. Then the Green–Saint-Venant strain tensor $\mathbb{E}$ as defined in (6.4) can be expressed in terms of the displacement

$$\mathbb{E} = \frac{1}{2}(\mathbb{C} - \mathbb{1}_{3\times3}) = \frac{1}{2}(D\Phi^T D\Phi - \mathbb{1}_{3\times3}) = \frac{DU + DU^T}{2} + \frac{1}{2}DU^T DU = \varepsilon(U) + \frac{1}{2}DU^T DU\,,$$

where the symmetrized gradient $\varepsilon(U) = \frac{1}{2}(DU + DU^T)$ is called the linearized strain tensor. Then, for the Saint-Venant–Kirchhoff material (6.7) the PDE (6.2) is linearized to

$$\begin{cases} -\operatorname{div}(\lambda \operatorname{tr}(\varepsilon(U))\mathbb{1}_{3\times3} + 2\mu\varepsilon(U)) = F_A & \forall x_A \in \Omega_A\,,\\ (\lambda \operatorname{tr}(\varepsilon(U))\mathbb{1}_{3\times3} + 2\mu\varepsilon(U))\,n_A = G_A & \forall x_A \in \Gamma_B^N\,, \end{cases}$$

which is a linear PDE of second-order in the displacement U. Corresponding to (6.9) we define the stored elastic, the potential, and the free energy by

$$\begin{aligned}
\mathcal{E}^{\mathrm{lin}}_{\mathrm{stored}}(U) &= \int_{\Omega_A} \frac{\lambda}{2}\operatorname{div}(U)^2 + \mu\varepsilon(U) : \varepsilon(U)\,\mathrm{d}x_A\,,\\
\mathcal{E}^{\mathrm{lin}}_{\mathrm{pot}}(U) &= \int_{\Omega_A} F_A \cdot U\,\mathrm{d}x_A + \int_{\Gamma_A^N} G_A \cdot U\,\mathrm{d}\mathcal{H}^2(x_A)\,,\\
\mathcal{E}^{\mathrm{lin}}_{\mathrm{free}}(U) &= \mathcal{E}^{\mathrm{lin}}_{\mathrm{stored}}(U) - \mathcal{E}^{\mathrm{lin}}_{\mathrm{pot}}(U)\,.
\end{aligned} \tag{6.10}$$

Existence of minimizing displacements of the free energy $\mathcal{E}^{\text{lin}}_{\text{free}}$ in the space $W^{1,2}(\Omega_A, \mathbb{R}^3)$ is a simple consequence of the direct method in the calculus of variations and Korn's inequality (2.2). Note that for linear elasticity a minimizing displacement U of the free energy satisfies

$$2\mathcal{E}^{\text{lin}}_{\text{stored}}(U) = \mathcal{E}^{\text{lin}}_{\text{pot}}(U) = -2\mathcal{E}^{\text{lin}}_{\text{free}}(U)\,,$$

which does, in general, not hold for nonlinear elasticity.

More generally, we could consider anisotropic materials. Then by linearizing the response function $\Sigma_A^{\text{resp,sym}}$ in the linearized strain $\varepsilon(U)$, we can write

$$\Sigma_A^{\text{resp,sym}}(x_A, \mathbb{C}) = C\varepsilon(U) = \left(\sum_{k,l=1,2,3} C_{ijkl}\varepsilon(U)_{kl} \right)_{i,j=1,2,3}\,, \tag{6.11}$$

where $C \in \mathbb{R}^{3\times3\times3\times3}$ is a fourth-order tensor and $\sigma := C\varepsilon(U) \in \mathbb{R}^{3\times3}$ is called the linear stress tensor. Since $\varepsilon(U)$ and σ are symmetric, we can deduce that $C_{ijkl} = C_{ijlk} = C_{jikl} = C_{klij}$ for all $i, j, k, l = 1, 2, 3$. Thus, the effective degrees of freedom of C are reduced to 21. In this thesis, we essentially consider isotropic materials. However, a composition of two different isotropic materials behaves anisotropic.

6.2 Homogenization

Now, we consider a rapidly oscillating material distribution, which determines a microstructure on the macroscopic reference domain. In [Bab76], the aspect ratio of microcells to the macroscale was discussed. Then, in the limit of vanishing aspect ratio, the theory of mathematical homogenization explains the macroscopic behavior of the material. In [BLP78], periodic microstructures were investigated. General compactness theorems were established in [Mur78, Tar79, MT97]. For a more detailed introduction into the field of mathematical homogenization, we refer the reader to Allaire's famous book [All02].

As above, we take into account a reference domain $\Omega_A \subset \mathbb{R}^3$ of an elastic body. Moreover, we assume that a force F_A acting on Ω_A is given. For simplicity, we neglect boundary forces. Here, we restrict to linear elasticity and denote by

$$M^4_{\text{sym}} := \left\{ C = (C_{ijkl})_{i,j,k,l=1,2,3} \in \mathbb{R}^{3\times3\times3\times3} \ : \ C_{ijkl} = C_{klij} = C_{jikl} = C_{ijlk} \right\}$$

the set of fourth-order elasticity tensors (see (6.11)). Furthermore, for lower and upper bounds $\alpha, \beta > 0$, we consider the space of admissible Hooke's laws

$$M_{\alpha,\beta} = \left\{ C \in M^4_{\text{sym}} \ : \ C\xi : \xi \geqslant \alpha|\xi|^2\,,\ C^{-1}\xi : \xi \geqslant \beta|\xi|^2 \ \forall \xi \in \mathbb{R}^{3\times3}_{\text{sym}} \right\}.$$

Since $C \in M_{\alpha,\beta}$ satisfies $\alpha|\xi|^2 \leqslant C\xi : \xi \leqslant \beta^{-1}|\xi|^2$, we suppose $\alpha\beta \leqslant 1$.

Now, we define convergence of a sequence of material distributions on $M_{\alpha,\beta}$ in a sense s.t. for arbitrary forces the corresponding equilibrium displacements converge.

Definition 6.2.1 (H-convergence)**.** A sequence $(C^h)_h \subset L^\infty(\Omega_A, M_{\alpha,\beta})$ converges in the sense of homogenization (simply H-converges) to $C^* \in L^\infty(\Omega_A, M_{\alpha,\beta})$ if for all $F_A \in \left(W^{1,2}_0(\Omega_A, \mathbb{R}^3)\right)'$ the sequence $(U_h)_h \subset W^{1,2}_0(\Omega_A, \mathbb{R}^3)$ of weak solutions to the state equation

$$\begin{cases} -\operatorname{div}(C^h\varepsilon(U_h)) = F_A & \text{in } \Omega_A\,, \\ U_h = 0 & \text{on } \partial\Omega_A \end{cases}$$

satisfies $U_h \rightharpoonup U_*$ in $W_0^{1,2}(\Omega_A, \mathbb{R}^3)$ for $h \to 0$ and $C^h\varepsilon(U_h) \rightharpoonup C^*\varepsilon(U_*)$ in $L^2(\Omega_A)$ for $h \to 0$, where $U_* \in W_0^{1,2}(\Omega_A, \mathbb{R}^3)$ is the weak solution to

$$\begin{cases} -\mathrm{div}(C^*\varepsilon(U_*)) = F_A & \text{in } \Omega_A\,, \\ U_* = 0 & \text{on } \partial\Omega_A\,. \end{cases}$$

We call C^* the homogenized or effective elasticity tensor.

In the following, we consider two cases, namely those of a periodic and a one-dimensional structure, where the homogenized elasticity tensors can be computed explicitly.

Periodic Homogenization. First, we restrict to a specific sequence of a period material distribution with decreasing cell size. Such a scenario is depicted in Figure 6.2.

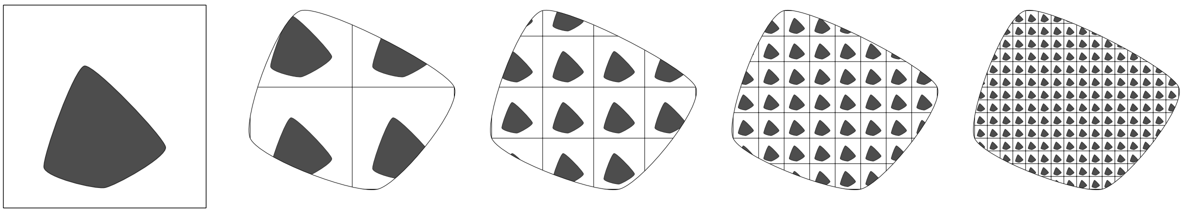


Figure 6.2: Sketch of a periodic microcell in 2D, which generates a corresponding sequence $(C^h)_{h=\frac{1}{2},\frac{1}{4}\frac{1}{8},\frac{1}{16},\dots}$ of elasticity tensors on the domain Ω_A.

For $p \in [1, \infty]$ and $m \in \mathbb{N}_+$, we introduce spaces of periodic functions

$$\begin{aligned} L^p_\#((0,1)^3) &:= \left\{ f \in L^p((0,1)^3) \ : \ f \text{ periodic on } (0,1)^3 \right\}\,, \\ W^{m,p}_\#((0,1)^3) &:= \left\{ f \in W^{m,p}((0,1)^3) \ : \ f \text{ periodic on } (0,1)^3\,,\ \int_{(0,1)^3} f \,\mathrm{d}x = 0 \right\}\,. \end{aligned}$$

Now, we obtain H-convergence, and the homogenized elasticity tensor can be computed explicitly.

Theorem 6.2.2 (Periodic Homogenization)**.** *Let $C^1 \in L^\infty_\#((0,1)^3, M_{\alpha,\beta})$ and $C^h(x) := C^1(\frac{x}{h})$ for $h \in (0,1)$. Then the sequence $(C^h)_h$ H-converges to C^*, which can be computed into directions $\xi_i, \xi_j \in \mathbb{R}^{3\times 3}_{sym}$ by*

$$C^*\xi_i : \xi_j = \int_{(0,1)^3} C^1(\xi_i + \varepsilon(U_i)) : (\xi_j + \varepsilon(U_j))\,\mathrm{d}x\,,$$

where for $k \in \{i,j\}$ the displacement $U_k \in W^{1,2}_\#((0,1)^3, \mathbb{R}^3)$ is the weak solution to

$$-\mathrm{div}(C^1(\xi_k + \varepsilon(U_k))) = 0 \quad \text{in } (0,1)^3\,.$$

Proof. See [All02, Chapter 1.1.4]. □

Homogenization in 1D. So far, we have considered domains in 3D, since this is our main case of interest, of which we especially make use of in Chapter 7. However, the definition of H-convergence transfers analogously to arbitrary dimensions. Next, we consider the one-dimensional case, where the homogenization limit can be characterized in general.

Theorem 6.2.3 (Homogenization in 1D)**.** *Let $I \subset \mathbb{R}$ be a compact interval. We consider a sequence $(C^h)_h \subset L^\infty(I)$ and assume uniform bounds $\alpha \leqslant C^h(t) \leqslant \beta$ for a.e. $t \in I$. Then there exists $B^* \in L^\infty(I)$ s.t. $(C^h)^{-1} \overset{*}{\rightharpoonup} B^*$ in L^∞ for $h \to 0$. and the sequence $(C^h)_h$ H-converges to $C^* := (B^*)^{-1}$.*

Proof. See [All02, Chapter 1.2.3]. □

Note that in this one-dimensional case the involved differential equations are just of type $(C^h U')' = F_A$. In the case of a periodic material, *e.g.*, on the interval $[0,1]$ generated from $C^1 = \alpha\chi_{(0,\lambda)} + \beta\chi_{(\lambda,1)}$ the weak-* limit is known and thus the homogenized tensor is explicitly given by the harmonic mean

$$C^* = \frac{\alpha\beta}{\alpha(1-\lambda)+\beta\lambda}\,.$$

Finally, we remark that also in the case of so-called laminated structures, where the sequence of elasticity tensors is generated by oscillations into a single direction, an explicit formula for the homogenized elasticity tensor can be derived.

6.3 Elastic Shape Optimization

Later, in Chapter 7 and Chapter 8, we study specific shape optimization problems. Roughly speaking, in elastic shape optimization, we are concerned with finding a subdomain $O \subset \Omega_A$ optimizing the mechanical stability within a set $\mathscr{O}$ of admissible subdomains. Here, following [PRW12], we formally introduce a particular class of shape optimization problems and present a corresponding phase-field relaxation.

State Equation. First, we consider a fixed subdomain $O \in \mathscr{O}$, which we identify with its characteristic function $\chi_O \colon \Omega_A \to \{0,1\}$. We assume that forces F_A, G_A are acting on the reference domain Ω_A. Then, we seek for a deformation of O minimizing the free energy, which we then denote by $\Phi(\chi_O)$. Here, we do not discuss assumptions on the domain O to guarantee the existence of $\Phi(\chi_O)$. Instead, we make use of the so-called ersatz material approach by substituting a soft material with a small factor $\tau > 0$ on the complementary set $\Omega_A \backslash O$. This allows considering the elastic problem on the full domain Ω_A, where, according to (6.9), we define energy functionals

$$\begin{aligned}
\mathcal{E}^\tau_{\text{stored}}(\chi_O,\Phi) &= \int_O \mathbb{W}(x_A, D\Phi(x_A))\,\mathrm{d}x_A + \int_{\Omega_A\backslash O} \tau\, \mathbb{W}(x_A, D\Phi(x_A))\,\mathrm{d}x_A \\
&= \int_{\Omega_A} (\chi_O + (1-\chi_O)\tau)\ \mathbb{W}(x_A, D\Phi(x_A))\,\mathrm{d}x_A\,, \\
\mathcal{E}_{\text{pot}}(\Phi) &= \int_{\Omega_A} F_A\cdot\Phi\,\mathrm{d}x_A + \int_{\Gamma_A^N} G_A\cdot\Phi\,\mathrm{d}\mathscr{H}^2(x_A)\,, \\
\mathcal{E}^\tau_{\text{free}}(\chi_O,\Phi) &= \mathcal{E}^\tau_{\text{stored}}(\chi_O,\Phi) - \mathcal{E}_{\text{pot}}(\Phi)\,.
\end{aligned}$$

Then we ask for a deformation of the full domain Ω_A solving the so-called state equation

$$\Phi(\chi_O) \in \operatorname*{arg\,min}_{\Phi\in\mathcal{A}} \mathcal{E}^\tau_{\text{free}}(\chi_O,\Phi)\,,$$

where $\mathcal{A}$ is a suitable space of deformations encoding boundary conditions and regularity assumptions. Nevertheless, in the case of nonlinear elasticity, the uniqueness of global minimizers of the relaxed free energy $\mathcal{E}^\tau_{\text{free}}$ is not guaranteed, and a set of minimizers has to be considered. In particular, in the numerical implementation, we have to cope with an even larger set of local minimizers.

Cost functional. Now, still using the ersatz material approach with factor $\tau > 0$, to measure the mechanical stability of O, we take into account a cost functional $\mathcal{J}^\tau_{\text{expl}} \colon \mathscr{O} \times \mathcal{A} \to \mathbb{R}$ explicitly depending on the domain and the deformation. Then we define a total cost functional $\mathcal{J}^\tau_{\text{tot}} \colon \mathscr{O} \to \mathbb{R}$ by

$$\mathcal{J}^\tau_{\text{tot}}(\chi_O) = \mathcal{J}^\tau_{\text{expl}}(\chi_O, \Phi(\chi_O))$$

and ask for the optimal subdomain $O \subset \mathscr{O}$ minimizing $\mathcal{J}^\tau_{\text{tot}}$. In the case of nonlinear elasticity, in [PRW12] the three functionals

$$\begin{aligned}
\mathcal{J}^\tau_{\text{stored}}(\chi_O,\Phi) &= 2\mathcal{E}^\tau_{\text{stored}}(\chi_O,\Phi)\,, \\
\mathcal{J}^\tau_{\text{pot}}(\chi_O,\Phi) &= \mathcal{E}_{\text{pot}}(\Phi)\,, \\
\mathcal{J}^\tau_{\text{free}}(\chi_O,\Phi) &= -2\mathcal{E}^\tau_{\text{free}}(\chi_O,\Phi)
\end{aligned}$$

have been compared as possible for $\mathcal{J}^{\tau}_{\text{expl}}$. For linear elasticity, the counterparts $\mathcal{E}^{\text{lin}}_{\text{pot}}(U(\chi_O))$, $2\mathcal{E}^{\tau,\text{lin}}_{\text{stored}}(\chi_O, U(\chi_O))$, and $-2\mathcal{E}^{\tau,\text{lin}}_{\text{free}}(\chi_O, U(\chi_O))$ coincide for the unique minimizing displacement $U(\chi_O)$. Without any restrictions on the set $\mathcal{O}$, the minimizer for any of these cost functionals is given by the full domain $O = \Omega_A$. Typically, we impose a volume constraint by defining the set of admissible subdomains by

$$\mathcal{O} = \{O \subset \Omega_A \ : \ \mathcal{V}(\chi_O) \leqslant V\}\,,$$

where $\mathcal{V}(\chi_O) = \mathscr{L}(O)$ denotes the Lebesgue measure of O and $V \in (0, \mathscr{L}(\Omega_A))$. Alternatively, the functional $\lambda\mathcal{V}(\chi_O)$ can be added as a penalty to the cost $\mathcal{J}^{\tau}_{\text{tot}}$ for some constant $\lambda \in \mathbb{R}_+$, which can vice versa be interpreted as a Lagrange multiplier.

6.3.1 Perimeter Regularization and Phase-Field Approximation

In general, shape optimization problems of the above type are ill-posed, even for the ersatz material approach. Under the assumption that the subdomain O is measurable, the characteristic function χ_O belongs to the space $L^\infty(\Omega_A, \{0,1\})$. Unfortunately, the limit χ^* of a minimizing sequence $\chi_k \overset{*}{\rightharpoonup} \chi^*$ in L^∞ with $\chi_k \in L^\infty(\Omega_A, \{0,1\})$ does not necessarily take values in $\{0,1\}$, but in the interval $[0,1]$, and thus, χ^* cannot be identified with a subdomain O^*. A possible relaxation in [ABFJ97] takes into account the homogenization method, as described in Section 6.2. Here, we consider another type of regularization by adding a perimeter term to the cost functional.

Definition 6.3.1 (Perimeter)**.** For $\chi \in BV(\Omega_A, \{0,1\})$ we define the perimeter in Ω_A as

$$\text{Per}_{\Omega_A}(\chi) := |D\chi|_{TV}(\Omega_A)\,.$$

Note that for a smooth set O we have $\text{Per}_{\Omega_A}(\chi) = \mathscr{H}^2(\partial O)$.

Then, for a regularization parameter $\eta > 0$, we consider the total cost functional

$$\mathcal{J}^{\eta,\tau}_{\text{tot}}(\chi_O) = \mathcal{J}^{\tau}_{\text{expl}}(\chi_O, \Phi(\chi_O)) + \eta\text{Per}_{\Omega_A}(\chi_O)\,,$$

which we aim to minimize over all χ in a suitable space $\mathcal{O} \subset BV(\Omega_A, \{0,1\})$. Such a regularized shape optimization problem was, *e.g.*, investigated in [AB93] for heat diffusion as state equation and in [PRW12] for nonlinear elasticity.

For the numerical implementation, it is furthermore advantageous to approximate a characteristic function by a phase-field function $v \in W^{1,2}(\Omega_A, [-1,1])$, where the value $v = 1$ represents the domain, the value $v = -1$ the complementary set, and values between -1 and 1 are allowed to represent a diffuse interface. Then, for a smooth approximation of the perimeter functional in terms of the phase-field variable, we consider the Modica–Mortola [MM77] functional $\mathcal{A}^\epsilon \colon W^{1,2}(\Omega_A) \to \mathbb{R}$ defined as

$$\mathcal{A}^\epsilon(v) = \frac{1}{2}\int_{\Omega_A} \epsilon|\nabla v|^2 + \frac{1}{\epsilon}W(v)\,\mathrm{d}x_A\,, \tag{6.12}$$

where we choose W as the double-well function

$$W(v) = \frac{9}{16}(v^2-1)^2\,. \tag{6.13}$$

Here, the parameter $\epsilon > 0$ is related to the interface width between two phases. In the limit $\epsilon \to 0$, the following convergence result was established.

Theorem 6.3.2 (Γ-Convergence of Modica–Mortola Functional)**.** *For the sequence of functionals* $(\mathcal{A}^\epsilon)_{\epsilon>0}$ *as defined in* (6.12)*, we have* Γ*-convergence w.r.t. strong convergence in* $L^1(\Omega_A)$ *to the functional*

$$\mathcal{A}^0(v) := \begin{cases} \mathit{Per}_{\Omega_A}\left(\chi_{\{x_A\in\Omega_A\,:\,v(x_A)=1\}}\right) & \textit{if } v(x_A) \in \{-1,1\} \textit{ for a.e. } x_A \in \Omega_A\,, \\ \infty & \textit{otherwise}\,. \end{cases}$$

Proof. See [Bra06, Theorem 7.3]. □

Finally, to define the total cost functional $\mathcal{J}_{\text{tot}}^{\eta,\tau}$ in terms of the phase-field variable v, the characteristic function χ_O has to be approximated in the integrands of the stored elastic energy $\mathcal{E}_{\text{stored}}^{\tau}$ and the volume $\mathcal{V}(O) = \int_{\Omega_A} \chi_O \, \mathrm{d}x_A$. In [PRW12], for both functionals, a quadratic approximation with $\chi(v) := \frac{1}{4}(v+1)^2$ was applied. Because of the Γ-convergence result in Theorem 6.3.2, this choice does not matter in the limit $\epsilon \to 0$. For example, in [BGHR16], a sharp interface limit for a generic class of cost functionals was derived in the case of linear elasticity, provided that the phase-field approximation admits uniform coercivity and continuity bounds. However, for a concrete ϵ and intermediate values of $v \in (-1, 1)$, which are always given in the implementation, it appears more natural to define two different approximations of χ_O. For the volume functional, we use $\chi(v) := \frac{1}{2}(v+1)$, which is weak-* continuous w.r.t. convergence in L^∞. In contrast, regarding the homogenization result for one-dimensional material parameters as discussed in Theorem 6.2.3, a harmonic averaging for the stored elastic energy is suitable. Since we, in particular, consider three-dimensional domains and for computational simplicity, we frequently choose a second or a fourth-order polynomial to approximate the characteristic function for the stored elastic energy.

6.3.2 Computing the Shape Derivative

Later, to numerically compute a minimizer of a cost functional $\mathcal{J}_{\text{tot}}$, we apply a first-order method like a gradient descent or Quasi-Newton scheme. This requires to compute the first variation of the cost functional, which is for the above phase-field approximation given by the chain rule as

$$\frac{d}{dv}\mathcal{J}_{\text{tot}}^{\eta,\tau}(v)(\hat{v}) = \frac{d}{dv}\left(\mathcal{J}_{\text{expl}}^{\eta,\tau}(v,\Phi(v))\right) = \partial_v \mathcal{J}_{\text{expl}}^{\eta,\tau}(v,\Phi(v))(\hat{v}) + \partial_\Phi \mathcal{J}_{\text{expl}}^{\eta,\tau}(v,\Phi(v))(\partial_v\Phi(v)(\hat{v}))\,.$$

Unfortunately, the shape sensitivity $\partial_v\Phi(v)(\hat{v})$ is numerically expensive to compute. Therefore, a well-established approach (see, *e.g.*, [HPUU08]) is to take into account a so-called adjoint problem. First, we fix a notation for partial derivatives of second-order.

Remark 6.3.3 (Second-Order Partial Derivatives). For a functional $\mathcal{F}$, we use the notation

$$\partial^2_{X_iX_j}\mathcal{F}(X_1,\dots,X_n)(\widehat{X_j})(\hat{X_i}) := \partial_{X_i}\left(\partial_{X_j}\mathcal{F}(X_1,\dots,X_n)(\widehat{X_j})\right)(\hat{X_i})\,.$$

Now, the adjoint problem is given by

$$\partial^2_{\Phi,\Phi}\mathcal{E}_{\text{free}}^{\tau}(v,\Phi(v))(\hat{\Phi})(A) = -\partial_\Phi \mathcal{J}_{\text{expl}}^{\eta,\tau}(v,\Phi(v))(\hat{\Phi}) \quad \forall \hat{\Phi} \in \mathcal{A}\,,$$

which has to be solved in the variable $A \in \mathcal{A}$. Since the necessary condition for the state equation $\Phi(v) \in \arg\min_{\Phi\in\mathcal{A}} \mathcal{E}_{\text{free}}^{\tau}(v,\Phi)$ is given by $\partial_\Phi \mathcal{E}_{\text{free}}^{\tau}(v,\Phi(v)) = 0$, the inverse function theorem allows computing the shape sensitivity by

$$\partial_v\Phi(v) = -(\partial^2_{\Phi,\Phi}\mathcal{E}_{\text{free}}^{\tau}(v,\Phi(v)))^{-1}\partial^2_{v,\Phi}\mathcal{E}_{\text{free}}^{\tau}(v,\Phi(v))\,.$$

Then, together with the solution to the adjoint problem, we can compute the first variation of the cost functional by

$$\frac{d}{dv}\mathcal{J}_{\text{tot}}^{\eta,\tau}(v)(\hat{v}) = \partial_v \mathcal{J}_{\text{expl}}^{\eta,\tau}(v,\Phi(v))(\hat{v}) + \partial^2_{v,\Phi}\mathcal{E}_{\text{free}}^{\tau}(v,\Phi(v)))(A)(\hat{v})\,. \tag{6.14}$$

Similarly, the shape derivative can be determined for other shape representations, *e.g.*, for a level-set approximation [AJT04], where the differentiability of the signed distance function is used. In the case of nonlinear elasticity, as it was considered in [PRW12], the nonuniqueness of solutions to the state equation has to be coped. Moreover, in Chapter 8, we take into account bending isometries and thus, we have to consider a suitable Lagrangian to solve the corresponding state equation, which requires adapting the expression (6.14) accordingly.

Chapter 7

Simultaneous Elastic Shape Optimization

This chapter is motivated by a biomechanical application in bone tissue engineering. Usually, a broken bone is able to regenerate, where metal implants in the form of plates and screws are well-established to support the healing process. Here, we investigate the case of large scale bone loss, which is, *e.g.*, a consequence of removing osteosarcoma (a cancerous tumor in a bone). Recently, the construction and appropriateness of additional substitutes consisting of biologically degradable polymers are explored. The usage of such degradable materials in medicine is already quite common, *e.g.*, for threads to close incisions in the skin. An example of the application to bone substitutes is studied in [PCW$^+$16], where polycaprolactone is taken into account. Today, 3D printers allow producing a huge variety of polymer scaffolds instantaneously. Since we are dealing with large scale bone loss, such a polymer scaffold is required to be resistant against certain exterior forces. Furthermore, the substitute has a specific microstructure, s.t. during the regeneration process, new bone tissue first grows into the void part of the scaffold. Later, the polymer is degraded, and the bone will completely regenerate. For a more detailed overview of the medical background, we refer the reader to [DPRS19, Section 2].

In the following, our goal is to optimize the microstructure of a polymer implant in the above situation. Therefore, we formulate a suitable shape optimization problem. First, we assume that the microstructure of the scaffold is periodic s.t. we are only concerned with optimizing a single periodic microcell. We consider affine loads acting on the microcell corresponding to macroscopic bending and torsion forces, *e.g.*, for a substitute of a section of the tibia, a realistic loading scenario consists of unilateral compression and shear. Since deformations of the considered object are expected to be small, we can restrict to linear elasticity. Then, we take into account both the mechanical stability of the polymer scaffold and the complementary set, where new bone tissue will grow first. Thus, we arrive at a simultaneous elastic shape optimization, since within a given domain (the periodic microcell) an object, as well as its complementary set, has to be optimized w.r.t. mechanical stability.

In Section 7.1, we formally derive the corresponding shape optimization problem, where we start, based on the theory of homogenization, with a formulation of state equations related to a multiple load scenario. Then we define a suitable cost functional penalizing the less stiff object by taking into account the relevant entries of the effective elasticity tensor. For a mathematically rigorous formulation, we propose in Section 7.2 a perimeter regularization of the characteristic function variable representing the domain splitting. Furthermore, an ersatz material approach assuming soft instead of void material on the complementary object is applied to guarantee the existence of solutions to the state equation, since Korn's inequality can be applied on the full domain. We are able to prove the existence of minimizing characteristic functions for this relaxed formulation. The numerical scheme is implemented via a phase-field approximation, which we describe in Section 7.3. Besides the real application, we study in Section 7.4 different load scenarios and material properties of both objects. In Section 7.5, we propose extensions of our model by additionally incorporating volume constraints and diffusion constraints on certain entries of the homogenized diffusion tensor of the bone phase. The latter guarantees that bone can grow appropriately. Finally, a conclusion is given in Section 7.6.

Remark 7.0.1. All results presented in this chapter are joint work with Patrick Dondl, Patrina Poh, and Martin Rumpf and have been published in [DPRS19].

7.1 Simultaneous Elastic Shape Optimization of a Periodic Microcell

As a reference domain representing the (scaled) microcell, we consider the unit cube $\Omega = [0,1]^d$ ($d \geqslant 2$), which allows us to model periodic materials in $\mathbb{R}^d$, where $d = 3$ is the application relevant case. We consider a splitting of the reference domain Ω of an elastic object into two disjoint subdomains $\mathcal{O}^0$ and $\mathcal{O}^1$, *i.e.*,

$$\overline{\Omega} = \overline{\mathcal{O}^0} \cup \overline{\mathcal{O}^1}\,, \quad \mathcal{O}^0 \cap \mathcal{O}^1 = \emptyset\,.$$

In the application, $\mathcal{O}^1$ represents the polymer scaffold and $\mathcal{O}^0$ the complementary set, where bone will grow first. A possible domain splitting is depicted in Figure 7.1. We denote by $m \in \{0,1\}$ an index corresponding to the subdomain $\mathcal{O}^m$, and represent the two disjoint objects by characteristic functions χ^m.

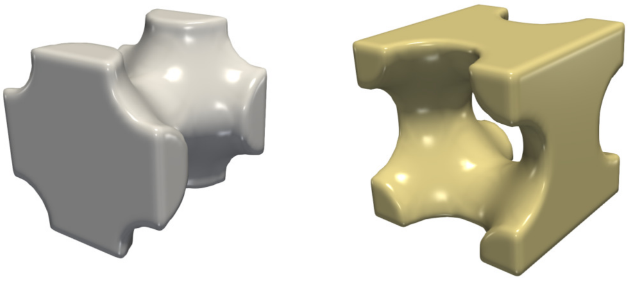


Figure 7.1: Example of a domain splitting of the unit cube $[0,1]^3$ into disjoint sets $\mathcal{O}^0$ and $\mathcal{O}^1$.

7.1.1 State Equations

In the following, we investigate affine displacements $U_l^{\mathrm{aff},m} \colon \Omega \to \mathbb{R}^{d\times d}$ representing a multiple load scenario with $L \in \mathbb{N}_+$ affine loads on the microcell, *i.e.*,

$$U_l^{\mathrm{aff},m}(x) = \xi_l^m x \quad \text{for } l \in \{1,\dots,L\}\,, \tag{7.1}$$

where $\xi_l^m \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$. Usually, we consider equal loads $\xi_l^0 = \xi_l^1$ for both subdomains, where for the application we have in mind combinations of compressions and shears of type

$$\xi_{\mathrm{compr}} = \beta e_i^T e_i\,, \quad \xi_{\mathrm{shear}} = \beta(e_i^T e_j + e_j^T e_i)\,,$$

for $\{e_1,\dots,e_d\}$ being the canonical basis in $\mathbb{R}^d$ and some $\beta \in \mathbb{R}$. To measure stiffness of the objects $\mathcal{O}^m$ in the directions given by the affine displacements, we assume linear elasticity, because deformations are expected to be small. We denote by $C^m = (C^m_{ijkl})_{i,j,k,l=1,\dots,d}$ the corresponding elasticity tensors and assume for simplicity that both materials are isotropic, and thus, determined by the Lamé–Navier parameters $\mu^m > 0$ and $\lambda^m > 0$, *i.e.*, for a displacement $U \colon \Omega \to \mathbb{R}^d$ we have

$$C^m \varepsilon(U) : \varepsilon(U) = 2\mu^m \varepsilon(U) : \varepsilon(U) + \lambda^m \mathrm{div}(U)\mathrm{div}(U)\,.$$

Based on the theory of periodic homogenization (see Section 6.2), we consider the elastic energies

$$\mathcal{E}^m(\chi^m, U_l^{\mathrm{tot},m}) = \int_\Omega \chi^m\, C^m \varepsilon(U_l^{\mathrm{tot},m}) : \varepsilon(U_l^{\mathrm{tot},m})\,\mathrm{d}x = \int_{\mathcal{O}^m} C^m \varepsilon(U_l^{\mathrm{tot},m}) : \varepsilon(U_l^{\mathrm{tot},m})\,\mathrm{d}x\,, \tag{7.2}$$

for affine-periodic displacements $U_l^{\mathrm{tot},m} \colon \mathcal{O}^m \to \mathbb{R}^d$ with

$$U_l^{\mathrm{tot},m} = U_l^m + U_l^{\mathrm{aff},m}\,.$$

Since the actual unknown variable is the periodic counterpart U_l^m, we consider the elastic energy (7.2) only in dependence of the periodic part, and indicate the fixed affine part as an index

$$\begin{aligned}\mathcal{E}_l^m(\chi^m, U_l^m) &= \int_\Omega \chi^m \, C^m \varepsilon(U_l^{\mathrm{tot},m}) : \varepsilon(U_l^{\mathrm{tot},m}) \, \mathrm{d}x \\ &= \int_\Omega \chi^m \, C^m \varepsilon(U_l^m + U_l^{\mathrm{aff},m}) : \varepsilon(U_l^m + U_l^{\mathrm{aff},m}) \, \mathrm{d}x \\ &= \int_\Omega \chi^m \, C^m \left(\varepsilon(U_l^m) + \xi_l^m\right) : \left(\varepsilon(U_l^m) + \xi_l^m\right) \, \mathrm{d}x \,.\end{aligned} \tag{7.3}$$

For the moment, we do not introduce function spaces and just assume that for a prescribed characteristic function χ^m a unique minimizing periodic displacement of the elastic energy $\mathcal{E}_l^m(\chi^m, \cdot)$ in (7.3) exists, which we denote by $U_l^m(\chi^m)$. Note that the displacements here are restricted to the associated object $\mathcal{O}^m$, which is different for a rigorous definition for a relaxed formulation in Section 7.2, where a hard-soft material approximation allows to consider displacements of the full domain Ω. However, to prove such an existence result in this context, we would have to specify precise regularity assumptions on $\mathcal{O}^m$ s.t. Korn's inequality can be applied. Instead, we take the above definitions rather formally and recall (*cf*. Theorem 6.2.2) that for a given characteristic function χ^m the entries of the homogenized elasticity tensor $C^{m,*}(\chi^m)$ can be computed by

$$C^{m,*}(\chi^m)\xi : \xi = \min_{U\colon \Omega \to \mathbb{R}^d \text{ periodic}} \int_{\mathcal{O}^m} \chi^m C^m \; (\varepsilon(U) + \xi) \; : \; (\varepsilon(U) + \xi) \; \mathrm{d}x \tag{7.4}$$

for all $\xi \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$. Thus, minimizing the elastic energy $\mathcal{E}_l^m(\chi^m, \cdot)$ over periodic displacements means computing the entry

$$C^{m,*}(\chi^m)\xi_l^m : \xi_l^m = \mathcal{E}_l^m(\chi^m, U_l^m(\chi^m)) \,.$$

Note that the object $\mathcal{O}^m$ is stiff w.r.t. to the load $U_l^{\mathrm{aff},m}$ if the corresponding entry of the homogenized tensor is large. Next, we propose a cost functional taking these values into account.

7.1.2 Cost Functional

To measure the overall stiffness of a domain $\mathcal{O}^m$, we take into account a continuous function $g^m\colon \mathbb{R}_+^L \to \mathbb{R}$, which should weight the entries of the homogenized elasticity tensor and is therefore supposed to be monotone decreasing in each argument. Thus, we define for both subdomains, respectively, the cost associated with the set of loading conditions as

$$\mathcal{G}^m(\chi^m) := g^m \left(C^{m,*}(\chi^m)\xi_1^m : \xi_1^m, \dots, C^{m,*}(\chi^m)\xi_L^m : \xi_L^m \right) \,. \tag{7.5}$$

For simplicity, we consider an equal load scenario for both subdomains and use the same weighting function. In our implementation, we choose an l^p-norm of the inverse values

$$g^m(E_1, \dots, E_L) = g(E_1, \dots, E_L) = \left(\sum_{l=1,\dots,L} E_l^{-p} \right)^{\frac{1}{p}} \,.$$

for some $p \in [1, \infty)$. For $p \to \infty$, the resulting cost converges to the maximal inverse total energy

$$\max_{l=1,\dots,L} E_l^{-1} = \left(\min_{l=1,\dots,L} E_l \right)^{-1} ,$$

and thus, represents a worst-case optimization problem, where solely the loading scenario with the smallest elastic energy is taken into account.

Finally, we define a total cost functional in dependence of a characteristic function χ representing the domain splitting via $\chi^0 = \chi$ and $\chi^1 = 1 - \chi$. This functional should prefer both subdomains to be stiff and thus penalize large values w.r.t. to the weighting function g. For that purpose, we select the less stiffer object by choosing the maximum value of $\mathcal{G}^m(\chi^m)$, *i.e.*, our simultaneous elastic shape optimization problem is given by minimizing

$$\mathcal{J}_{\text{tot}}(\chi) := \max\left(\mathcal{G}^0(\chi), \mathcal{G}^1(1-\chi)\right) \tag{7.6}$$

over all periodic characteristic function $\chi\colon \Omega \to \{0,1\}$. Compared to other shape optimization problems, where only one subdomain is optimized, no volume constraint or penalty is needed, since there is a competition of both subdomains in the sense that increasing the stiffness of one domain is only possible with a payoff in the cost of the complementary subdomain.

7.2 Hard-Soft Approximation and Perimeter Regularization

Next, we derive a mathematically rigorous formulation for a relaxation of the cost functional (7.6). In particular, we choose appropriate function spaces for the characteristic function and the periodic parts of the displacements. Then, based on a similar result in [AB93] for a scalar-valued problem and in [PRW12] for the existence of minimizing phase-fields in the case of nonlinear elastic shape optimization, we prove the existence of minimizing characteristic functions.

We remember that the cost functional (7.6) is defined for a single characteristic function χ to model the domain splitting. Now, we assume χ to be in the space of functions of bounded variation with periodic boundary conditions. Furthermore, since any periodically extended translation has the same cost, we choose a fixed center of mass $c \in \Omega$, *i.e.*,

$$\chi \in BV_{\#,c}(\Omega,\{0,1\}) := \left\{\chi \in BV(\Omega,\{0,1\}) \ :\ \chi \text{ periodic on } \Omega\,,\ \int_\Omega \chi(x_i - c_i)\,\mathrm{d}x = 0 \text{ for } i = 1,\ldots,d\right\}.$$

For the elastic problem, as in (7.1) we take into account a set of affine displacements $\left(U_l^{\text{aff},m}\right)_{l=1,\ldots,L}$ and periodic parts

$$U_l^m \in W_\#^{1,2}(\Omega,\mathbb{R}^d) = \left\{U \in W^{1,2}(\Omega,\mathbb{R}^d) \ :\ U \text{ periodic on } \Omega\,,\ \int_\Omega U\,\mathrm{d}x = 0\right\}.$$

Next, we take into account an ersatz material approach by replacing the void phase on the complementary set $\Omega\backslash\overline{O^m}$ by a very soft phase, which allows to consider the elastic problems on the full domain Ω instead of on the subdomains O^m. More precisely, the characteristic function χ^m for each object O^m is approximated by

$$\chi^m + \tau(1-\chi^m)$$

for some small constant $\tau > 0$. Then, for $m \in \{0,1\}$ and $l \in \{1,\ldots,L\}$, corresponding to the minimization problem (7.4), we define elastic energies $\mathcal{E}_l^{m,\tau}\colon BV_{\#,c}(\Omega,\{0,1\}) \times W_\#^{1,2}(\Omega,\mathbb{R}^d) \to \mathbb{R}$ as

$$\mathcal{E}_l^{m,\tau}(\chi, U_l^m) = \int_\Omega (\chi + \tau(1-\chi))\, C^m \varepsilon(U_l^m + U_l^{\text{aff},m}) : \varepsilon(U_l^m + U_l^{\text{aff},m})\,\mathrm{d}x\,. \tag{7.7}$$

Now, in this function space setup, for a fixed characteristic function, we can guarantee the existence and uniqueness of a minimizing displacement.

Lemma 7.2.1 (Existence of Unique Minimizing Displacements). *Let $\chi \in BV_{\#,c}(\Omega,\{0,1\})$.*

1. *There exists a unique minimizer $U_l^m(\chi) \in W_\#^{1,2}(\Omega,\mathbb{R}^d)$ of $\mathcal{E}_l^{m,\tau}(\chi,\cdot)$.*

2. *Furthermore, for every sequence $(\chi_k)_{k\in\mathbb{N}} \subset BV_{\#,c}(\Omega,\{0,1\})$ with $\chi_k \overset{*}{\rightharpoonup} \chi$ in BV, there exists a subsequence $(\chi_{k_n})_{n\in\mathbb{N}}$ s.t. $\mathcal{E}_l^{m,\tau}(\chi_k^m, U_l^m(\chi_{k_n})) \to \mathcal{E}_l^{m,\tau}(\chi^m, U_l^m(\chi))$ for $n \to \infty$.*

Proof.

1. Since $\tau > 0$ and χ only takes values in $\{0,1\}$, we have uniform bounds on the coefficients
$$0 < \alpha < \|(\chi + \tau(1-\chi))\, C^m\|_\infty < \beta\,.$$
Existence of a unique minimizer is a direct consequence of the Lax–Milgramm theorem, where coercivity of the corresponding bilinear form follows by Korn's inequality (2.2) combined with Poincaré's inequality.

2. For the second statement, we make use of Γ-convergence of $\mathcal{E}_l^{m,\tau}(\chi_k,\cdot)$ to $\mathcal{E}_l^{m,\tau}(\chi,\cdot)$ w.r.t. the weak $W^{1,2}$-topology for $k \to \infty$. To see this, we recall that weak-* convergence in BV implies strong convergence in L^1 (see Theorem 2.2.4) and there is a subsequence (here again indexed by k) s.t. $\chi_k \to \chi$ pointwise a.e. Then the Γ-liminf inequality is a direct consequence of Theorem 2.3.4. For the Γ-limsup inequality we can choose for any $U_l^m \in W_\#^{1,2}$ the constant recovery sequence $U_{l,k}^m = U_l^m$. Indeed, by the upper bound of the coefficients, the integrands $(\chi_k + \tau(1-\chi_k))\, C^m \varepsilon(U_l^m + U_l^{\mathrm{aff},m}) : \varepsilon(U_l^m + U_l^{\mathrm{aff},m})$ are bounded by the L^1-function $\beta|\varepsilon(U_l^m + U_l^{\mathrm{aff},m})|^2$. Thus, because of the pointwise convergence of the subsequence, the Γ-limsup inequality follows by Lebesgue's dominated convergence theorem. Furthermore, the elastic energies $(\mathcal{E}_l^{m,\tau}(\chi_k,\cdot))_{k\in\mathbb{N}}$ are equi-coercive because of the upper bound on the coefficients. Then convergence of a subsequence follows by the Fundamental Theorem of Γ-convergence (2.3.3).

□

Corresponding to (7.5), the cost $\mathcal{G}^{m,\tau} : BV_{\#,c}(\Omega,\{0,1\}) \to \mathbb{R}$ associated with the set of loading conditions for a specific subdomain is given by
$$\mathcal{G}^{m,\tau}(\chi^m) := g^m\left(C^{m,*}(\chi^m)\xi_1^m : \xi_1^m, \dots, C^{m,*}(\chi^m)\xi_L^m : \xi_L^m\right)\,.$$
Finally, we define the total cost functional $\mathcal{J}_{\mathrm{tot}}^{\eta,\tau} : BV_{\#,c}(\Omega,\{0,1\}) \to \mathbb{R}$ as
$$\mathcal{J}_{\mathrm{tot}}^{\eta,\tau}(\chi) = \max\left(\mathcal{G}^{0,\tau}(\chi), \mathcal{G}^{1,\tau}(1-\chi)\right) + \eta|D\chi|_{TV}(\Omega)\,. \tag{7.8}$$
Here, to regularize the interface between the subdomains, we add the perimeter $\eta|D\chi|_{TV}(\Omega)$ for some constant $\eta > 0$. In the following theorem, we provide the existence of minimizing characteristic functions.

Theorem 7.2.2 (Existence of Optimal Subdomain Splitting)**.** *For $\eta > 0$ and $\tau > 0$, there exists a minimizer $\chi \in BV_{\#,c}(\Omega,\{0,1\})$ of the functional $\mathcal{J}_{tot}^{\eta,\tau}$.*

Proof. First, we take a minimizing sequence $(\chi_k)_{k\in\mathbb{N}} \subset BV_{\#,c}(\Omega,\{0,1\})$ of the functional $\mathcal{J}_{\mathrm{tot}}^{\eta,\tau}$. This sequence is uniformly bounded in $BV(\Omega,\{0,1\})$ because of the perimeter term in the functional $\mathcal{J}_{\mathrm{tot}}^{\eta,\tau}$. Thus, there exists a subsequence for simplicity again denoted by $(\chi_k)_{k\in\mathbb{N}}$ s.t. $\chi_k \overset{*}{\rightharpoonup} \chi$ in $BV_{\#,c}(\Omega,\{0,1\})$. By Lemma 7.2.1, we obtain convergence $\mathcal{E}_l^{m,\tau}(\chi_k^m, U_l^m(\chi_k)) \to \mathcal{E}_l^{m,\tau}(\chi^m, U_l^m(\chi))$ for $k \to \infty$. Since g and the maximum function are continuous, we get in the limit $\mathcal{J}_{\mathrm{tot}}^{\eta,\tau}(\chi) = \lim_{k\to\infty} \mathcal{J}_{\mathrm{tot}}^{\eta,\tau}(\chi_k) = \inf_{BV_{\#,c}(\Omega,\{0,1\})} \mathcal{J}_{\mathrm{tot}}^{\eta,\tau}$. □

7.3 Phase-Field Approximation and Finite Element Discretization

We recall from Section 6.3 that a phase-field approach is quite common in the literature (see, *e.g.*, [PRW12]) to compute a minimizer of an elastic shape optimization problem numerically. Here, we adopt this ansatz by approximating the characteristic function $\chi \in BV_{\#,c}(\Omega,\{0,1\})$ by a phase-field function
$$v \in W_{\#,c}^{1,2}(\Omega,[-1,1]) := \left\{ v \in W^{1,2}(\Omega,[-1,1]) \,:\, v \text{ periodic on } \Omega\,,\ \int_\Omega v(x_i - c_i)\,\mathrm{d}x = 0 \text{ for } i = 1,\dots,d \right\}\,.$$
In the following, we define counterparts of the elastic energies (7.7) and the cost functional (7.8) in terms of the phase-field variable v. Then the core ingredient of our numerical scheme consists in computing the first derivative of the cost functional.

First, for the phase-field function $v \in W_{\#,c}^{1,2}(\Omega,[-1,1])$, we define approximations of the characteristic functions by
$$\chi^0(v) = \frac{1}{16}(1+v)^4\,, \quad \chi^1(v) = \chi^0(-v) = \frac{1}{16}(1-v)^4\,.$$

Then, for $m \in \{0,1\}$ and $l \in \{1,\dots,L\}$, the elastic energies are given by

$$\mathcal{E}_l^{m,\tau}(v, U_l^m) = \int_\Omega (\chi^m(v) + \tau(1-\chi^m(v)))\, C^m \varepsilon(U_l^m + U_l^{\mathrm{aff},m}) : \varepsilon(U_l^m + U_l^{\mathrm{aff},m})\,\mathrm{d}x\,.$$

Analogously to Lemma 7.2.1, there exist unique displacements $U_l^m(\chi) \in W_\#^{1,2}(\Omega,\mathbb{R}^d)$ minimizing the energy $\mathcal{E}_l^{m,\tau}(v,\cdot)$ and thus solving the linear equation

$$\partial_{U_l^m}\mathcal{E}_l^{m,\tau}(v, U_l^m(v))(U_l^m) = 0 \quad \forall U_l^m \in W_\#^{1,2}(\Omega,\mathbb{R}^d)\,. \tag{7.9}$$

We observe that the cost for a specific subdomain can be written in dependence of the equilibrium displacement $U_l^m(v)$ by

$$\begin{aligned}\mathcal{G}^{m,\tau}(v) :&= g^m\left(C^{m,*}(\chi^m)\xi_1^m : \xi_1^m, \dots, C^{m,*}(\chi^m)\xi_L^m : \xi_L^m\right)\\ &= g^m\left(\mathcal{E}_1^{m,\tau}(v, U_1^m(v)), \dots, \mathcal{E}_L^{m,\tau}(v, U_L^m(v))\right)\,.\end{aligned}$$

To approximate the perimeter functional in v, we recall the Modica–Mortola functional [MM77]

$$\mathcal{A}^\epsilon(v) := \frac{1}{2}\int_\Omega \epsilon|\nabla v|^2 + \frac{1}{\epsilon}W(v)\,\mathrm{d}x\,,$$

where ϵ describes the width of the diffused interface between the two subdomains and we set $W(v) := \frac{9}{16}(v^2-1)^2$. Then we replace the perimeter $|D\chi|_{TV}(\Omega)$ by the phase-field energy $\mathcal{A}^\epsilon(v)$. Furthermore, the maximum function is approximated by a smooth function Max_α. In our computations, we choose

$$\mathrm{Max}_\alpha(x,y) := \frac{1}{2}\left(x + y + \sqrt{|x-y|^2+\alpha}\right) \tag{7.10}$$

for a small $\alpha > 0$. Altogether, we define a cost functional in terms of v as

$$\begin{aligned}\mathcal{J}_{\mathrm{tot}}^{\eta,\tau}(v) &= \mathrm{Max}_\alpha\left(\mathcal{G}^{0,\tau}(v)\,,\,\mathcal{G}^{1,\tau}(v)\right) + \eta\mathcal{A}^\epsilon(v)\\ &= \mathcal{J}_{\mathrm{expl}}^{\eta,\tau}\left(v, U_1^0(v),\dots,U_L^0(v),U_1^1(v),\dots,U_L^1(v)\right)\,,\end{aligned}$$

where a cost functional $\mathcal{J}_{\mathrm{expl}}^{\eta,\tau}$ explicitly depending on phase-fields and displacements is given by

$$\begin{aligned}&\mathcal{J}_{\mathrm{expl}}^{\eta,\tau}\left(v, U_1^0,\dots,U_L^0,U_1^1,\dots,U_L^1\right)\\ &= \mathrm{Max}_\alpha\left(g^0(\mathcal{E}_1^{0,\tau}(v,U_1^0),\dots,\mathcal{E}_L^{0,\tau}(v,U_L^0))\,,\,g^1(\mathcal{E}_1^{1,\tau}(v,U_1^1),\dots,\mathcal{E}_L^{1,\tau}(v,U_L^1))\right) + \eta\mathcal{A}^\epsilon(v)\,.\end{aligned}$$

Now, our numerical algorithm to compute a (local) minimizer of the cost functional $\mathcal{J}_{\mathrm{tot}}^{\eta,\tau}$ requires to compute the first derivative.

Lemma 7.3.1 (Computation of the Shape Derivative)**.** *The derivative of $\mathcal{J}_{tot}^{\eta,\tau}$ along a direction $\hat{v} \in W^{1,2}(\Omega)$ is given by*

$$\frac{d}{dv}\mathcal{J}_{tot}^{\eta,\tau}(v)(\hat{v}) = \partial_v\mathcal{J}_{expl}^{\eta,\tau}\left(v, U_1^0(v),\dots,U_L^0(v),U_1^1(v),\dots,U_L^1(v)\right)(\hat{v})\,.$$

Proof. First, we have that

$$\begin{aligned}\frac{d}{dv}\mathcal{J}_{\mathrm{tot}}^{\eta,\tau}(v)(\hat{v}) &= \partial_v\mathcal{J}_{\mathrm{expl}}^{\eta,\tau}\left(v, U_1^0(v),\dots,U_L^0(v),U_1^1(v),\dots,U_L^1(v)\right)(\hat{v})\\ &+ \sum_{m=0,1}\sum_{l=1}^{L}\partial_{U_l^m}\mathcal{J}_{\mathrm{expl}}^{\eta,\tau}\left(v, U_1^0(v),\dots,U_L^0(v),U_1^1(v),\dots,U_L^1(v)\right)\left(\partial_v U_l^m(v)(\hat{v})\right)\,.\end{aligned}$$

Then we make use of the solutions $A_l^m \in W_\#^{1,2}(\Omega, \mathbb{R}^d)$ to the adjoint problems

$$\partial^2_{U_l^m, U_l^m} \mathcal{E}_l^{m,\tau} \left(v, U_l^m(v)\right) (\widehat{U_l^m})(A_l^m) = -\partial_{U_l^m} \mathcal{J}_{\text{expl}}^{\eta,\tau} \left(v, U_1^0(v), \dots, U_L^0(v), U_1^1(v), \dots, U_L^1(v)\right) (\widehat{U_l^m}) \tag{7.11}$$

for all $\widehat{U_l^m} \in W_\#^{1,2}(\Omega, \mathbb{R}^d)$, which allows to compute

$$\frac{d}{dv} \mathcal{J}_{\text{tot}}^{\eta,\tau}(v)(\hat{v}) = \partial_v \mathcal{J}_{\text{expl}}^{\eta,\tau}(v, U_1^0(v), \dots, U_L^0(v), U_1^1(v), \dots, U_L^1(v))(\hat{v}) + \sum_{m=0,1} \sum_{l=1}^{L} \partial^2_{v, U_l^m} \mathcal{E}_l^{m,\tau}(v, U_l^m(v))(A_l^m)(\hat{v})\,.$$

Now, since $\partial_{U_l^m} \mathcal{E}_l^{m,\tau}(v, U_l^m(v)) = 0$, we observe for the right hand side of the adjoint equation (7.11) that

$$\begin{aligned}
&\partial_{U_l^m} \mathcal{J}_{\text{expl}}^{\eta,\tau} \left(v, U_1^0(v), \dots, U_L^0(v), U_1^1(v), \dots, U_L^1(v)\right) (\widehat{U_l^m}) \\
=&\partial_m \mathrm{Max}_\alpha \left(g^0(\mathcal{E}_1^{0,\tau}(v, U_1^0(v)), \dots, \mathcal{E}_L^{0,\tau}(v, U_L^0(v)))\,,\; g^1(\mathcal{E}_1^{1,\tau}(v, U_1^1(v)), \dots, \mathcal{E}_L^{1,\tau}(v, U_L^1(v)) \right) \\
&\quad Dg^m \left(\mathcal{E}_1^{m,\tau}(v, U_1^m(v)), \dots, \mathcal{E}_L^{m,\tau}(v, U_L^m(v))\right) \partial_{U_l^m} \mathcal{E}_l^{m,\tau}(v, U_l^m(v)) \\
=&0\,.
\end{aligned}$$

Thus, we can conclude that $A_l^m = 0$ for all adjoint solutions. Consequently, the derivative of the cost functional simplifies to

$$\frac{d}{dv} \mathcal{J}_{\text{tot}}^{\eta,\tau}(v)(\hat{v}) = \partial_v \mathcal{J}_{\text{expl}}^{\eta,\tau}(v, U_1^0(v), \dots, U_L^0(v), U_1^1(v), \dots, U_L^1(v))(\hat{v})\,.$$

□

For the numerical discretization in 3D ($d = 3$), we use a cuboid mesh, *i.e.*, the unit cube Ω is uniformly divided into $(N-1)^3$ cuboid elements with N^3 nodes. On this mesh, we define the space $\mathcal{V}_h$ of piecewise trilinear, continuous functions. Then we consider discrete phase-fields $v_h \in \mathcal{V}_h$ and discrete displacement $U_{l,h}^m \in \mathcal{V}_h^3$. In analogy to the continuous case, we restrict to the space of discrete, affine periodic functions. Furthermore, the elastic energies are approximated by a tensor product Simpson quadrature. To implement the periodicity, we identify the nodal values of the discrete phase-field and the discrete displacements on corresponding pairs of nodes.

Concerning the solver, the average value conditions on $U_{l,h}^m$ are imposed via a Lagrange multiplier approach. The corresponding linear systems for the elasticity problems (7.9) are solved using the conjugate gradient method with diagonal preconditioning. Solving the adjoint equations (7.11) is not necessary, since we have already figured out that all adjoint solutions are zero.

The actual shape optimization problem in the unknown phase-field v_h is solved by using the `IPOPT` package [WB06]. Therefore we provide an implementation of the cost functional $\mathcal{J}_{\text{tot}}(v_h)$ and its first derivative. Moreover, the `IPOPT` solver allows incorporating the pointwise constraints $-1 \leqslant v_h(x) \leqslant 1$ for all nodes x and the center of mass condition $\int_\Omega \frac{v_h+1}{2}(x_i - \frac{1}{2})\, dx = 0$ for $i = 1, 2, 3$.

7.4 Numerical Results for Optimal Periodic Microcells

In the following, we present our computational results for optimal microstructures in 3D. Especially, we study different load scenarios and the influence of the material parameters.

First, we comment on the choice of the various parameters, which we have to determine for our numerical scheme. We always initialize the phase-field with random values in the interval $[-1, 1]$ on a mesh with 17^3 vertices. Then the solution on this coarse mesh is prolongated to a finer mesh, where it is used as an initialization. Here, all results are computed on a mesh with 65^3 vertices. For a grid size h, we choose $\epsilon = 2h$ for the phase-field parameter in the Modica–Mortola functional, and the penalty parameter is set to $\eta = 2$. For the ersatz material approach, we choose on the complementary set a factor $\tau = 10^{-4}$. The exponent in the weight function g is chosen to be $p = 2$. For the smooth approximation of the maximum function in (7.10), we choose $\alpha = 10^{-5}$.

We take into account several load scenarios by investigating different combinations of compression and shear loads, where we choose for both subdomains the same loads. More precisely, the corresponding affine displacements are given by $U^{\mathrm{aff},m}(x) = \xi x$ for a symmetric matrix $\xi \in \mathbb{R}^{3\times3}_{\mathrm{sym}}$. We denote by $\{e_1, e_2, e_3\}$ the canonical basis in $\mathbb{R}^3$. Then, for some $\beta \in \mathbb{R}$, compression loads are given by $\xi_{ii} = \beta e_i^T e_i$ and shear loads by $\xi_{ij} = \beta(e_i^T e_j + e_j^T e_i)$. Here, we choose $\beta = -0.25$. Then, we compute the corresponding components of the homogenized elasticity tensors by $C^{m,*}_{iiii} = \beta^{-2} C^{m,*} \xi_{ii} : \xi_{ii}$ (compressive stresses caused by compressive strains) and $C^{m,*}_{ijij} = \beta^{-2} C^{m,*} \xi_{ij} : \xi_{ij}$ (shear strains induced shear stresses).

7.4.1 Different Load Scenarios for Equal Material Parameters

First, we consider equal material parameters $(E^0, \nu^0) = (10, 0.25) = (E^1, \nu^1)$. In Figure 7.2, three different load scenarios are compared:

1. three compression modes ($C^{m,*}_{1111}$, $C^{m,*}_{2222}$, $C^{m,*}_{3333}$),
2. two compression modes combined with a single shear mode ($C^{m,*}_{1111}$, $C^{m,*}_{2222}$, $C^{m,*}_{2323}$),
3. and one compression mode combined with two shear modes ($C^{m,*}_{1111}$, $C^{m,*}_{1212}$, $C^{m,*}_{1313}$).

We observe significant differences in the components of the objective functional. Indeed, those entries of the effective elasticity tensor present in the objective functional indicate a substantially stronger stiffness. Nevertheless, in all cases, the interface between the two subdomains is of the same topology as the Schwarz P surface. Especially in the case of three compression modes, the interface also seems geometrically very close to the Schwarz P surface. For our numerical discretization, we compare an approximation of the Schwarz P surface given as the discrete minimizer of the phase-field area functional $\mathcal{A}^\epsilon$. We obtain values $C^{m,*}_{iiii} = 2.7811$ ($i = 1, 2, 3$) and $C^{m,*}_{ijij} = 2.481$ ($i, j = 1, 2, 3,\ i \neq j$), which significantly differ compared to the optimizer for three compression modes and a difference of approximately 3% for the phase-field area functional $\mathcal{A}^\epsilon$. In the literature [TD04, Sil07], the subdomain splitting associated with the Schwarz P surface as the interface has been investigated concerning its optimality in the context of PDE constrained optimization for a scalar-valued problem.

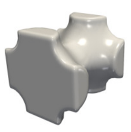
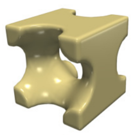
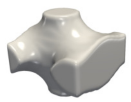
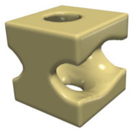
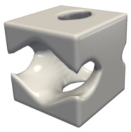
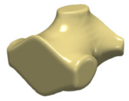
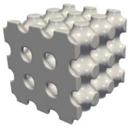
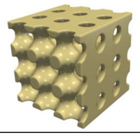
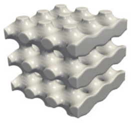
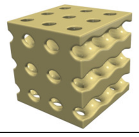
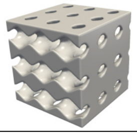
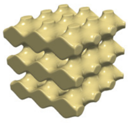
	3× compr		2× compr, 1× shear		1× compr, 2× shear	
single cell						
3^3 cells						
	m=0	m=1	m=0	m=1	m=0	m=1
$C^{m,*}_{1111}$	2.825	2.825	2.3657	2.3657	3.745	3.745
$C^{m,*}_{2222}$	2.825	2.825	3.8584	3.8584	2.3035	2.3035
$C^{m,*}_{3333}$	2.825	2.825	2.1651	2.1651	2.3035	2.3035
$C^{m,*}_{1212}$	2.4851	2.4851	3.0126	3.0126	2.8256	2.8256
$C^{m,*}_{1313}$	2.4851	2.4851	1.1134	1.1134	2.8256	2.8256
$C^{m,*}_{2323}$	2.4851	2.4851	2.7998	2.7998	1.6268	1.6268
volume	0.5	0.5	0.5	0.5	0.5	0.5

Figure 7.2: Comparison of optimal microstructures and relevant induced components of the effective elasticity tensors for different load scenarios indicated above. In the top row we depict the subdomains on the fundamental cell of the microstructure and below a 3 × 3 × 3 composition pronouncing the periodicity. Those components of the tensor which are part of the corresponding objective functional are highlighted in grey.

7.4.2 Influence of the Perimeter Term

Next, in Figure 7.3, we show the effect of the perimeter functional by considering different values for the penalty parameter $\eta = 2, 4, 10$. Here, we investigate a load scenario with three shear loads ($C^{m,*}_{1212}, C^{m,*}_{1313}, C^{m,*}_{2323}$). For small η, we obtain a laminate type optimal configuration, whereas, for larger η, the interface is again similar to the Schwarz P surface. On the intermediate range of the parameter η, we obtain an optimal microstructure with an interface similar to a gyroid minimal surface, which is also taken into account as a possible microstructure in bone tissue engineering in [KHM$^+$11]. We observe that this intermediate range is comparatively small for the load scenario with three shear loads. For the other load scenarios studied in Figure 7.2, we also obtain an interface similar to the Schwarz P surface for large η, but for smaller values $\eta = 0.1$, the numerical optimization scheme still converges to similar solutions. Depending on the initialization, for even smaller values $\eta = 0.001$, the method does not converge because of a lack of regularization, but laminate structures never appear. This indicates that for the load scenario with three shear loads, the optimal solution is indeed a (nested) laminate structure.

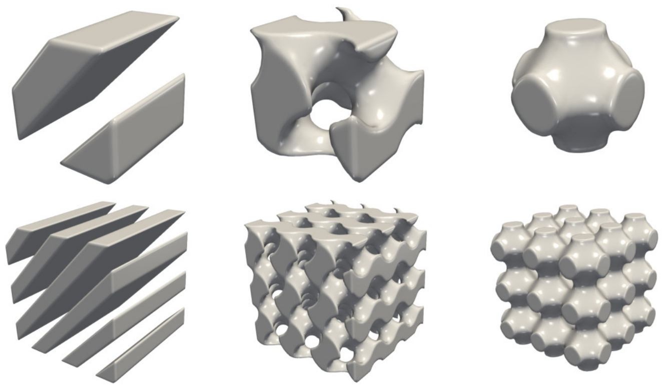


Figure 7.3: Optimal microstructures for different values of the perimeter parameter η (from left to right: $\eta = 2, 4, 10$). In the top row we depict a single fundamental cell and below a $3 \times 3 \times 3$ block.

7.4.3 Influence of Weighting Function

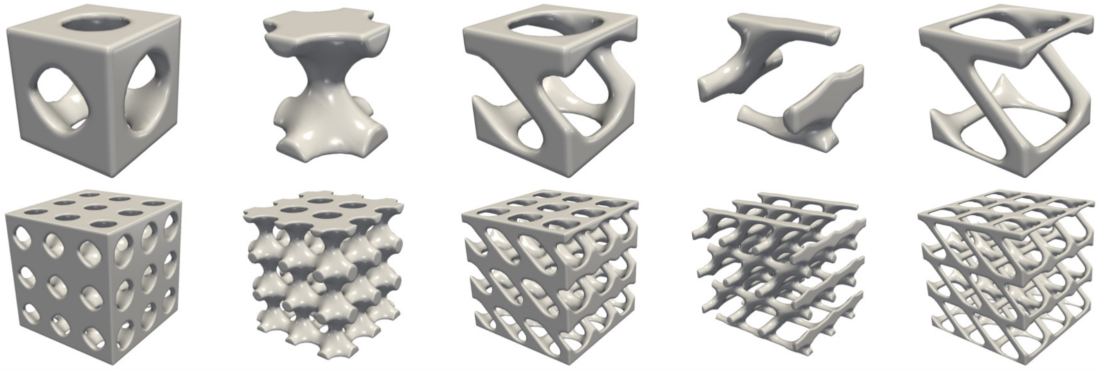
So far, for the weight function g, we have always chosen $p = 2$. In Table 7.1, we show for the load scenario with two compression loads and one shear load the relevant entries of the effective elasticity tensor. For increasing p, we observe a successive balancing of the different components of the objective functional. In particular, the largest component $C^{m,*}_{2222}$ of the effective elasticity tensor is slightly decreasing, while the smallest component $C^{m,*}_{1111}$ is slightly increasing.

p	2		4		8		16	
	m=0	m=1	m=0	m=1	m=0	m=1	m=0	m=1
$C^{m,*}_{1111}$	2.3657	2.3657	2.4438	2.4384	2.4847	2.4808	2.5053	2.5056
$C^{m,*}_{2222}$	3.8584	3.8584	3.8408	3.8429	3.8286	3.8291	3.8286	3.828
$C^{m,*}_{2323}$	2.7998	2.7998	2.6764	2.6857	2.6139	2.6206	2.5768	2.5766

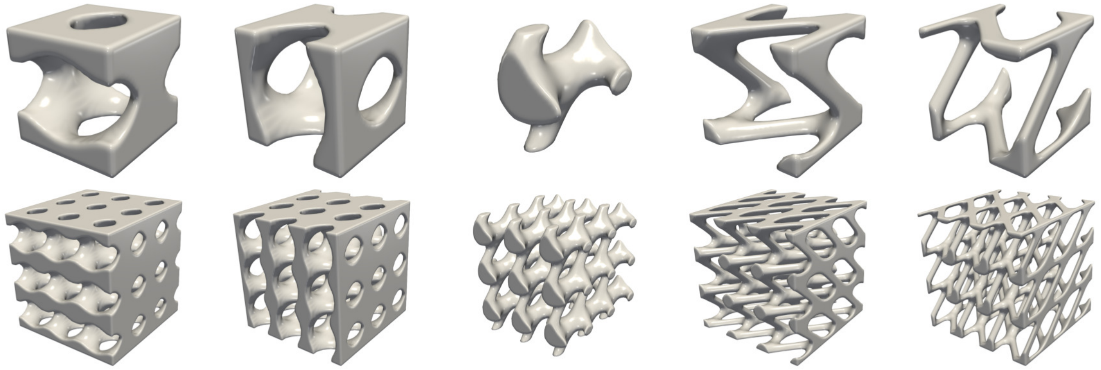
Table 7.1: Stiffness moduli of the optimal subdomain splitting for different values of p.

7.4.4 Varying Young's Modulus

Next, we study the influence of Young's modulus by considering $E^0 = 20, 40, 80, 160, 320$, where we always choose $(E^1, \nu^1) = (10, 0.25)$ and $\nu^0 = 0.25$. We observe that the structures become thinner for the subdomain with increasing values of Young's modulus, since the difference in stiffness of the materials has to be compensated by a higher volume fraction of the other subdomain. In Figure 7.4, we show results obtained for different load scenarios.



	m=0	m=1	m=0	m=1	m=0	m=1	m=0	m=1	m=0	m=1
$C^{m,*}_{1111}$	3.6587	3.3904	4.7996	4.2686	6.413	5.9585	6.837	7.1062	8.8662	8.6567
$C^{m,*}_{2222}$	4.9667	4.3946	7.0545	5.4043	7.6653	6.7548	7.2202	7.4892	9.3866	9.0007
$C^{m,*}_{3333}$	3.5702	3.4632	5.0673	4.6461	2.3783	5.9513	3.9501	7.6961	2.562	8.4978
$C^{m,*}_{1212}$	3.2681	4.0275	3.3935	5.4312	2.7089	8.1685	0.67487	9.0034	1.2522	10.1784
$C^{m,*}_{1313}$	2.2299	3.1636	2.3708	4.6865	1.006	6.5492	1.1444	9.4359	0.46034	10.0924
$C^{m,*}_{2323}$	3.5124	4.1468	4.283	5.8455	6.0737	8.1456	9.7432	10.0662	8.8804	10.1808
vol	0.41037	0.58963	0.32744	0.67256	0.22565	0.77435	0.15367	0.84633	0.10194	0.89806



	m=0	m=1	m=0	m=1	m=0	m=1	m=0	m=1	m=0	m=1
$C^{m,*}_{1111}$	6.3411	4.9912	9.6988	5.8266	10.39	6.7704	10.264	8.0092	10.773	9.1178
$C^{m,*}_{2222}$	3.103	2.8423	4.3565	3.6762	3.2102	5.2916	2.4592	7.0516	2.6306	8.3817
$C^{m,*}_{3333}$	3.1044	2.8416	4.4754	3.6902	3.2021	5.2907	2.4592	7.0516	2.6306	8.3817
$C^{m,*}_{1212}$	3.9041	4.1364	4.8936	5.4934	6.4294	7.7477	8.0385	10.026	9.2991	11.802
$C^{m,*}_{1313}$	3.9046	4.1362	4.8983	5.5215	6.4298	7.7457	8.0385	10.026	9.2991	11.802
$C^{m,*}_{2323}$	1.2109	1.7901	1.2961	2.7966	1.0188	5.1458	0.5915	8.1734	0.36466	10.502
vol	0.41917	0.58083	0.3454	0.6546	0.25082	0.74918	0.16297	0.83703	0.10518	0.89482

Figure 7.4: Comparison of optimal microstructures for varying values of Young's modulus (from left to right $E^0 = 20, 40, 80, 160, 320$). We take into account load configurations with two compression loads and one shear load (top) and one compression load and two shear loads (bottom). We depict the subdomain $\mathcal{O}^0$ on the fundamental cell of the microstructure and a $3 \times 3 \times 3$ composition. Those components of the tensor which are part of the corresponding objective functional are again highlighted in gray.

7.4.5 Realistic Material Parameters for Bone and Polymer

Now, for the actual application to optimize the polymer scaffold, we remark that real bone is substantially stiffer than the bioresorbable polymer with a 15 times larger value of Young's modulus, and realistic Poisson ratios are $\nu^B = 0.1$ for bone and $\nu^P = 0.3$ for the polymer. In Figure 7.5, we show the optimal bone and polymer subdomains for a load scenario with one compression load and two shear loads, which corresponds to possible movements of a tibia. Furthermore, for each load, we plot the von Mises stresses on the boundary of the corresponding subdomains in the fundamental cell, which are given by $\sigma_l^{vM} = \sqrt{\frac{1}{2}\sum_{1\leqslant i<j\leqslant 3}(\lambda_l^i - \lambda_l^j)^2}$, where $\lambda_l^1, \lambda_l^2, \lambda_l^3$ are the eigenvalues of the linear stress tensor $\sigma_l = (\chi(v) + \tau(1 - \chi(v)))C\varepsilon(U_l)$.

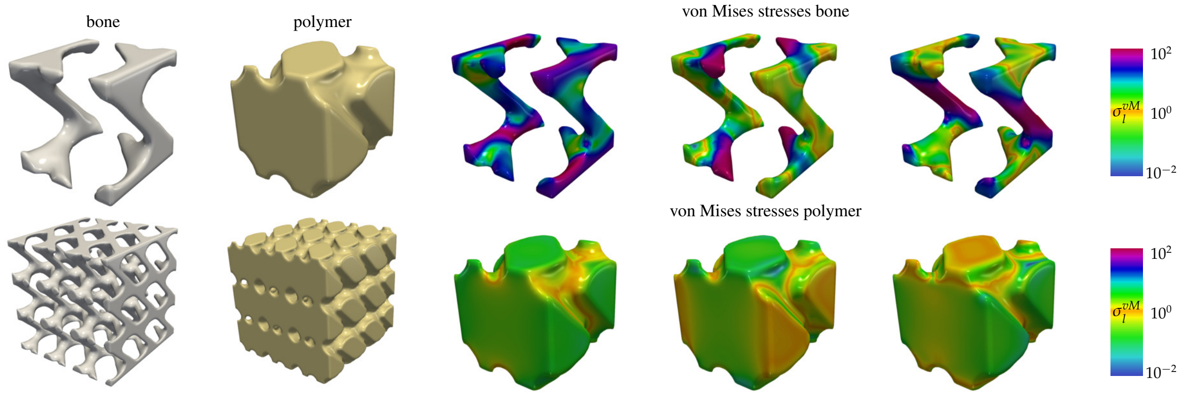


Figure 7.5: Optimal bone and polymer microstructures for realistic material parameters and a load scenario with one compression and two shear loads. Fore each load, we show the corresponding von Mises stresses color-coded in an HSV model with a logarithmic scale.

7.4.6 The Two-Dimensional Case

We briefly comment on the 2D case ($d = 2$). In Figure 7.6, we show the numerical result for a scenario with two uniaxial compression loads. The optimal domain splitting is given by diamond-shaped regions. Due to the hard-soft approximation, this is a mechanically admissible configuration. For a hard-void shape optimization model and two uniaxial compression loads in the vertical and horizontal direction, no mechanically favorable splitting of the unit square $[0, 1]^2$ into two subdomains is possible. Indeed, a uniaxial load requires a truss with a nonvanishing interior connecting the components of the boundary opposite in the loading direction. A truss configuration simultaneously in the horizontal and vertical direction for both subdomains is thus topologically impossible.

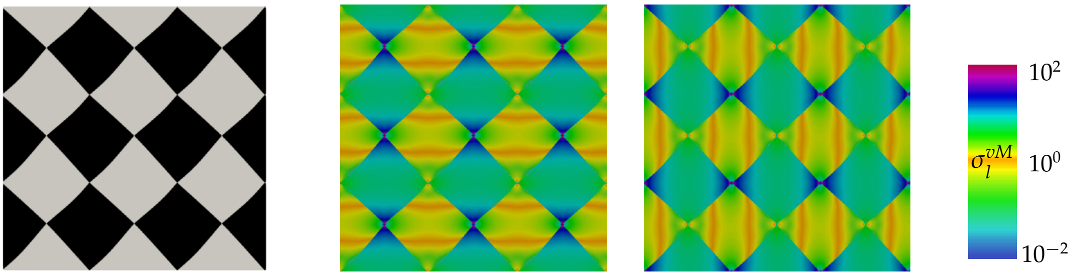


Figure 7.6: For a 2D domain and a hard-soft approximation with $\tau = 10^{-4}$, we depitct an optimal decomposition for a load scenario with two loads corresponding to the compression modes ($C_{1111}^{m,*}$ and $C_{2222}^{m,*}$). A block of 3 × 3 cells is plotted with the two subdomains in white and black together with a color plot of the von Mises stresses.

7.5 Extensions of the Model by Diffusion and Volume Constraints

So far, our proposed scaffold design in Figure 7.5 is optimized w.r.t. the mechanical stability of the scaffold itself and the complementary set filled with bone. Nevertheless, the very low porosity would seriously impede vascularization and therefore prevent the regeneration of the bone. Therefore we propose to extend our model by additionally enforcing either a volume constraint or a diffusion constraint for the complementary set of the scaffold.

Volume Constraint. More precisely, for a volume constraint, we minimize the objective functional $\mathcal{J}_{\text{tot}}^{\eta,\tau}$ defined in (7.8) over all $\chi \in BV_{\#,c}(\Omega, \{0,1\})$ with $\int_\Omega \chi \, \mathrm{d}x \geqslant V^0$, meaning that $V^0 \in (0,1)$ is a lower bound for the volume fraction of the corresponding domain $\mathcal{O}^0$.

Diffusion Constraint. For a diffusion constraint, we take into account linear diffusion with a scalar-valued diffusion coefficient $a^m \in \mathbb{R}$. Then, similar to linear elasticity (7.4), it is well-known from the theory of periodic homogenization [All02] that the homogenized diffusion tensor $A^{m,*} \in \mathbb{R}^{d \times d}$ of the resulting microstructure is uniquely described by

$$A^{m,*}F \cdot F = \min_{f^m \in W^{1,2}_\#(\Omega)} \int_\Omega \chi^m a^m \; (F + \nabla f^m) \; \cdot \; (F + \nabla f^m) \; \mathrm{d}x$$

for all $F \in \mathbb{R}^d$. Then we impose that certain entries of $A^{0,*}$ are bounded from below, s.t. we can guarantee a transfer in the corresponding direction.

Adaption of the Numerical Optimization Method. A description of both constraints in our phase-field model is straightforward. Furthermore, the IPOPT package is capable to include these constraints if we can provide $\int_\Omega v \, \mathrm{d}x$, $A^{m,*}$, and the corresponding derivatives in the phase-field variable v.

For our numerical simulations, we take into account the load scenario and the material parameters as in Figure 7.5. In Figure 7.7, we study the effect for different volume constraints. For the diffusion constraints, we choose $a^m = 1$ for the corresponding coefficients. In Figure 7.8, we depict the results for a single diffusion constraint on the entry $A^{*,0}_{11}$ of the homogenized diffusion tensor, and in Figure 7.9, we incorporate three diffusion constraints on the entries $A^{*,0}_{ii}$ simultaneously. Indeed, both approaches lead to a larger porosity.

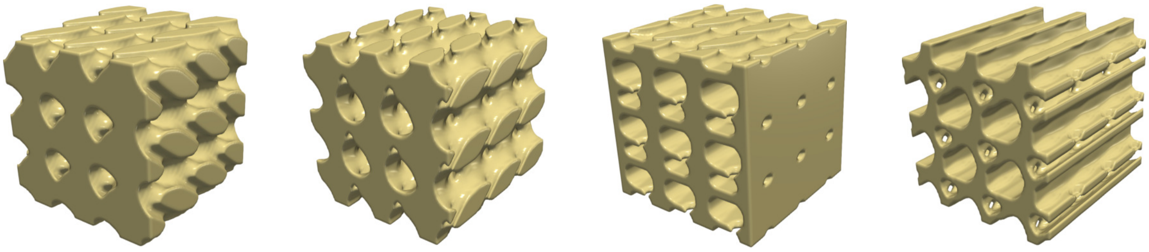
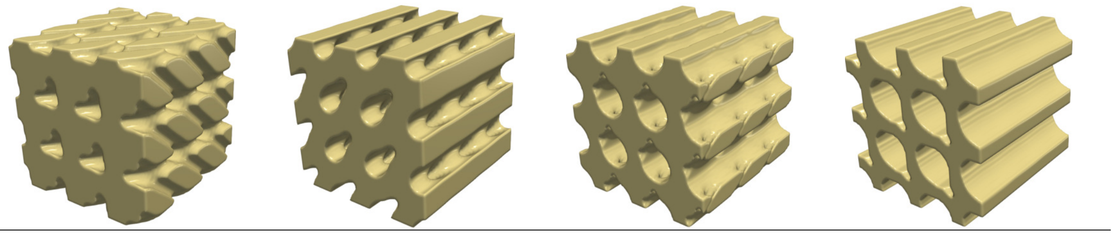


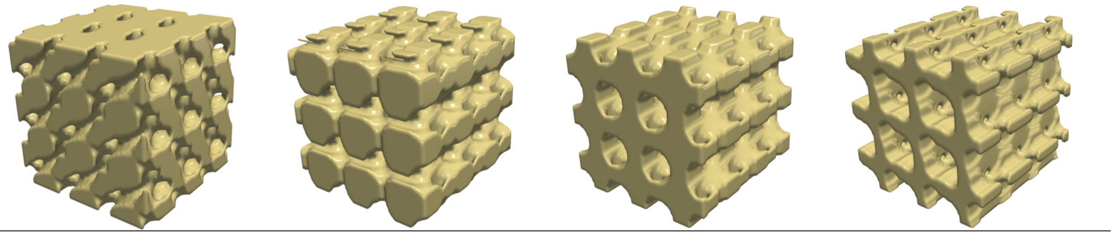
Figure 7.7: Optimal microstructures for increasing volume constraints $\int_\Omega \chi \, \mathrm{d}x \geqslant V^0$ with $V^0 = 0.2, 0.3, 0.4, 0.5$.

We finally remark that our model does not contain a specific size for a single microcell, since the involved theory of homogenization has to be understood as a limiting model with cell size converging to zero. Additional thickness constraints on the domain were included via a level set approach in [AJM16]. Now, in our context, the definition that $\mathcal{O}^0$ has a thickness larger than a constant c can be interpreted as $\mathcal{O}^0$ has pore size larger than c. Thus, to fix a precise size for our microcell, we propose to compute on the reference object $[0,1]^d$ the smallest value c s.t. a certain thickness constraint is guaranteed. Then if a specific pore size on the microcell is required to allow vascularization, we can scale the reference object accordingly.



	m=0	m=1	m=0	m=1	m=0	m=1	m=0	m=1
$C^{m,*}_{1111}$	11.699	8.7464	28.639	7.9429	44.299	6.9517	59.643	5.9105
$C^{m,*}_{2222}$	2.2194	7.379	1.0122	5.3709	1.652	4.5596	0.62499	3.46
$C^{m,*}_{3333}$	2.2164	7.3773	1.0121	5.3041	1.9344	4.5951	0.62558	3.4585
$C^{m,*}_{1212}$	9.2414	9.7276	7.4141	8.1408	5.9764	6.5545	5.1434	5.3435
$C^{m,*}_{1313}$	9.2424	9.7271	7.4139	8.0834	6.2822	6.6355	5.1432	5.3429
$C^{m,*}_{2323}$	0.407	7.681	3.9817	6.5508	0.51959	3.3189	2.3766	2.1144
vol	0.16782	0.83218	0.25479	0.74521	0.34192	0.65808	0.4379	0.5621
$\mathcal{J}/\mathcal{J}_0$	1.344		1.528		1.773		2.159	

Figure 7.8: Optimal microstructures for increasing diffusion constraints $A^{*,0}_{11} \geqslant \alpha$ with $\alpha = 0.1, 0.2, 0.3, 0.4$. In the last row we compare the relative increase of the total cost functional $\mathcal{J}^{\eta,\tau}_{\text{tot}}$ compared to the result in Figure 7.5 without any constraints.



	m=0	m=1	m=0	m=1	m=0	m=1	m=0	m=1
$C^{m,*}_{1111}$	7.6929	6.3258	27.732	6.4087	43.756	5.9029	59.169	3.7485
$C^{m,*}_{2222}$	9.102	4.4464	13.841	4.0532	13.539	3.4021	12.942	1.9032
$C^{m,*}_{3333}$	9.4654	4.9992	13.342	4.1119	13.129	3.6912	12.713	1.7949
$C^{m,*}_{1212}$	12.392	7.2419	8.2786	6.457	14.148	5.3252	19.611	3.1959
$C^{m,*}_{1313}$	12.143	7.3398	9.1809	6.3218	11.098	5.3953	15.817	2.904
$C^{m,*}_{2323}$	7.0137	5.7203	2.4781	3.6596	3.9284	2.5161	5.9591	0.8262
vol	0.25404	0.74596	0.30301	0.69699	0.37452	0.62548	0.52089	0.47911
$\mathcal{J}/\mathcal{J}_0$	1.843		1.910		2.109		3.377	

Figure 7.9: Optimal microstructures for diffusion constraints $A^{*,0}_{11} \geqslant \alpha$, $A^{*,0}_{22} \geqslant 0.1$, $A^{*,0}_{33} \geqslant 0.1$ with $\alpha = 0.1, 0.2, 0.3, 0.4$. In the last row we compare the relative increase of the total cost functional $\mathcal{J}^{\eta,\tau}_{\text{tot}}$ compared to the result in Figure 7.5 without any constraints.

7.6 Conclusion and Outlook

Motivated by a biomechanical application of designing optimal polymer scaffolds for bone regeneration, we have proposed an elastic shape optimization problem by taking into account the homogenized elasticity tensors of the domain with the polymer implant and the complementary set, where new bone tissue growths first. Compared to [KHM$^+$11], where minimal surfaces were proposed as possible scaffolds, we have obtained significantly different structures, in particular, for a realistic load scenario with one compression and two shear loads. Furthermore, we have investigated additional volume and diffusion constraints, where especially the latter one appears to be biomechanically relevant.

We briefly discuss a possible extension of our model. So far, we have assumed that the microstructure of the polymer scaffold is periodic and thus, we have optimized only a single microcell. More generally, we could consider realistic patient-specific implant geometries on the macroscale, where the polymer implant has to be inserted, and then we could ask for the optimal scaffold by allowing each microcell to vary. Compared to the numerical implementation of a two-scale model as it was considered in [CGRS14, GR16, CGLR17] and where on each quadrature point of the macroscopic grid a microcell was adapted, in the application for bone tissue engineering, the computed object must be printable by a 3D printer. First steps into that direction were established in [Sch19]. There, the microcells were considered on a (cuboid) element of the macroscopic grid, and the printability condition was incorporated via Dirichlet boundary conditions on the microcells. Then the optimization scheme consisted of an alternating update of the (homogenized) elasticity tensors on the macroscale and the optimal design on certain blocks of microcells for the displacements on the macroscale. However, the concept of the homogenized elasticity tensor was not precisely reflected in the discretization, since the microstructures were considered on a fixed scale instead of quadrature points. Alternatively, the displacements on the macroscale could be computed by taking into account the full grid containing all degrees of freedom. For practical applications, this would imply a huge grid size, s.t. solving the corresponding linear systems requires, *e.g.*, multigrid methods.

Finally, the regeneration of bone and degradation of the polymer implant is highly complex in reality. Here, we have supposed that three phases can describe this dynamic process. More precisely, we have assumed that first, the implant is inserted. Subsequently, new bone tissue grows into the void part while the implant is still present, and afterwards, the polymer starts to degrade. Certainly, this is a substantial simplification. A first time-dependent model was proposed for a one-dimensional space domain in [PVB$^+$18], where the minimal value of the effective mechanical stiffness over the regeneration time was maximized.

Chapter 8

Shape Design of Thin Elastic Objects

In Chapter 6, we have described deformations of elastic bodies as solutions to suitable partial differential equations. Here, we focus on a special class of so-called thin elastic objects, which can be characterized by a small thickness and a regular and orientable two-dimensional midsurface. Considering the limit of vanishing thickness, Γ-convergence results have been established to express the 3D deformation of the thin object only by a 2D deformation of its midsurface. A membrane theory describes tangential distortion on the surface, and a bending theory takes into account isometric deformations. Computing such deformations numerically has been intensively studied in the literature, where numerous discretization approaches have been applied, in particular, for models combining membrane and bending energy functionals. Pure bending isometries of plates have been numerically approximated in [Bar13] by making use of the discrete Kirchhoff triangle (DKT) element. Furthermore, in Chapter 6, we have discussed certain shape optimization problems to optimize the material distribution on the reference domain of an elastic body to guarantee maximal mechanical stability w.r.t. an external force.

In this chapter, we study shape optimization problems to optimize the material distribution on a thin elastic object, where we, for simplicity, restrict to parametric surfaces. We consider a load scenario only consisting of a single force acting on the thin elastic object. To describe deformations, we take into account different types of elastic energies, in particular, we deal with nonlinear elasticity. Then a numerical discretization scheme to compute equilibrium deformations is based on the discrete Kirchhoff triangle element. A special focus is on pure bending isometries, which we can efficiently approximate due to the degrees of freedom for derivatives at nodal positions similar to the approach in [Bar13]. For a total cost functional depending on the material distribution, we consider the potential energy and enforce a constraint on the amount of hard material. Moreover, we apply a phase-field model and use the Modica–Mortola functional to penalize the width of the diffuse interface between the hard and soft subdomains.

This chapter is organized as follows. First, in Section 8.1, we define thin elastic shells for parametric surfaces and derive certain state equations. Moreover, we recall the discrete Kirchhoff triangle element. In Section 8.2, we study shape optimization problems for both linear and nonlinear elasticity, where the stored elastic energy consists of a membrane and a bending energy part. Furthermore, we investigate shape optimization problems for pure bending isometries. In Section 8.3, we consider a one-dimensional model of elastic beams in 2D, s.t. an isometric deformation can be expressed in terms of the phase, which simplifies the corresponding state equation to an unconstrained ordinary differential equation. In [HRS19], we used this reformulation to compute the optimal material distribution in a special setting explicitly. Here, we summarize this theoretical classification result, which is confirmed and extended to more general scenarios by our numerical simulations. Finally, in Section 8.4, we consider isometric deformations of two-dimensional objects and obtain different optimal designs even though we apply a one-dimensional boundary condition.

Remark 8.0.1 (Collaborations and Publications). The results presented in Section 8.3 are joint work with Peter Hornung and Martin Rumpf and have been published in [HRS19].

8.1 Thin Elastic Shells

Here, we introduce thin elastic shells and refer to [Cia08, CM08] for a comprehensive overview. Roughly speaking, a thin elastic shell is an elastic body in $\mathbb{R}^3$, which can be described by a 2D surface (the midsurface) and a thickness $\delta > 0$. Thus, we first recall some basics from differential geometry, where we restrict to parametric surfaces. Considering the limit $\delta \to 0$, we are interested to understand the 3D deformation Φ of the elastic object just by a 2D deformation of its midsurface. Then we discuss several discretization methods to compute equilibrium deformations of the corresponding state equations. Finally, we recall the discrete Kirchhoff triangle element, which we take into account for the shape optimization problems in this chapter.

8.1.1 Differential Geometry for Parametric Surfaces

In the following, we introduce basic differential geometric objects and especially focus on expressing these objects on the chart domain. For a general introduction to Riemannian geometry, we refer the reader to [dC92]. Here, we restrict to two-dimensional embedded surfaces in $\mathbb{R}^3$ and refer the reader to [Bär01].

We consider a manifold $\mathcal{M} = \psi(\omega)$ that is given as the image of a single chart $\psi\colon \omega \to \mathbb{R}^3$, where $\omega \subset \mathbb{R}^2$ is an open and bounded domain with Lipschitz boundary. For the moment, we assume that $\psi \in C^2(\overline{\omega}, \mathbb{R}^3)$, but later we discuss the regularity assumptions on ψ more precisely. We denote by $\xi \in \omega$ coordinates in the chart domain and by $p = \psi(\xi) \in \mathcal{M}$ coordinates on the manifold. Furthermore, ψ is assumed to be an injective immersion, *i.e.*, for all $\xi \in \omega$, the two vectors $\partial_1\psi(\xi)$ and $\partial_2\psi(\xi)$ are linearly independent and span the tangent space $T_p\mathcal{M} = \operatorname{span}(\partial_1\psi(\xi), \partial_2\psi(\xi))$ at $p = \psi(\xi)$. Thus, the unit normal at p is given by

$$n(p) = n(\psi(\xi)) = \frac{\partial_1\psi(\xi) \times \partial_2\psi(\xi)}{|\partial_1\psi(\xi) \times \partial_2\psi(\xi)|}\,.$$

First Fundamental Form. In general, we say that $\mathcal{M}$ is a Riemannian manifold, if for each $p \in \mathcal{M}$ there is a scalar product $g(p)\colon T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$, which is smooth in $p \in \mathcal{M}$. Since in our case, $\mathcal{M}$ is embedded in $\mathbb{R}^3$, we can define the first fundamental form by the Euclidean scalar product $g(p)(V, W) = V \cdot W$ for $V, W \in T_p\mathcal{M} \subset \mathbb{R}^3$. To represent the first fundamental form on the chart domain, we first note that vector fields $V, W\colon \mathcal{M} \to T\mathcal{M} = \{(p, Z) \,:\, p \in \mathcal{M}, Z \in T_p\mathcal{M}\}$ can be expressed in the basis $(\partial_1\psi(\xi), \partial_2\psi(\xi))$ as $V(p) = D\psi(\xi)v(\xi)$, $W(p) = D\psi(\xi)w(\xi)$. Then we define $g(\xi)(v, w) = g(p)(D\psi(\xi)v, D\psi(\xi)w)$ for $v, w \in \mathbb{R}^2$ and obtain

$$g(\xi) = (D\psi(\xi))^T D\psi(\xi) = (g(\xi)_{ik})_{i,k=1,2} = \left(\sum_{j=1,2} \partial_i\psi_j(\xi)\partial_k\psi_j(\xi) \right)_{i,k=1,2}\,.$$

Furthermore, $g(\xi) \in \mathbb{R}^{2\times 2}$ is invertible and we denote its inverse by $g(\xi)^{-1} = \left(g(\xi)^{ik}\right)_{i,k=1,2}$. Note that the first fundamental form admits the integral transformation rule

$$\int_{\mathcal{M}} f(p)\, \mathrm{d}\mathscr{H}^2(p) = \int_\omega \sqrt{\det g(\xi)}\, f \circ \psi(\xi)\, \mathrm{d}\xi$$

for $f \in L^1(\mathcal{M})$ and thus, especially allows measuring the area of the manifold. Next, we introduce certain differential operators on $\mathcal{M}$. First, for a smooth function $f\colon \mathcal{M} \to \mathbb{R}^d$, the differential $df(p)\colon T_p\mathcal{M} \to \mathbb{R}^d$ is given by $df(p)(V) = \frac{d}{dt}(f \circ \gamma)|_{t=0}$, where $\gamma\colon (-\varepsilon, \varepsilon) \to \mathcal{M}$ is a smooth curve satisfying $\gamma(0) = p$ and $\gamma'(0) = V$ and it can be verified that this definition does not depend on γ. Analogously, the differential of a smooth function $f\colon \mathcal{M} \to \mathcal{N}$ onto a manifold $\mathcal{N}$ is defined, which is a mapping between the tangent spaces $df(p)\colon T_p\mathcal{M} \to T_{f(p)}\mathcal{N}$. For a scalar valued function $f\colon \mathcal{M} \to \mathbb{R}$, the gradient $\nabla_{\mathcal{M}} f(p) \in T_p\mathcal{M}$ is defined by the relation $g(p)(\nabla_{\mathcal{M}} f(p), V) = df(p)(V)$ for all $V \in T_p\mathcal{M}$, which leads to

$$\nabla_{\mathcal{M}} f(p) = D\psi(\xi)g(\xi)^{-1}\nabla(f \circ \psi)(\xi)\,.$$

Defining the divergence $\operatorname{div}_{\mathcal{M}}$ as the adjoint operator acting on vector fields $V(p) = D\psi(\xi)v(\xi)$, we obtain

$$\operatorname{div}_{\mathcal{M}}(D\psi(\xi)v(\xi)) = \frac{1}{\sqrt{\det g(\xi)}} \operatorname{div}\left(\sqrt{\det g(\xi)}v(\xi)\right)\,.$$

Then, the Laplace–Beltrami operator is defined by $\Delta_{\mathcal{M}} f(p) = \mathrm{div}_{\mathcal{M}} \nabla_{\mathcal{M}} f(p)$. Moreover, we introduce the Christoffel symbols of first and second kind by

$$\Gamma_{ijk} := \partial^2_{ij}\psi \cdot \partial_k \psi = \frac{1}{2}\left(\partial_j g_{ki} + \partial_i g_{kj} - \partial_k g_{ij}\right) , \quad \Gamma^m_{ij} := \sum_{k=1,2} g^{mk}\Gamma_{ijk} .$$

Second Fundamental Form. Note that the normal can be considered as a vector valued function $n\colon \mathcal{M} \to \mathcal{S}^2 \subset \mathbb{R}^3$. Since $T_{n(p)}\mathcal{S}^2 = (\mathrm{span}(n(p)))^{\perp} = T_p\mathcal{M}$, the differential $S(p) = dn(p)$ at a point $p \in \mathcal{M}$ is thus a linear map $S(p)\colon T_p\mathcal{M} \to T_{n(p)}\mathcal{S}^2 = T_p\mathcal{M}$, which is called the shape operator or Weingarten map. It can be verified that $S(p)$ is self-adjoint w.r.t. the first fundamental form. Then the associated bilinear form $h(p)\colon T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ with

$$h(p)(V, W) := g(p)(S(p)(V), W) = g(p)(V, S(p)(W))$$

is called the second fundamental form and can be represented on the chart domain by a matrix $h(\xi) \in \mathbb{R}^{2\times 2}$ with entries $h_{ij}(\xi) := h(p)(\partial_i\psi(\xi), \partial_j\psi(\xi))$, which leads to

$$h(\xi) = D(n \circ \psi)(\xi) \cdot D\psi(\xi) = -D^2\psi(\xi) \cdot n(\psi(\xi)) .$$

Also the shape operator has a matrix representation $S(\xi) = g(\xi)^{-1}h(\xi) \in \mathbb{R}^{2\times 2}$ on the chart domain, s.t. $S(p)(\partial_j\psi(\xi)) = \sum_{i=1,2} S_{ij}(\xi)\partial_i\psi(\xi)$. Then we call $K(\xi) = \det(S(\xi))$ the Gauss curvature and $H(\xi) = \mathrm{tr}(S(\xi))$ the mean curvature.

Isometric Chart Maps. Next, we consider a special class of chart maps given by isometries. In general, an isometry $\psi\colon \omega \to \mathbb{R}^3$ can be defined as a length-preserving map. Above, we have for simplicity assumed that $\psi\colon \omega \to \mathcal{M}$ is a chart of a Riemannian manifold with C^2 regularity. However, there might be a huge difference to C^1 isometries, which we want to point out.

Definition 8.1.1 (Isometry). A map $\psi \in C^1(\omega, \mathbb{R}^3)$ is called isometry if $g(\xi) = \mathbb{1}_{2\times 2}$ for all $\xi \in \omega$.

Now, the famous Nash–Kuiper theorem states that any short immersion can be uniformly approximated by C^1 isometries.

Theorem 8.1.2 (Nash–Kuiper). *Let $\omega \subset \mathbb{R}^2$ be open and bounded, and let $u \in C^\infty(\overline{\omega}, \mathbb{R}^3)$ with $Du^TDu \leqslant \mathbb{1}_{2\times 2}$ and rank$(Du) = 2$ everywhere. Then for every $\varepsilon > 0$ there exists $\psi \in C^1(\overline{\omega}, \mathbb{R}^3)$ with $D\psi^TD\psi = \mathbb{1}_{2\times 2}$ and $\|u - \psi\|_{L^\infty} < \varepsilon$.*

Proof. See [Nas54], [Kui55]. □

In contrast, for C^2 isometries we have the following properties.

Proposition 8.1.3 (Properties of C^2 Isometries). *Let $\psi \in C^2(\omega, \mathbb{R}^3)$ be an isometry.*

1. *For the Christoffel symbols, we have $\Gamma_{ijk} = \partial_{ij}\psi \cdot \partial_k\psi = 0$ for all $i, j, k = 1, 2$.*

2. *For the Gauss curvature, we have $K = 0$.*

3. *We have equalities $|D^2\psi| = |\Delta\psi| = |h| = |H|$.*

Proof. See [Bar15, Proposition 8.2]. □

Furthermore, the Hartman–Nirenberg theorem states that C^2 isometries behave rigidly in the following sense.

Theorem 8.1.4 (Hartman–Nirenberg). *Let $\omega \subset \mathbb{R}^2$ be open and bounded. Furthermore, let $\psi \in C^2(\omega, \mathbb{R}^3)$ s.t. $D\psi(\xi)^TD\psi(\xi) = \mathbb{1}_{2\times 2}$ for all $\xi \in \omega$. Then ψ is developable,* i.e.*, for any $\xi \in \omega$, one of the following holds:*

1. *There exists $U \subset \omega$ open with $\xi \in U$ and ψ is affine on U.*

2. *There exist $a, b \in \partial\omega$ with $\xi \in [a, b]$ and ψ is affine on the line segment $[a, b]$.*

Proof. See [HN59]. □

A generalization was established by Hornung [Hor11], who proved that isometries $\psi \in W^{2,2}_{\mathrm{iso}}(\omega, \mathbb{R}^3) := \{u \in W^{2,2}(\omega, \mathbb{R}^3) : Du^TDu = \mathbb{1}_{2\times 2}$ *a.e.*$\}$ are developable and can be approximated in the strong $W^{2,2}$-topology by functions in $W^{2,2}_{\mathrm{iso}}(\omega, \mathbb{R}^3) \cap C^\infty(\overline{\omega}, \mathbb{R}^3)$.

Deformations between Parametric Surfaces. Now, we consider two manifolds $\mathcal{M}_A = \psi_A(\omega)$ and $\mathcal{M}_B = \psi_B(\omega)$, which are parametrized over the same chart domain $\omega \subset \mathbb{R}^2$. Then, a deformation between the two manifolds is given by $\phi = \psi_B \circ \psi_A^{-1} \colon \mathcal{M}_A \to \mathcal{M}_B$. In Figure 8.1, we show a sketch of this configuration.

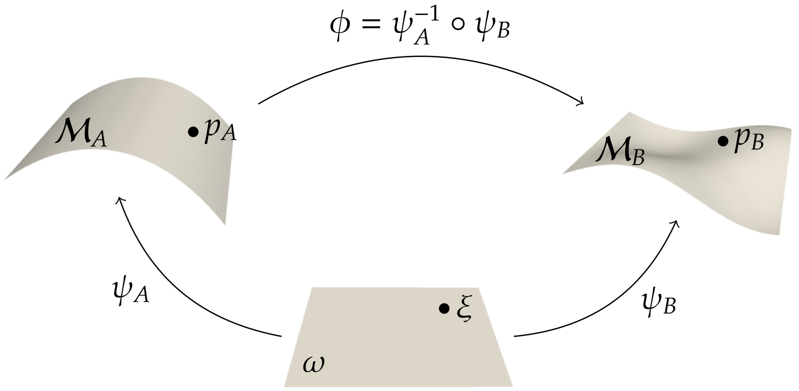


Figure 8.1: Sketch of a deformation ϕ between parametric surfaces $\mathcal{M}_A$ and $\mathcal{M}_B$, which are parametrized over the same chart domain ω.

For the moment, we regard both manifolds $\mathcal{M}_A$ and $\mathcal{M}_B$ and the deformation ϕ as fixed. We define the Cauchy–Green strain tensor $G(\xi) \in \mathbb{R}^{2\times 2}$ at a point $\xi \in \omega$ by the relation $g_B(\xi)(v,w) = g_A(\xi)(G(\xi)v,w)$ for all $v,w \in \mathbb{R}^2$, s.t. we obtain $G(\xi) = g_A(\xi)^{-1}g_B(\xi)$. We recall that the matrix representation of the shape operator of $\mathcal{M}_A$ on the chart domain is given by $S_A(\xi) = g_A(\xi)^{-1}h_A(\xi)$. To compare $S_A(\xi)$ with the corresponding shape operator on $\mathcal{M}_B$, for $p_A \in \mathcal{M}_A$, we take into account the pull-back $S_B^*(p_A) \colon T_{p_A}\mathcal{M}_A \to T_{p_A}\mathcal{M}_A$ given by

$$g_A(p_A)\left(S_B^*(p_A)(V), W\right) = h_B(\phi(p_A))\left(d\phi(p_A)(V), d\phi(p_A)(W)\right) \,,$$

and we define the relative shape operator $S^{\mathrm{rel}}(p_A) \colon T_{p_A}\mathcal{M}_A \to T_{p_A}\mathcal{M}_A$ by $S^{\mathrm{rel}}(p_A) := S_A(p_A) - S_B^*(p_A)$. Then, a matrix representation $S^{\mathrm{rel}}(\xi) \in \mathbb{R}^{2\times 2}$ of the relative shape operator on the chart domain is given by $S^{\mathrm{rel}}(\xi)(v,w) = g_A(\xi)^{-1}\left(h_A(\xi)(v,w) - h_B(\xi)(v,w)\right)$. Finally, in analogy to Definition 8.1.1, we say that a deformation $\phi = \psi_B \circ \psi_A^{-1} \colon \mathcal{M}_A \to \mathcal{M}_B$ is an isometry if for all $V, W \in T_{p_A}\mathcal{M}_A$

$$g_B(\phi(p_A))\left(d\phi(p_A)(V), d\phi(p_A)(W)\right) = g_A(p_A)(V,W) \,, \tag{8.1}$$

which can be transferred to the chart domain to the equivalent relation $g_A(\xi) = g_B(\xi)$. As above, an isometry implies length-preservation. In the following, we consider $\mathcal{M}_A$ as a reference domain always regarded to be fixed, whereas the deformed domain $\mathcal{M}_B$ is obtained as a solution of a specific equilibrium problem under certain load conditions. Therefore, we indicate the operators G and S^{rel} in dependence of the deformation ϕ or the chart map ψ_B, *i.e.*, we write G_ϕ, S^{rel}_ϕ or G_{ψ_B}, $S^{\mathrm{rel}}_{\psi_B}$.

Thin Elastic Shells. Finally, we give the definition of a thin elastic shell.

Definition 8.1.5 (Thin Elastic Shell)**.** A thin elastic shell is an elastic body $\mathcal{S}^\delta \subset \mathbb{R}^3$ of the following type

$$\mathcal{S}^\delta = \left\{ x \in \mathbb{R}^3 \;:\; x = p + \tau n(p) \text{ with } p \in \mathcal{M},\ \tau \in \left(-\frac{\delta}{2}, \frac{\delta}{2}\right) \right\} \,, \tag{8.2}$$

where $\mathcal{M} = \psi(\omega) \subset \mathbb{R}^3$ is a regular and orientable two-dimensional surface, which can be parametrized by a single chart $\psi \colon \overline{\omega} \to \mathbb{R}^3$ for an open and bounded domain $\omega \subset \mathbb{R}^2$ with Lipschitz boundary. Furthermore, we assume that there is no self-intersection, *i.e.*, for $p, \tilde{p} \in \mathcal{M}$ and $\tau, \tilde{\tau} \in \left(-\frac{\delta}{2}, \frac{\delta}{2}\right)$, the relation $p + \tau n(p) = \tilde{p} + \tilde{\tau} n(\tilde{p})$ implies that $(p,\tau) = (\tilde{p}, \tilde{\tau})$. Then we call $\mathcal{M}$ the midsurface and $\delta > 0$ the thickness of the shell.

Remark 8.1.6. More generally, we could consider $\mathcal{M} \subset \mathbb{R}^3$ as an arbitrary regular and orientable two-dimensional surface, but here we restrict to parametric surfaces. In this simplified case, the orientability constraint follows directly.

8.1.2 Two-Dimensional Models for Elastic Deformations of Thin Shells

In the following, we fix a thin elastic shell $\mathcal{S}_A^\delta = \psi_A(\omega)$ with midsurface $\mathcal{M}_A$ as a reference domain. For a force $F_A \colon \mathcal{S}_A^\delta \to \mathbb{R}^3$ acting on $\mathcal{S}_A^\delta$, an equilibrium deformation $\Phi\colon \mathcal{S}_A^\delta \to \mathbb{R}^3$ is described by minimizing the free energy

$$\mathcal{E}_{\text{free}}(\Phi) = \int_{\mathcal{S}_A^\delta} \mathbb{W}_{3D}(D\Phi) - F_A \cdot \Phi \, \mathrm{d}x_A \,,$$

where $\mathbb{W}_{3D} : \mathcal{S}_A^\delta \times \mathbb{R}_+^{3\times 3} \to \mathbb{R}$ is assumed to be a hyperelastic energy density function as we have introduced in Section 6.1.1. Furthermore, we assume that Φ is clamped at a fixed part

$$\Gamma_A^\delta = \left\{ x_A \in \mathbb{R}^3 \ : \ x_A = p_A + \tau_A n_A(p_A) \text{ with } p_A \in \Gamma_A \,,\ \tau_A \in \left(-\frac{\delta}{2}, \frac{\delta}{2}\right) \right\} \subset \mathcal{S}_A^\delta$$

for $\Gamma_A \subset \partial\mathcal{M}_A$. Note that even if the deformed midsurface $\mathcal{M}_B = \Phi(\mathcal{M}_A)$ is a Riemannian manifold, it is unclear that the deformed object $\Phi(\mathcal{S}_A^\delta)$ is itself a thin elastic shell of type (8.2), since in general $\Phi(p_A + \tau_A n_A(p_A)) \neq \Phi(p_A) + \tau_B n_B(\Phi(p_A))$. However, considering the limit $\delta \to 0$, we are interested in understanding the 3D deformation Φ just by a 2D deformation ϕ of the the midsurface $\mathcal{M}_A$, or alternatively by a chart map $\psi_B = \phi \circ \psi_A$ parameterizing the deformed midsurface $\mathcal{M}_B = \psi_B(\omega)$. In particular, we ask for an appropriate energy functional, which characterizes the 2D deformations as corresponding (local) minimizer. In the following, we summarize two approaches to obtain such a limit energy functional. First, the models of Koiter's type make additional assumptions on the 3D deformation, which directly allows a 2D description. Furthermore, a suitable framework to study the limit of minimizing deformations of the free energy for $\delta \to 0$, is established by Γ-convergence, which we have introduced in Section 2.3.

Koiter Type Models. We start with the simple and commonly used Mindlin–Reissner model (see, *e.g.*,[Bra07]) in plate theory, *i.e.*, we consider the flat case $\mathcal{M}_A = \omega \subset \mathbb{R}^2$ and may assume that $\psi_A = \text{id}$. For a point $x_A \in \mathcal{S}_A^\delta$, we use the notation $x_A = (\xi, z)$ with $\xi \in \omega$ and $z \in \left(-\frac{\delta}{2}, \frac{\delta}{2}\right)$. The force $F_A(\xi) = (0, 0, f_n(\xi))^T$ is supposed to act only into the orthogonal direction. Now, in the Mindlin–Reissner model, it is assumed that the displacement $U\colon \mathcal{S}_A^\delta \to \mathbb{R}^3$ has the form

$$U(x_A) = U(\xi, z) = \begin{pmatrix} -z\theta(\xi) \\ w(\xi) \end{pmatrix},$$

where $\theta\colon \omega \to \mathbb{R}^2$ represents the normal stretch and $w\colon \omega \to \mathbb{R}$ the transversal bending displacement. Starting from linear elasticity with the free energy defined in (6.10) and assuming that the normal stress $\sigma_{33} = 0$ vanishes, we can derive that

$$\begin{aligned}
&\mathcal{E}_{\text{free}}^{\text{lin}}(U) = \mathcal{E}_{\text{free}}^{\text{lin,MR}}(\theta, w) \\
&\quad = \delta \frac{E}{2(1+\nu)} \int_\omega |\nabla w - \theta|^2 \, \mathrm{d}\xi + \delta^3 \frac{E}{24(1+\nu)} \int_\omega \varepsilon(\theta) : \varepsilon(\theta) + \frac{\nu}{1-2\nu} \operatorname{div}(\theta)^2 \, \mathrm{d}\xi \ - \ \delta \int_\omega f_n w \, \mathrm{d}\xi \,.
\end{aligned}$$

In addition, in the Kirchoff–Love plate model, it is assumed that deformed normals are orthogonal to the deformed midsurface, which implies $\nabla w = \theta$. Thus, the free energy can be expressed solely in terms of w as

$$\mathcal{E}_{\text{free}}^{\text{lin}}(U) = \mathcal{E}_{\text{free}}^{\text{lin,KL}}(w) = \delta^3 \frac{E}{24(1+\nu)} \int_\omega \|D^2 w\|_F^2 + \frac{\nu}{1-2\nu} (\Delta w)^2 \, \mathrm{d}\xi \ - \ \delta \int_\omega f_n w \, \mathrm{d}\xi \,.$$

Now, for the general case of a generic shell, a similar model was investigated by Koiter [Koi66]. Considering the undeformed and deformed midsurfaces $\mathcal{M}_A$ and $\mathcal{M}_B$ parametrized over the same chart domain $\omega \subset \mathbb{R}^2$, the corresponding thin elastic objects $\mathcal{S}_A^\delta$ and $\mathcal{S}_B^\delta$ are obtained as images of extended chart maps from a thickened chart domain $\omega \times \left(-\frac{\delta}{2}, \frac{\delta}{2}\right) \subset \mathbb{R}^3$ given by

$$\psi_A^{3D}(\xi, z) = \psi_A(\xi) + z n_A(\xi) \,, \qquad \psi_B^{3D}(\xi, z) = \psi_B(\xi) + z v(\xi) \,.$$

Now, the Kirchhoff–Love assumption transfers to $v = n_B$. Again starting from linear elasticity (6.10) and denoting by $u = \psi_B - \psi_A \colon \omega \to \mathbb{R}^3$ the displacement of the midsurface w.r.t. the chart domain, the stored elastic energy in terms of u can be expressed by

$$\begin{aligned}\mathcal{E}^{\mathrm{lin}}_{\mathrm{stored}}(U) = \mathcal{E}^{\mathrm{lin,Koi}}_{\mathrm{stored}}(u) =& \frac{\delta}{2}\int_\omega \sqrt{\det g_A}\, \mathbb{H}(g_B - g_A)^{\mathrm{lin}} : (g_B - g_A)^{\mathrm{lin}}\,\mathrm{d}\xi \\ &+\frac{\delta^3}{24}\int_\omega \sqrt{\det g_A}\, \mathbb{H}(h_B - h_A)^{\mathrm{lin}} : (h_B - h_A)^{\mathrm{lin}}\,\mathrm{d}\xi\,,\end{aligned} \tag{8.3}$$

where $\mathbb{H} \in \mathbb{R}^{3\times3\times3\times3}$ is a fourth order tensor with entries

$$\mathbb{H}^{ijkl} = \frac{4\lambda\mu}{\lambda+2\mu} g_A^{ij} g_A^{kl} + 2\mu(g_A^{ik}g_A^{jl} + g_A^{il}g_A^{jk})\,,$$

and the linearizations of the first and second fundamental forms in the displacement u are explicitly given by

$$\begin{aligned}(g_B - g_A)^{\mathrm{lin}} =& (D\psi_A)^T Du + (Du)^T D\psi_A\,,\\ (h_B - h_A)^{\mathrm{lin}}_{ij} =& -\partial_{ij}u \cdot n_A + \frac{1}{\sqrt{\det g_A}}\left(\partial_1 u \cdot (\partial_{ij}\psi_A \times \partial_2\psi_A) + \partial_2 u \cdot (\partial_1\psi_A \times \partial_{ij}\psi_A)\right)\\ &+ \frac{\partial_{ij}\psi_A \cdot n_A}{\sqrt{\det g_A}}\left(\partial_1 u \cdot (\partial_2\psi_A \times n_A) + \partial_2 u \cdot (n_A \times \partial_1\psi_A)\right)\,.\end{aligned}$$

As proposed in [Koi66], the energy functional (8.3) motivates the definition of a nonlinear Koiter shell model in terms of the chart map ψ_B parameterizing the deformed domain with stored elastic energy given by

$$\begin{aligned}\mathcal{E}^{\mathrm{nl,Koi}}_{\mathrm{stored}}(\psi_B) =& \frac{\delta}{2}\int_\omega \sqrt{\det g_A}\, \mathbb{H}(g_B - g_A) : (g_B - g_A)\,\mathrm{d}\xi\\ &+\frac{\delta^3}{24}\int_\omega \sqrt{\det g_A}\, \mathbb{H}(h_B - h_A) : (h_B - h_A)\,\mathrm{d}\xi\,.\end{aligned}$$

We notice that the part involving the first fundamental forms, which is called membrane energy, is scaled with a factor δ, whereas the part involving the second fundamental forms, which is referred as bending energy, is scaled with a factor δ^3.

More generally, we consider mixed models with a weighted sum of a membrane energy depending on the Cauchy–Green strain tensor G_{ψ_B} and a bending energy depending on the relative shape operator $S^{\mathrm{rel}}_{\psi_B}$. Such a model was, *e.g.*, applied in [IBRS13]. Thus, for suitable density functions $\mathbb{W}_{\mathrm{mem}}$ and $\mathbb{W}_{\mathrm{ben}}$, we define a stored elastic energy by

$$\mathcal{E}^{\mathrm{nl,mix}}_{\mathrm{stored}}(\psi_B) = \frac{\delta}{2}\int_\omega \sqrt{\det g_A}\, \mathbb{W}_{\mathrm{mem}}(G_{\psi_B})\,\mathrm{d}\xi + \frac{\delta^3}{24}\int_\omega \sqrt{\det g_A}\, \mathbb{W}_{\mathrm{ben}}(S^{\mathrm{rel}}_{\psi_B})\,\mathrm{d}\xi\,.$$

Indeed, we immediately see that the limiting energy functionals rigorously observed by Γ-convergence are of a similar structure and permit the same scaling in the thickness.

Γ-Convergence Results. In the following, we summarize certain Γ-convergence results leading to membrane and bending energy functionals acting on deformations of the 2D midsurface. First, since we are now considering the limit $\delta \to 0$ we indicate for a fixed $\delta > 0$ the stored elastic energy by $\mathcal{E}^{\delta}_{\mathrm{stored}}$, which is defined on a space of deformations $\phi\colon \mathcal{S}^\delta_A \to \mathbb{R}^3$. We note that the underlying domain $\mathcal{S}^\delta_A$ and consequently an associated function space for ϕ changes by varying δ. Thus, to generate a suitable setup for Γ-convergence, which in particular requires a sequence of functionals defined on a common function space, a transformation to an object $\mathcal{S}^1_A$ with unit thickness has to be applied. Then, in case of Γ-convergence the limiting functional is a priori also defined on deformations of $\mathcal{S}^1_A$, but in the subsequently presented results it turns out that these limits can be identified as energies on deformations of the midsurface $\mathcal{M}_A$.

For a membrane theory, in [LDR95, LDR96], the above mentioned rescaling for the sequence $\big(\frac{1}{\delta}\mathcal{E}^{\delta}_{\text{stored}}\big)_{\delta>0}$ of energy functionals was taken into account. For a homogeneous material and under p-growth assumption on the hyperelastic density function $\mathbb{W}_{3D}$ with $p \in (1,\infty)$, a Γ-convergence result was proven w.r.t. to the weak $W^{1,p}$ topology of deformations from the thickened chart domain, where the limit functional is given by

$$\mathcal{E}_{\text{mem}}(\psi_B) = \int_{\omega} \sqrt{\det g_A}\ \mathbb{W}^{qc}_{2D}(\xi, D\psi_B(\xi))\ \mathrm{d}\xi\,,$$

where the corresponding 2D stored energy function $\mathbb{W}_{2D}$ is given by

$$\mathbb{W}_{2D}(\xi, F) := \min_{z\in\mathbb{R}^3} \mathbb{W}_{3D}\,\big(\ (F|z)\ (\partial_1\psi_A(\xi)|\partial_2\psi_A(\xi)|n_A(\xi))^{-1}\ \big)$$

and $\mathbb{W}^{qc}_{2D}$ is the quasi convex envelope of $\mathbb{W}_{2D}$. Note that especially the density function of a Saint-Venant–Kirchhoff material is not polyconvex, but in this case $\mathbb{W}^{qc}_{2D}$ can be computed explicitly.

For a bending theory, we consider a sequence $\big(\frac{1}{\delta^3}\mathcal{E}^{\delta}_{\text{stored}}\big)_{\delta>0}$ of appropriately scaled energy functionals. First, because of the scaling of order three, note that a finite value of the limiting functional can only be expected on the set of minimizers of the membrane energy $\mathcal{E}_{\text{mem}}$. Thus, according to the definition of a smooth isometry in (8.1), for $m \in \mathbb{N}_+$. we define the space of $W^{m,2}$-isometries by

$$W^{m,2}_{\text{iso}}(\mathcal{M}_A, \mathbb{R}^3) := \left\{\phi \in W^{m,2}(\mathcal{M}_A, \mathbb{R}^3)\ :\ g_A(p_A) = g_B(\phi(p_A)) \text{ for } a.e.\ p_A \in \mathcal{M}_A\right\}\,.$$

In [FJM02], a Γ-convergence result was derived in the plate case. There, a central inside is the rigidity result

$$\min_{Q\in SO(3)} \int_{\omega} |D\psi - Q|^2\ \mathrm{d}\xi \leqslant c \int_{\omega} \operatorname{dist}^2(D\psi, SO(3))\ \mathrm{d}\xi \tag{8.4}$$

for $\psi \in W^{1,2}(\omega, \mathbb{R}^3)$, which can be regarded as nonlinear version of Korn's inequality (2.2). Then, provided that $\mathbb{W}_{3D}(M) \geqslant c\ \operatorname{dist}^2(M, SO(3))$ for some constant $c > 0$, the estimate (8.4) can be applied to obtain a Γ-limit w.r.t. to the strong $W^{1,2}$-topology, which is given by

$$\mathcal{E}^{\text{plate}}_{\text{ben}}(\psi) = \begin{cases} \dfrac{1}{24}\displaystyle\int_{\omega} \min_{z\in\mathbb{R}^3} Q_2 \left(\begin{array}{c|c} h(\xi) & \mathrm{z} \\ 0 & \end{array} \right) \mathrm{d}\xi & \text{if } \psi \in W^{2,2}_{\text{iso}}(\omega, \mathbb{R}^3)\,, \\ \infty & \text{otherwise}\,, \end{cases} \tag{8.5}$$

where Q_2 is the quadratic form $Q_2(M) := D^2\mathbb{W}_{3D}(\mathbb{1}_{3\times3})(M)(M)$. For a Saint-Venant–Kirchhoff material, the inner minimization problem can be computed explicitly, s.t. the integrand for $W^{2,2}$-isometries is given by $2\mu\,\operatorname{tr}(h_B^2) + \frac{\lambda\mu}{\mu+\frac{\lambda}{2}}\operatorname{tr}(h_B)^2$. In [FJMM03], the result was extended to general shells, where the limit energy functional takes into account the relative shape operator and is given by

$$\mathcal{E}_{\text{ben}}(\phi) = \begin{cases} \dfrac{1}{24}\displaystyle\int_{\mathcal{M}_A} \min_{z\in\mathbb{R}^3} Q_2(S^{\text{rel}}_{\phi}(p_A) + z\otimes n_A(p_A))\ \mathrm{d}\mathscr{H}^2(p_A) & \text{if } \phi \in W^{2,2}_{\text{iso}}(\mathcal{M}_A, \mathbb{R}^3)\,, \\ \infty & \text{otherwise}\,. \end{cases} \tag{8.6}$$

Finally, we mention that the here presented Γ-convergence results only hold under certain additional regularity assumptions on $\mathbb{W}_{3D}$, which we do not specify in detail and instead refer the reader to the literature mentioned above.

8.1.3 Overview of Computational Methods for Thin Elastic Shells

In the literature, there are many possibilities to discretize thin elastic shells and the corresponding deformation energies. Here, we give a brief overview of different approaches, where we restrict to two-dimensional models as we have described above. The main numerical difficulty is that due to curvature terms appearing in the bending energy at least an approximation of second derivatives is required.

$W^{2,2}$-Conforming Finite Elements. Classical finite elements, *e.g.*, globally continuous functions on a triangular mesh that are polynomials restricted to any element, unfortunately, do not belong to the space $W^{2,2}$. Thus, for a conforming finite-dimensional approximation of deformations more elaborated discretizations have to be investigated. A globally C^1-regular finite element is for example given by the Argyris finite element (see, *e.g.*, [Bra07]), which requires degrees of freedom for first and second derivatives at nodes and for normal derivatives at edges. Besides a highly computational effort, an extension of the Argyris element to curved domains is not straightforward. Instead, for triangular meshes, a suitable alternative is given by taking into account a larger support of nodalwise basis functions. In [COS00], so-called loop subdivision elements were used to compute elastic deformations for a linear Koiter type model. On quadrilateral meshes arbitrary regularity can be obtained by nonuniform rational B-splines (NURBS) (see, *e.g.*, [HCB05]).

Nonconforming Finite Elements. An alternative is given by nonconforming finite element functions, which do not necessarily belong to the space $W^{2,2}$, but solutions to the discretized elasticity problem admit similar error estimates, s.t. convergence in the limit is guaranteed. Later, we make use of the discrete Kirchhoff triangle (DKT) element [BBH80], which was originally proposed to solve the linear plate bending problem. We discuss the DKT element in detail in Section 8.1.4.

Discrete Differential Geometry. The spirit of discrete differential geometry (DDG) is actually to understand the discrete object, *e.g.*, in form of a triangular mesh, itself as a discrete surface, by making sense of differential geometric quantities like the Riemannian metric or the mean curvature. These objects are then defined only on elements, edges, or at nodes and thus cannot be considered as functions on the whole domain. By comparing dihedral angles of neighborhood triangles, a discrete bending energy was introduced in [GHDS03]. Combined with a nonlinear membrane energy, this was extensively applied in works by Heeren and coworkers [HRS$^+$14, Hee16]. Computationally, methods from discrete differential geometry have turned out to be extremely efficient. However, convergence, *e.g.*, of the mean curvature for a sequence of triangular meshes, can only be established in an integrated sense [War08].

The Engineering Point of View. In engineering applications, it is often convenient to combine a pure membrane with a pure bending energy model. Then the corresponding stiffness matrices are just assembled elementwise, where only the associated degrees of freedom are taken into account for the membrane and bending part, respectively. For example, the DKT-CST element [BH81] combines the DKT element with the constant strain triangle element. Alternatively, quadratic in-plane displacements for the membrane part lead to the DKTP element [DMM86]. A coupling of the two parts of the deformation energy is guaranteed by certain commonly defined degrees of freedom, but other values are rather meaningless for either the membrane or the bending energy. Similar as in DDG, there is no corresponding overall function defined on the discrete mesh, but related to a mixed method, the membrane and bending stress can be regarded separately as functions.

Bending Isometries. Minimizing the pure bending energy under an isometry constraint numerically has been rather less studied. According to the discretization, the isometry constraint has to be formulated appropriately. Using DDG, this corresponds to the condition that length and angles are preserved for all triangles. In [WBH$^+$07], the Willmore energy was approximated by a quadratic curvature energy, which is related to a nonconforming Crouzeix–Raviart finite element discretization. Instead of describing the triangular mesh by nodal positions, in [WDAH10], edge length and dihedral angles were used as degrees of freedom. Then the isometries can be approximated by only allowing the dihedral angles to vary. This approach was studied in [Sas19]. Concerning the above discussion about the regularity of isometries related to the Nash–Kuiper theorem 8.1.2 and the Hartman–Nirenberg theorem 8.1.4, the deformations in DDG are even not of class C^1. However, a notion of discrete developable surfaces on quadrilateral meshes was established in [RHSH18]. In [Bar13], a numerical approximation scheme for large bending isometries was provided by making use of the DKT element. There, the isometry constraint was enforced only at nodal positions, which can easily be formulated due to the derivative degrees of freedom. Then the regularity error estimates for the DKT element still allow to prove convergence of minimizers. For a computational scheme, a discrete $W^{2,2}$ gradient flow using a linearization of the isometry constraint was proposed.

8.1.4 Discrete Kirchhoff Triangle Element

Here, we recall the DKT element on plates and refer the reader to [BBH80, Bar15, Bra07] for more detailed introductions. For simplicity, we assume that $\omega \subset \mathbb{R}^2$ is polygonal, s.t. we can directly consider a triangulation $\mathcal{T}_h$ of ω. Otherwise, ω could be approximated by a polygonal domain. We denote by $\mathcal{N}_h$ the set of nodes in $\mathcal{T}_h$. For a triangle T in $\mathcal{T}_h$, let $P_k(T)$ be the space of polynomials of order $k \in \mathbb{N}$. In analogy, we consider for an edge E the space $P_k(E)$. Furthermore, we define for a triangle T the space $P_{3,\text{red}}(T)$ of polynomials of order three reduced by one degree of freedom by

$$P_{3,\text{red}}(T) := \left\{ w \in P_3(T) \ : \ w(z_T) = \frac{1}{3} \sum_{z \in \mathcal{N}_h \cap T} w(z) + \nabla w(z)(z_T - z) \right\} ,$$

where $z_T = \frac{1}{3} \sum_{z \in \mathcal{N}_h \cap T} z$ denotes the center of mass. We denote by $\Gamma_\omega \subset \partial\omega$ the Dirichlet boundary. Then, we define the following finite element spaces.

Definition 8.1.7 (DKT Finite Element Spaces)**.**

1. $\mathcal{W}_h(\omega) := \left\{ w_h \in W^{1,2}_{\Gamma_\omega}(\omega) \ : \ w_h|_T \in P_{3,\text{red}}(T) \ \forall T \in \mathcal{T}_h \text{ and } \nabla w_h(z) \text{ is continuous for all } z \in \mathcal{N}_h \right\}$,
2. $\Theta_h(\omega) := \left\{ \theta_h \in \left(W^{1,2}_{\Gamma_\omega}(\omega) \right)^2 \ : \ \theta_h|_T \in (P_2(T))^2 \ \forall T \in \mathcal{T}_h \text{ and } \theta_h \cdot n \in P_1(E) \text{ for every edge } E \text{ in } \mathcal{T}_h \right\}$.

Furthermore, we introduce a discrete gradient operator.

Definition 8.1.8 (DKT Gradient Operator)**.** We define a discrete gradient operator

$$\nabla_h \colon \mathcal{W}_h(\omega) \to \Theta_h(\omega) , \quad w_h \mapsto \nabla_h w_h = \theta_h(w_h) ,$$

where $\theta_h(w_h) \in \Theta_h(\omega)$ is the uniquely defined function that satisfies for each triangle $T \in \mathcal{T}_h$ with nodes z_0, z_1, z_2

1. $\theta_h(w_h)(z_i) = \nabla w_h(z_i)$ for $0 \leqslant i \leqslant 2$ and
2. $\theta_h(w_h)(z_{ij}) \cdot (z_j - z_i) = \nabla w_h(z_{ij}) \cdot (z_j - z_i)$ for $0 \leqslant i, j \leqslant 2$ with $z_{ij} = z_{ji} = \frac{1}{2}(z_i + z_j)$.

Now, we call a function $w_h \in \mathcal{W}_h(\omega)$ a DKT function. Then the approximative second derivative of w_h is given by $\nabla\theta_h(w_h)$. Note that w_h can be determined by the values $w_h(z)$ and the derivatives $\nabla w_h(z)$ at nodes z, and thus, has three degrees of freedom per node. In Figure 8.2, we depict the gradient operator on a single triangular element. Finally, we define function spaces of displacements satisfying clamped boundary conditions on Γ_ω by

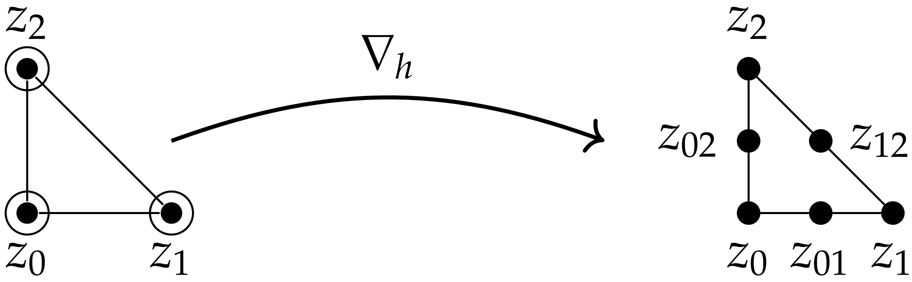


Figure 8.2: Sketch of the DKT gradient operator ∇_h. For a single triangle, it maps a cubic function $w_h \in P_{3,\text{red}}$ defined by the values $w_h(z_i)$ and derivatives $\nabla w_h(z_i)$ at the three nodes z_0, z_1, z_2 to a quadratic function.

$$\begin{aligned} W^{2,2}_{\Gamma_\omega}(\omega) &:= \left\{ w \in W^{2,2}(\omega) \ : \ w|_{\Gamma_\omega} = 0 \text{ and } \nabla w|_{\Gamma_\omega} = 0 \right\} , \\ \mathcal{W}_{h,\Gamma_\omega}(\omega) &:= \{ w_h \in \mathcal{W}_h \ : \ w_h(z) = 0 \text{ and } \nabla w_h(z) = 0 \ \forall z \in \mathcal{N}_h \cap \Gamma_\omega \} . \end{aligned}$$

Various error estimates for the DKT element have been established. For instance, the linear plate bending problem can be approximated under $W^{3,2}$-regularity assumption on the solution displacement s.t. the error of the approximative second derivative in the L^2-norm is of order one (see [Bar15, Theorem 8.2]).

8.2 Shape Design for Mixed Membrane-Bending Models

In the following, we study the optimal material distribution on thin elastic shells via a phase-field approximation. Moreover, deformations are described by models including a membrane and a bending energy part, and we investigate both linear and nonlinear elasticity.

8.2.1 Shape Optimization Problem for a Phase-Field Approximation

To describe a material distribution on the chart domain ω and thus on the midsurface $\mathcal{M}_A$ of the reference object, we consider a characteristic function $\chi \in BV(\omega, \{0,1\})$. More precisely, on the subdomains

$$\mathcal{O}^1 = \{\xi \in \omega \,:\, \chi(\xi) = 1\}\,, \quad \mathcal{O}^0 = \{\xi \in \omega \,:\, \chi(\xi) = 0\}\,,$$

we assume that the elastic material is described by parameters E^m for Young's modulus and ν^m for the Poisson ratio for $m \in \{0,1\}$. For simplicity, we restrict to the case $\nu^1 = \nu^0$. Now, as in Section 6.3, we could formulate a shape optimization problem in terms of the characteristic function. As we have described in Section 6.3.1, we use a phase-field variable $v \in W^{1,2}(\omega, [-1,1])$ to approximate the characteristic function χ. Here, we directly formulate a shape optimization problem in terms of the phase-field variable, since our numerical computation scheme is based on this approximation approach. Then, we define an interpolation $E(v)$ of Young's modulus depending on the phase-field variable by $E(v) = \chi(v)E^1 + (1-\chi(v))E^0$, where we set $\chi(v) = \frac{1}{16}(1+v)^4$. We recall that the corresponding Lamé–Navier parameters μ^m, λ^m are determined by (6.6). In analogy, we define interpolations $\mu(v)$ and $\lambda(v)$ depending on the phase-field variable.

State Equations. Here, we define stored elastic energy functionals both for linear an nonlinear elasticity. We recall that the chart map $\psi_B = \psi_A + u$ parameterizing the deformed midsurface can be recovered by the displacement u. Thus, we formulate all energies in terms of the displacement.

First, for a nonlinear membrane energy, we take into account the hyperelastic energy density function (6.8) and define

$$\mathbb{W}_{\mathrm{mem}}(v,S) = \frac{\mu(v)}{2}\,\mathrm{tr}(S) + \frac{\lambda(v)}{4}\det(S) - \left(\frac{\mu(v)}{2} + \frac{\lambda(v)}{4}\right)\log(\det(S)) - \mu(v) - \frac{\lambda(v)}{4}$$

for $S \in \mathbb{R}^{2\times 2}_{\mathrm{sym},+}$. In [HRWW12], this density function was applied for thin elastic objects. Then, the nonlinear membrane energy is given by

$$\mathcal{E}^{\mathrm{nl}}_{\mathrm{mem}}(v,u) = \int_\omega \sqrt{\det g_A}\; \mathbb{W}_{\mathrm{mem}}\left(v, g_A^{-1} g_B\right)\,\mathrm{d}\xi\,.$$

For a nonlinear bending energy, we recall from (8.6) that the Γ-limit takes into account the relative shape operator, which has a matrix representation $g_A^{-1}(h_B - h_A)$ on the chart domain. Here, we simply choose the Frobenius norm

$$\mathbb{W}_{\mathrm{ben}}(v,S) = \frac{E(v)}{24}\|S\|_F^2 \tag{8.7}$$

and define the nonlinear bending energy by

$$\mathcal{E}^{\mathrm{nl}}_{\mathrm{ben}}(v,u) = \int_\omega \sqrt{\det g_A}\; \mathbb{W}_{\mathrm{ben}}\left(v, g_A^{-1}(h_B - h_A)\right)\,\mathrm{d}\xi\,.$$

For a pure bending model, the limiting functional (8.6) is restricted to the set of $W^{2,2}$-isometries, which minimize the membrane energy $\mathbb{W}_{\mathrm{mem}}(v,\cdot)$. Now, for a mixed model with both a membrane and a bending part, the membrane energy functional acts as a regularizer for the isometry constraint. Thus, without an isometry constraint, we define a stored elastic energy $\mathcal{E}^{\mathrm{nl,mix}}_{\mathrm{stored}} \colon W^{1,2}(\omega,[-1,1]) \times W^{2,2}(\omega,\mathbb{R}^3) \to \mathbb{R} \cup \{\infty\}$ by

$$\mathcal{E}^{\mathrm{nl,mix}}_{\mathrm{stored}}(v,u) = \delta\,\mathcal{E}^{\mathrm{nl}}_{\mathrm{mem}}(v,u) + \delta^3\,\mathcal{E}^{\mathrm{nl}}_{\mathrm{ben}}(v,u)\,.$$

In the case of linear elasticity, we use the linear Koiter type model as defined in (8.3), where we additionally allow the material to vary, *i.e.*, we define a linear membrane and a linear bending energy by

$$\begin{aligned}\mathcal{E}^{\mathrm{lin}}_{\mathrm{mem}}(v,u) &= \frac{1}{2}\int_\omega \sqrt{\det g_A}\ \mathbb{H}(v)(g_B - g_A)^{\mathrm{lin}} : (g_B - g_A)^{\mathrm{lin}}\,\mathrm{d}\xi\,,\\ \mathcal{E}^{\mathrm{lin}}_{\mathrm{ben}}(v,u) &= \frac{1}{24}\int_\omega \sqrt{\det g_A}\ \mathbb{H}(v)(h_B - h_A)^{\mathrm{lin}} : (h_B - h_A)^{\mathrm{lin}}\,\mathrm{d}\xi\,,\end{aligned}$$

with

$$\mathbb{H}^{ijkl}(v) = \frac{4\lambda(v)\mu(v)}{\lambda(v)+2\mu(v)} g_A^{ij} g_A^{kl} + 2\mu(v)\left(g_A^{ik} g_A^{jl} + g_A^{il} g_A^{jk}\right)\,,$$

and a linear stored elastic energy by

$$\mathcal{E}^{\mathrm{lin,mix}}_{\mathrm{stored}}(v,u) = \delta\,\mathcal{E}^{\mathrm{lin}}_{\mathrm{mem}}(v,u) + \delta^3\,\mathcal{E}^{\mathrm{lin}}_{\mathrm{ben}}(v,u)\,.$$

For a force $f_A : \mathcal{M}_A \to \mathbb{R}^3$ acting on $\mathcal{M}_A$, we set $f = f_A \circ \psi_A : \omega \to \mathbb{R}^3$ and define the potential energy by

$$\mathcal{E}_{\mathrm{pot}}(u) = \delta\int_\omega \sqrt{\det g_A}\ f\cdot u\,\mathrm{d}\xi = \delta\int_\omega \sqrt{\det g_A}\ f\cdot(\psi_B - \psi_A)\,\mathrm{d}\xi\,.$$

Finally, for a fixed material distribution described by v, the state equation is given by minimizing the free energy

$$\mathcal{E}_{\mathrm{free}}(v,u) = \mathcal{E}_{\mathrm{stored}}(v,u) - \mathcal{E}_{\mathrm{pot}}(u)$$

over all displacements $u \in W^{2,2}_{\Gamma_\omega}(\omega,\mathbb{R}^3)$ of the chart domain satisfying clamped boundary conditions on the Dirichlet boundary $\Gamma_\omega = \psi_A^{-1}(\Gamma_A)$.

Cost Functional. For the cost functional $\mathcal{J}_{\mathrm{expl}}$ explicitly depending on the phase-field variable v and the displacement u, we take into account the potential energy $\mathcal{J}_{\mathrm{expl}}(v,u) = \mathcal{E}_{\mathrm{pot}}(u)$. To measure the area of the set $\mathcal{O}^1$, we define $\mathcal{V}(v) := \int_\omega \sqrt{\det g_A}\ \frac{v+1}{2}\,\mathrm{d}\xi$ as the relaxation of the area functional in terms of the phase-field variable. Moreover, $\mathcal{A}^\epsilon$ is the Modica–Mortola functional as defined in (6.13), which approximates the perimeter functional for $\epsilon \to 0$ (*cf*. Theorem 6.3.2). Then, we consider a shape optimization problem by minimizing a total cost functional

$$\mathcal{J}^\eta_{\mathrm{tot}}(v) = \mathcal{J}^\eta_{\mathrm{expl}}(v,u(v)) = \mathcal{J}_{\mathrm{expl}}(v,u(v)) + \eta\mathcal{A}^\epsilon(v)\,, \tag{8.8}$$

over all phase-fields $v \in W^{1,2}(\omega,[-1,1])$ s.t. an area constraint $\mathcal{V}(v) = V$ holds for some constant $V \in (0,\mathcal{H}^2(\mathcal{M}_A))$. Here, for a fixed phase-field v, we denote by $u(v)$ a minimizer of the free energy with stored elastic energy either given by $\mathcal{E}^{\mathrm{lin,mix}}_{\mathrm{stored}}$ or $\mathcal{E}^{\mathrm{nl,mix}}_{\mathrm{stored}}$. As discussed in Section 6.3, in the case of nonlinear elasticity, the minimizer $u(v)$ is not necessarily unique and thus, a set of minimizers has to be considered.

8.2.2 Finite Element Discretization for Mixed Membrane-Bending Models

Now, we aim at computing minimizer of $\mathcal{J}^\eta_{\mathrm{tot}}$ with a numerical optimization scheme, which requires to evaluate the derivative

$$\frac{d}{dv}\mathcal{J}^\eta_{\mathrm{tot}}(v)(\hat{v}) = \partial_v\mathcal{J}^\eta_{\mathrm{expl}}(v,u(v))(\hat{v}) + \partial_u\mathcal{J}^\eta_{\mathrm{expl}}(v,u(v))(\partial_v u(v)(\hat{v}))\,.$$

To compute the shape sensitivity $\partial_v u(v)(\hat{v})$ we apply the same approach as described in Section 6.3.2 by solving a suitable adjoint problem. This leads to

$$\frac{d}{dv}\mathcal{J}^\eta_{\mathrm{tot}}(v)(\hat{v}) = \partial_v\mathcal{J}^\eta_{\mathrm{expl}}(v,u(v))(\hat{v}) + \partial^2_{v,u}\mathcal{E}_{\mathrm{stored}}(v,u(v))(A)(\hat{v})\,,$$

where the adjoint variable $A \in W^{2,2}_{\Gamma_\omega}(\omega, \mathbb{R}^3)$ solves the linear problem

$$\partial^2_{u,u}\mathcal{E}_{\text{stored}}(v, u(v))(\hat{u})(A) = -\partial_u \mathcal{J}_{\text{expl}}(v, u(v))(\hat{u}) \quad \forall \hat{u} \in W^{2,2}_{\Gamma_\omega}(\omega, \mathbb{R}^3)\,. \tag{8.9}$$

In the case of linear elasticity with stored elastic energy $\mathcal{E}^{\text{lin,mix}}_{\text{stored}}$ and a cost functional defined by the potential energy, the solution to the linear system (8.9) for the adjoint variable is precisely given by $A = -u(v)$.

For a numerical discretization to compute approximations of stationary points of the free energy $\mathcal{E}_{\text{free}}$, we take into account the DKT element. Since we restrict to parametric surfaces $\mathcal{M}_A$, we can consider a triangulation $\mathcal{T}_h$ of the chart domain ω. Then, for the DKT finite element space as in Definition 8.1.7, we call a function $\psi_h \in \mathcal{W}_h(\omega)^3$ a DKT chart map. We fix such a DKT chart map $\psi_{A,h} \in \mathcal{W}_h(\omega)^3$ as an approximation of ψ_A. Then, we formulate discrete energies in terms of a displacement $u_h \in \mathcal{W}_{h,\Gamma_\omega}(\omega)^3$ satisfying clamped boundary conditions on Γ_ω and define the DKT chart map discretizing the deformed domain by $\psi_{B,h} := \psi_{A,h} + u_h$. In Figure 8.3, we show a sketch of this discrete configuration. Note that the numerical approximations $\mathcal{M}_{A,h} := \psi_{A,h}(\omega)$ and $\mathcal{M}_{B,h} := \psi_{B,h}(\omega)$ of the midsurfaces $\mathcal{M}_A$ and $\mathcal{M}_B$ are images of vector-valued DKT functions. In particular, triangular elements on the discretized midsurfaces are curved. The actual discrete deformation ϕ_h is a concatenation of a DKT chart map and the inverse of a DKT chart map.

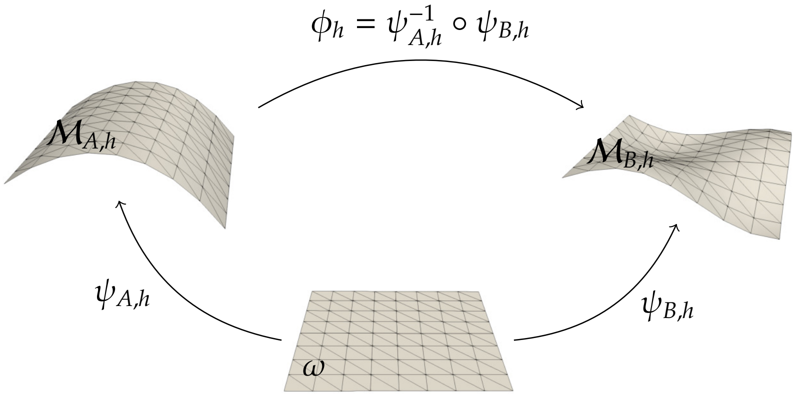


Figure 8.3: Sketch of the numerical approximation of a deformation by DKT chart maps.

Now, for first-order quantities, we simply evaluate the exact gradients of the DKT chart maps at quadrature points q, *i.e.*,

$$g_{A,h}(q) = D\psi_{A,h}(q)^T D\psi_{A,h}(q)\,, \qquad g_{B,h}(q) = D\psi_{B,h}(q)^T D\psi_{B,h}(q)\,,$$
$$n_{A,h}(q) = \frac{\partial_1\psi_{A,h}(q) \times \partial_2\psi_{A,h}(q)}{|\partial_1\psi_{A,h}(q) \times \partial_2\psi_{A,h}(q)|}\,, \qquad n_{B,h}(q) = \frac{\partial_1\psi_{B,h}(q) \times \partial_2\psi_{B,h}(q)}{|\partial_1\psi_{B,h}(q) \times \partial_2\psi_{B,h}(q)|}\,.$$

For second-order terms, we take into account the approximative second derivatives of the DKT chart maps. Note that $\nabla\nabla_h\psi_{A,h}(q)$ and $\nabla\nabla_h\psi_{B,h}(q)$ are in general not symmetric. Thus, we define

$$h_{A,h}(q) = \begin{pmatrix} (\nabla\nabla_h\psi_{A,h}(q))_{11} \cdot n_{A,h}(q) & \frac{1}{2}\left((\nabla\nabla_h\psi_{A,h}(q))_{12} + (\nabla\nabla_h\psi_{A,h}(q))_{21}\right) \cdot n_{A,h}(q) \\ \text{sym} & (\nabla\nabla_h\psi_{A,h}(q))_{22} \cdot n_{A,h}(q) \end{pmatrix}\,,$$
$$h_{B,h}(q) = \begin{pmatrix} (\nabla\nabla_h\psi_{B,h}(q))_{11} \cdot n_{B,h}(q) & \frac{1}{2}\left((\nabla\nabla_h\psi_{B,h}(q))_{12} + (\nabla\nabla_h\psi_{B,h}(q))_{21}\right) \cdot n_{B,h}(q) \\ \text{sym} & (\nabla\nabla_h\psi_{B,h}(q))_{22} \cdot n_{B,h}(q) \end{pmatrix}\,.$$

Next, we discretize the phase-field variable by functions in the finite element space

$$V^1_h(\omega) = \left\{ v_h \in C^0(\omega) \;:\; v_h|_T \text{ is affine } \forall T \in \mathcal{T}_h \right\}\,.$$

Then, we apply a Gaussian quadrature of degree 6 with $Q = 12$ quadrature points for each triangle element with weights ω, which allows us to define the following discrete counterparts

$$\begin{aligned}
\mathcal{E}^{\mathrm{nl}}_{\mathrm{mem},h}(v_h, u_h) &= \sum_{T\in\mathcal{T}_h} \sum_{q=1,\dots,Q} \omega(q) \sqrt{g_{A,h}(q)}\ \mathrm{W}_{\mathrm{mem}}\left(v_h(q), g_{A,h}(q)^{-1} g_{B,h}(q)\right) ,\\
\mathcal{E}^{\mathrm{nl}}_{\mathrm{ben},h}(v_h, u_h) &= \sum_{T\in\mathcal{T}_h} \sum_{q=1,\dots,Q} \omega(q) \sqrt{g_{A,h}(q)}\ \mathrm{W}_{\mathrm{ben}}\left(v_h(q), g_{A,h}(q)^{-1}(h_{B,h}(q) - h_{A,h}(q))\right) .
\end{aligned}$$

Assuming that a force f_h is explicitly given at quadrature points, we set

$$\begin{aligned}
\mathcal{E}_{\mathrm{pot},h}(u_h) &= \delta \sum_{T\in\mathcal{T}_h} \sum_{q=1,\dots,Q} \omega(q) \sqrt{g_{A,h}(q)}\ f_h(q) \cdot u_h(q) ,\\
\mathcal{E}^{\mathrm{nl}}_{\mathrm{free},h}(v_h, u_h) &= \delta\, \mathcal{E}^{\mathrm{nl}}_{\mathrm{mem},h}(v_h, u_h) + \delta^3\, \mathcal{E}^{\mathrm{nl}}_{\mathrm{ben},h}(v_h, u_h) - \mathcal{E}_{\mathrm{pot},h}(u_h) .
\end{aligned}$$

Finally, to solve the state equation $\partial_{u_h}\mathcal{E}^{\mathrm{nl}}_{\mathrm{free},h}(v_h, u_h) = 0$ for a fixed v_h, we apply Newton's method. In the case of linear elasticity, we use in analogy the symmetrized approximative second derivative to define a discrete version of $(h_B - h_A)^{\mathrm{lin}}$ at quadrature points. Then the state equation $\partial_{u_h}\mathcal{E}^{\mathrm{lin}}_{\mathrm{free},h}(v_h, u_h) = 0$ results in a linear system.

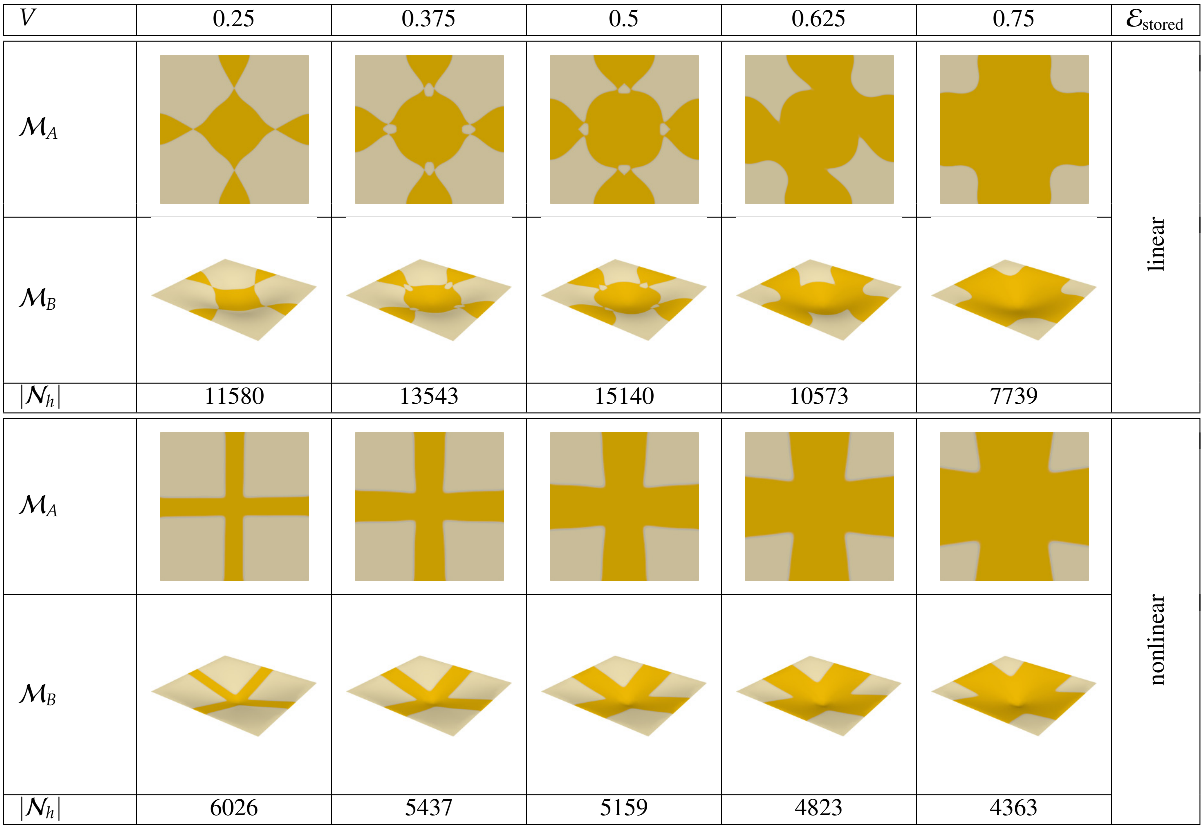
Finally, the shape optimization problem to minimize the fully discrete cost functional $\mathcal{J}^{\eta}_{\mathrm{tot},h}$ over all $v_h \in V^1_h(\omega, [-1, 1])$ s.t. $\mathcal{V}_h(v_h) = V$ is solved by using the `IPOPT` package [WB06]. To this end, we have to provide an implementation of $\mathcal{J}_{\mathrm{tot},h}$ and $\mathcal{V}_h$, as well as the first derivatives of these operators. Then the `IPOPT` solver allows to include a constraint on the amount of hard material and box constraints $-1 \leqslant v_h(z) \leqslant 1$ for all $z \in \mathcal{N}_h$ on the phase-field variable. To obtain a finer resolution of the diffuse interface, we use an adaptive refinement scheme via longest edge bisection. More precisely, after computing the solution v_h we mark those elements $T \in \mathcal{T}_h$ with $\fint_T \|\nabla v_h\|^2 \,\mathrm{d}x > \frac{1}{2}$. Then, we iteratively compute a solution $v_{h'}$ on the refined mesh $\mathcal{T}_{h'}$. The optimization method on the refined mesh $\mathcal{T}_{h'}$ is initialized with the linear prolongations of the solutions v_h and $u_h(v_h)$. For the parameter ϵ in the Modica–Mortola functional, we always choose $\epsilon = 2\min_{T\in\mathcal{T}_h} \mathrm{diam}(T)$, which is thus automatically adapted to the corresponding mesh size h. The longest edge bisection guarantees that the family $(\mathcal{T}_h)_h$ of refined triangular meshes is regular.

8.2.3 Numerical Results for Mixed Membrane-Bending Models

Now, we present our computational results, where the hard material is colored in orange. We always choose material parameters $E^1 = 100$, $E^0 = 1$ for Young's modulus, s.t. one material is substantially stiffer than the other. Moreover, we set the Poisson ratios to $\nu^1 = \nu^0 = 0$. In the following, we take into account reference domains of a flat square, a hemisphere, and a half cylinder. For the coarse initial meshes to discretize the chart maps, which are in our examples either given by the unit square or the unit disc, we use $|\mathcal{N}_h| = 289$ nodes. Then, depending on the specific example, we apply several adaptive refinement steps via longest edge bisection. In the case of the unit disc, new boundary nodes generated by the adaptive refinement are projected onto the boundary of the unit disc.

Centered Load on a Plate

First, we investigate the flat case for $\omega = \mathcal{M}_A = [0, 1]^2$ and $\psi_A = \mathrm{id}$. We consider a force $f = \left(0, 0, \beta\chi_{[0.45,0.55]^2}\right)$, which is acting into the normal direction and is supported on a square in the center of $\mathcal{M}_A$. The displacement is supposed to be clamped at the boundary $\partial\mathcal{M}_A$. As penalty parameter for the Modica–Mortola functional, we choose $\eta = 10^{-3}$ s.t. the contribution of $\eta\mathcal{A}^{\epsilon}$ is small in the total cost functional $\mathcal{J}^{\eta}_{\mathrm{tot}}$. Moreover, we choose different area constraints $V = \frac{k}{8}$ for $k = 2, 3, 4, 5, 6$. Then, depending on this area constraint, we set $\beta = -250V$ for the force to make the corresponding deformations comparable. For nonlinear elasticity, we consider $\delta = 10^{-2}$. Note that in the case of linear elasticity the associated linearized membrane energy is zero for the optimal displacement. Then, for the optimal design, scaling δ is equivalent to scaling the force f and the penalty parameter η for the Modica–Mortola functional. To obtain comparably large deformations as in the nonlinear case, we choose $\delta = 10^{-1}$. In Figure 8.4, we depict the results for linear elasticity after 9 adaptive refinement steps and for nonlinear elasticity after 7 adaptive refinement steps. In all results, we observe a cross type structure for the hard material.



V	0.25	0.375	0.5	0.625	0.75	$\mathcal{E}_{\text{stored}}$
$\mathcal{M}_A$						linear
$\mathcal{M}_B$						
$\lvert\mathcal{N}_h\rvert$	11580	13543	15140	10573	7739	
$\mathcal{M}_A$						nonlinear
$\mathcal{M}_B$						
$\lvert\mathcal{N}_h\rvert$	6026	5437	5159	4823	4363	

Figure 8.4: Optimal material distributions on a plate $\mathcal{M}_A = [0,1]^2$ for a centered load supported on $(0.45, 0.55)^2$ and acting into normal direction. We compare the results for different area constraints V. Top: Linear elasticity. Bottom: Nonlinear elasticity.

However, in the case of linear elasticity and for small amounts of hard material, the trusses become very thin at certain points, whereas for nonlinear elasticity, we obtain pure cross structures. For a more detailed analysis of these cross structures, we consider in Figure 8.5 the stresses and energy functionals for both types of crosses. More precisely, under the above load scenario we compute for an area constraint $V = 0.25$ the optimal design for linear and nonlinear elasticity and compare the potential energy functional for both solutions. We denote the corresponding phase-fields by v_L (linear) and v_{NL} (nonlinear). Then, we compute the minimizer of the free energy functional for the other cross structure, *i.e.*, we take into account v_L for nonlinear elasticity and v_{NL} for linear elasticity. We observe that in the case of linear elasticity, v_L has an approximately 25% lower potential energy than v_{NL}. Conversely, in the case of nonlinear elasticity, v_{NL} is approximately 30% better than v_L. Considering the distribution of the membrane stress for v_L, there is indeed a huge concentration at the four points, where the structure of the hard material becomes very thin. However, the linearized membrane stress is zero for this load scenario, since the force is acting into the normal direction and the linearizations of the first and second fundamental forms are given by

$$(g_B - g_A)^{\text{lin}} = \begin{pmatrix} 2\partial_1 u_1 & \partial_1 u_2 + \partial_2 u_1 \\ \partial_1 u_2 + \partial_2 u_1 & 2\partial_2 u_2 \end{pmatrix}, \quad (h_B - h_A)^{\text{lin}} = D^2 u_3 .$$

Now, both for v_L and v_{NL}, the linear bending stress is quite small around the four points. Thus, the thicker structure of v_{NL} in this region does not essentially improve the potential energy, but it is advantageous to use more hard material in the center and at the boundary of the plate. Moreover, we notice that the nonlinear bending energy is lower for v_{NL}, since the nonlinear deformation behaves much more rigid and the bending stress is concentrated at the center, where the force is acting. For these four computations, we have chosen a uniform triangular mesh with 16641 nodes. Besides, since we obtain the same optimal designs as in Figure 8.4, this indicates that our solution is not mesh-dependent.

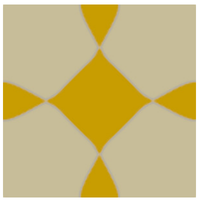
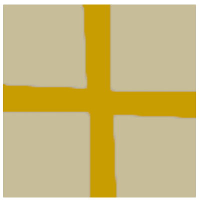
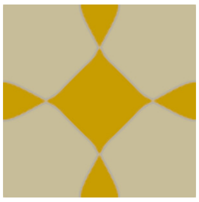
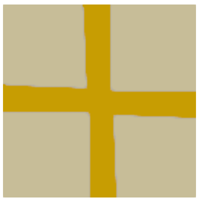
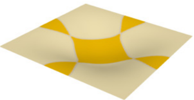
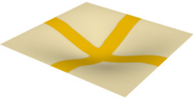
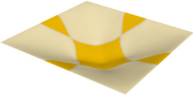
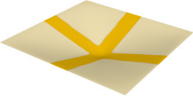
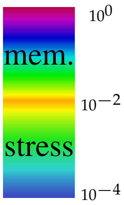
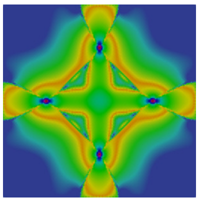
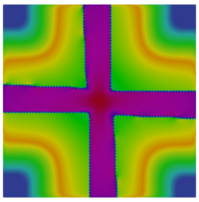
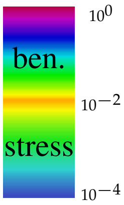
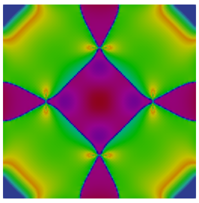
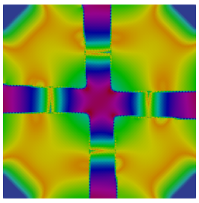
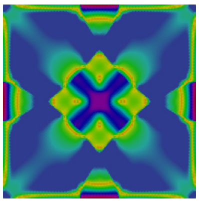
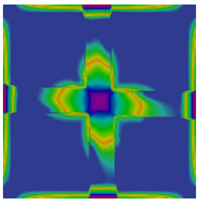
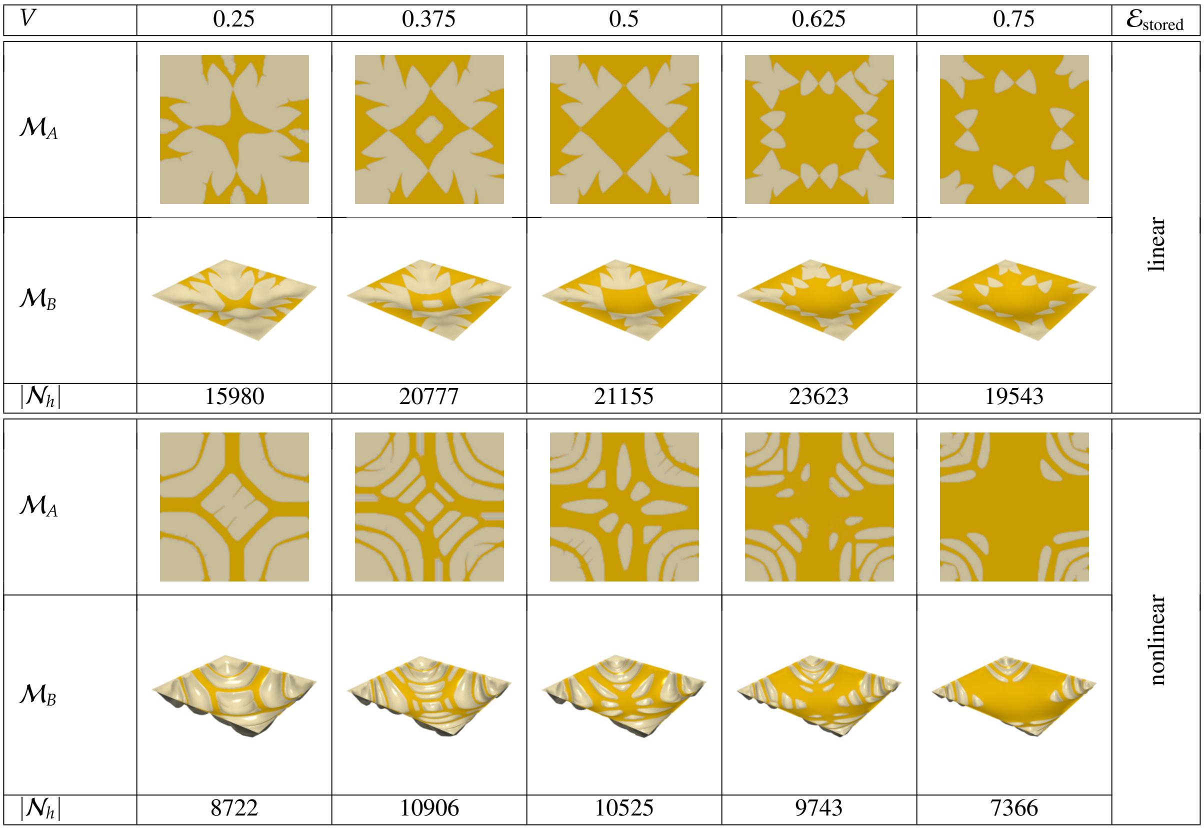
$\mathcal{E}_{\text{stored}}$	linear		nonlinear	
$\mathcal{M}_A$				
$\mathcal{M}_B$				
	linearized membrane stress is zero, since the force is acting into normal direction to the plate			
				
$\delta\,\mathcal{E}_{\text{mem}}$	0	0	0.03290	0.02212
$\delta^3\,\mathcal{E}_{\text{ben}}$	0.03186	0.04227	0.00266	0.00127
$\mathcal{E}_{\text{pot}}$	0.06372	0.08455	0.12701	0.08880

Figure 8.5: Comparison of two cross structures on a plate $\mathcal{M}_A = [0,1]^2$ for a centered load supported on $(0.45, 0.55)^2$ and acting into normal direction. We depict the corresponding membrane and bending stresses as averaged values over triangle elements using a color-code in logarithmic scaled HSV channel.

Constant Load on a Plate

Next, still for the flat case $\omega = \mathcal{M}_A = [0,1]^2$, we consider a force $f = (0,0,\beta)$ acting everywhere on the plate into normal direction for some constant β. Again, we assume clamped boundary conditions of the displacement on $\partial\mathcal{M}_A$. As above, we choose $\eta = 10^{-3}$ for the Modica–Mortola functional, $\delta = 10^{-2}$ for the thickness in the nonlinear case and $\delta = 10^{-1}$ for the thickness in the linear case. Furthermore, we compare different area constraints $V = \frac{k}{8}$ for $k = 2,3,4,5,6$ and set $\beta = -20V$ for the force. In Figure 8.6, we compare the results for linear elasticity after 9 adaptive refinement steps and nonlinear elasticity after 7 adaptive refinement steps. While for the centered load it has been sufficient to stabilize the area in the region, where the force is concentrated, by trusses connected to the boundary, for a constant load there is a need of microstructures to keep the deformation as small as possible in terms of the potential energy. Moreover, we observe significantly different results for linear and nonlinear elasticity.



V	0.25	0.375	0.5	0.625	0.75	$\mathcal{E}_{\text{stored}}$
$\mathcal{M}_A$						linear
$\mathcal{M}_B$						
$\lvert\mathcal{N}_h\rvert$	15980	20777	21155	23623	19543	
$\mathcal{M}_A$						nonlinear
$\mathcal{M}_B$						
$\lvert\mathcal{N}_h\rvert$	8722	10906	10525	9743	7366	

Figure 8.6: Optimal material distributions on a plate $\mathcal{M}_A = [0,1]^2$ for a constant load acting in normal direction and clamped boundary conditions on $\partial\mathcal{M}_A$. We compare the results for different area constraints V, where we take into account both linear and nonlinear elasticity.

Constant Load on a Hemisphere

Now, we investigate optimal material distributions on the upper hemisphere

$$\mathcal{M}_A = \{p_A = (p_1, p_2, p_3) \in \mathbb{R}^3 \ : \ p_1^2 + p_2^2 + p_3^2 = 1\,,\ p_3 \geqslant 0\}$$

parametrized by the unit disc $\omega = \{\xi \in \mathbb{R}^2 \ : \ \|\xi\| \leqslant 1\}$ as chart domain and the inverse of the stereographic projection $\psi_A(\xi_1, \xi_2) = \left(\frac{2\xi_1}{\xi_1^2+\xi_2^2+1}, \frac{2\xi_2}{\xi_1^2+\xi_2^2+1}, \frac{1-\xi_1^2-\xi_2^2}{\xi_1^2+\xi_2^2+1}\right)$ as chart map. We assume clamped boundary conditions on the left and right side, *i.e.*, we set $\Gamma_A = \{p_A = (p_1, p_2, p_3) \in \mathcal{M}_A \ : \ p_3 = 0\,,\ |p_1| \geqslant 0.9\}$. Here, we consider a single area constraint $V = 0.5\mathcal{V}_h(1)$. A force $f_A = (0, 0, \beta)$ with $\beta = 0.001$ is acting on the reference domain. For the thickness, we choose $\delta = 10^{-2}$. Then, we apply 8 adaptive refinement steps for linear elasticity and 6 adaptive refinement steps for nonlinear elasticity. In Figure 8.7, we compare different values for the parameter η to penalize the Modica–Mortola functional. For $\eta \to 0$, this should allow a larger perimeter for the optimal material distribution. Indeed, microstructures are emerging for smaller values of η. As for a constant load on the plate, we observe significantly different structures for linear and nonlinear elasticity, even though the force is relatively small.

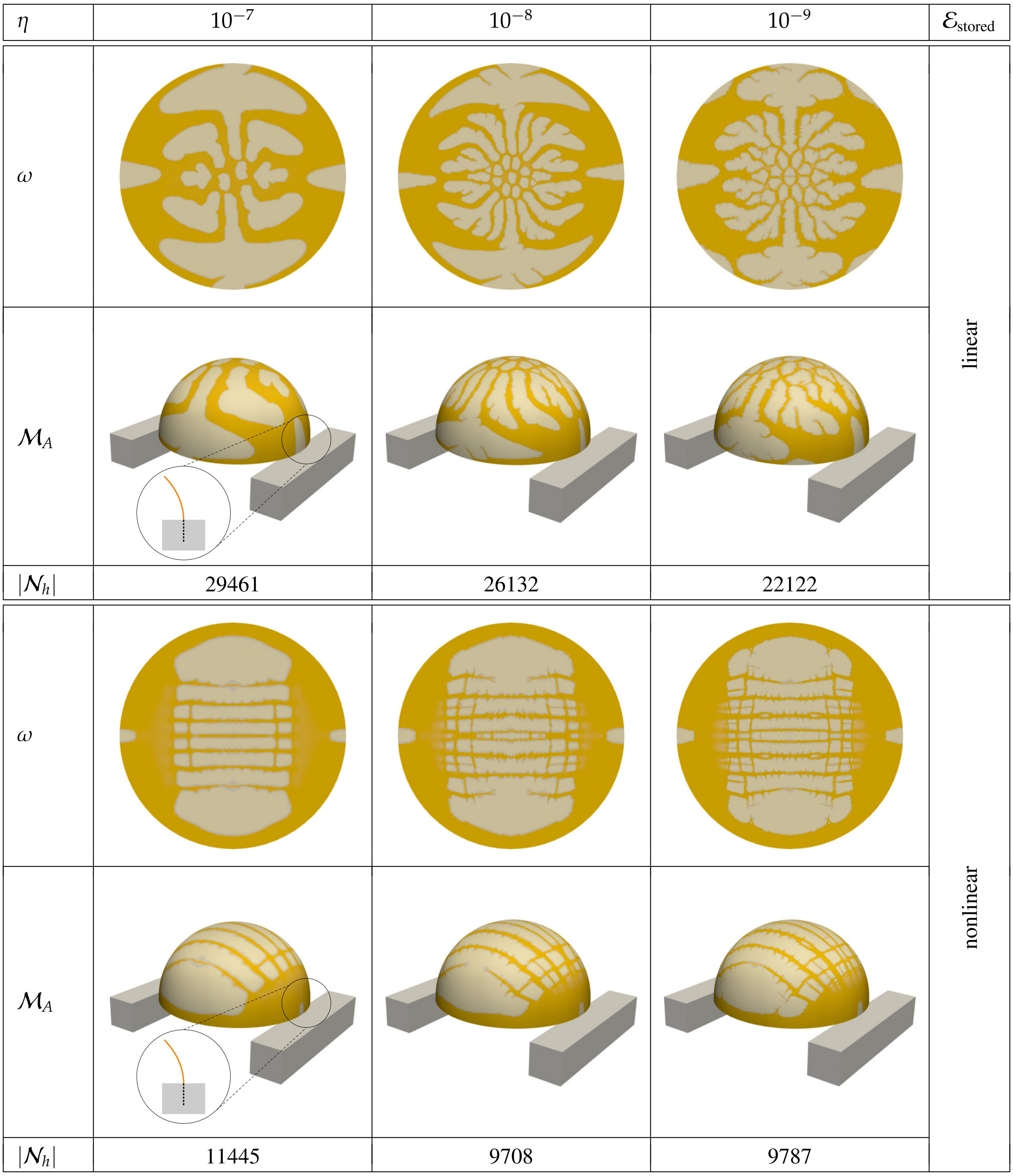


Figure 8.7: Optimal material distributions on a hemisphere. A constant load $f_A = (0, 0, 0.001)$ is applied, and an area constraint $V = 0.5\mathcal{V}_h(1)$ is enforced. We compare different values of the perimeter penalization term η. Top: Linear elasticity. Bottom: Nonlinear elasticity. Here, the left and right boundary are clamped as depicted for $\mathcal{M}_A$, where the clamped boundary condition is sketched for $\eta = 10^{-7}$.

Constant Load on a Half Cylinder

Finally, we consider the half cylinder

$$\mathcal{M}_A = \left\{ p_A = (p_1, p_2, p_3) \in \mathbb{R}^3 \;:\; p_2 \in [0,1]\,,\; p_1 > 0\,,\; \left(p_1 - \frac{1}{2\pi}\right)^2 + p_3^2 = \frac{1}{4\pi^2} \right\},$$

which we parametrize by $\omega = [0,1]^2$ as chart domain and $\psi_A(\xi_1, \xi_2) = \left(\frac{1}{2\pi}(1 - \cos(\pi\xi_1)), \xi_2, \frac{1}{2\pi}\sin(\pi\xi_1)\right)$ as chart map. We assume clamped boundary conditions on the left and right sides w.r.t. the p_2-direction, *i.e.*, we set $\Gamma_A = \{p_A = (p_1, p_2, p_3) \in \mathcal{M}_A \;:\; p_2 \in \{0,1\}\}$. Here, we restrict to nonlinear elasticity and study the effect for different thickness parameters δ. Then, we apply 6 adaptive refinement steps. We consider a single area constraint $V = 0.5\mathcal{V}_h(1)$. A force $f_A = (0,0,\beta)$ with $\beta = -10$ is acting on the reference domain. In Figure 8.8, we depict our numerically computed results. First, for a homogeneous material distribution with $v = 0$, we observe wrinkling effects for $\delta \to 0$. However, for the optimal material distribution, there is at least for the fixed choice of η, no increase of microstructure for $\delta \to 0$.

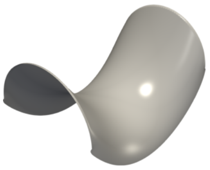
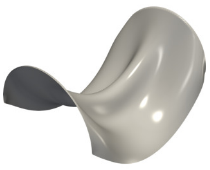
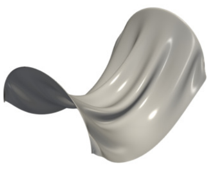
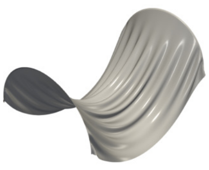
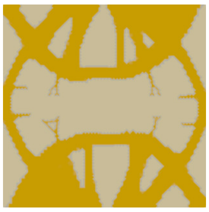
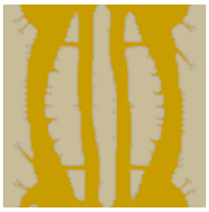
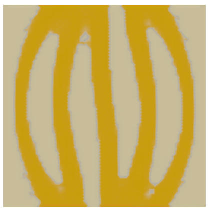
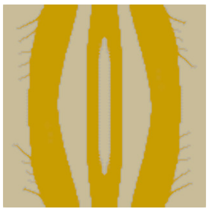
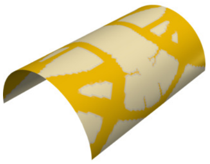
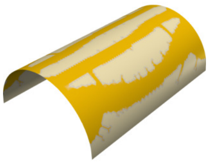
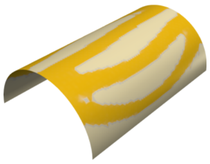
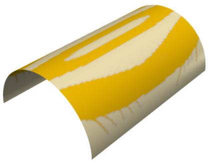
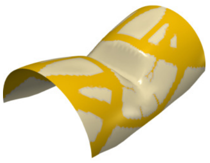
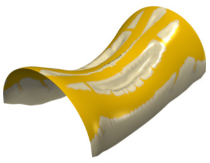
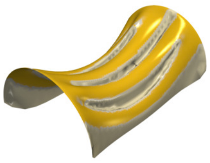
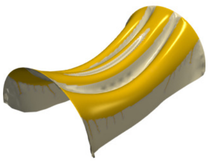
δ	10^{-1}	$10^{-1.5}$	10^{-2}	$10^{-2.5}$
$\mathcal{M}_B$ for homogeneous material $v = 0$				
ω				
$\mathcal{M}_A$				
$\mathcal{M}_B$				
$\lvert\mathcal{N}_h\rvert$	6161	6459	8476	5519

Figure 8.8: Optimal material distributions on a half cylinder, where the left and right sides are clamped. Here, a constant load $f_A = (0,0,-10)$ is acting on the reference domain. Furthermore, an area constraint $V = 0.5\mathcal{V}_h(1)$ on the amount of hard material is enforced. All results are computed for nonlinear elasticity. We compare different thickness parameters δ. Furthermore, we show solutions of the state equation for a homogeneous material distribution with $v_h = 0$.

8.3 Shape Design for Nonlinear Elastic Beams in 2D

In the following, we consider pure bending isometries of one-dimensional objects, which are obtained by dimension reduction of two-dimensional plates. We present a numerical scheme for the corresponding state equation and especially study local minimizer for a homogeneous material distribution. Then, as in Section 8.2, we investigate a shape optimization problem by computing the optimal material distribution on the one-dimensional object.

8.3.1 State Equation for Nonlinear Elastic Beams in 2D

We recall that the bending energy (8.5) was derived in [FJM02] via Γ-convergence. By a further dimension reduction, a similar Γ-convergence result was established in [MM03]. More precisely, the limit of the sequence $\frac{1}{\delta^2}\int_{(0,1)\times\delta S}\mathbb{W}(D\Phi)\,\mathrm{d}x_A$ was studied, where $S\subset\mathbb{R}^2$ is an open set and the hyperelastic energy density function is, *e.g.*, given by $\mathbb{W}(F)=\operatorname{dist}^2(F,SO(3))$. Instead, we briefly derive a one-dimensional model by taking into account the bending energy (8.5) for a two-dimensional plate $\mathcal{M}_A=(0,1)\times(-1,1)$, where a material distribution $\chi(\xi_1,\xi_2)=E(\xi_1)$ for a function $E\in L^\infty((0,1),[E^0,E^1])$ with $0<E^0<E^1<\infty$ is given. Moreover, we restrict the deformations to be of type $\phi(\xi_1,\xi_2)=(\gamma_1(\xi_1),\xi_2,\gamma_2(\xi_1))$ for some $\gamma\in W^{2,2}((0,1),\mathbb{R}^3)$ and assume clamped boundary condition at $\{0\}\times(-1,1)$. In this case, the isometry constraint $\nabla\phi^T\nabla\phi=\mathbb{1}$ simplifies to $|\gamma'|=1$. We denote by κ the curvature of γ. Then, for a stored elastic energy as in (8.5), a fixed material distribution E, and a force $f=\beta\,e_3=(0,0,\beta)^T$ acting in normal direction, the variational problem for the state equation becomes minimizing the free energy

$$\mathcal{E}_{\text{free}}(\gamma)=\int_0^1\frac{1}{2}E(t)\kappa(t)^2-\beta\gamma_2(t)\,\mathrm{d}t$$

over all $\gamma\in W^{2,2}((0,1),\mathbb{R}^3)$ with $\gamma(0)=0$ and $|\gamma'|=1$. Here, we neglect the thickness δ of the thin object, since scaling the bending energy $\mathcal{E}_{\text{ben}}(\gamma)=\int_0^1\frac{1}{2}E(t)\kappa(t)^2$ is equivalent to scaling the parameter β for the potential energy. The minimization problem in γ is still a constrained optimization problem involving second derivatives of the deformation. Now, in one dimension, we can make use of the phase

$$K(t)=\int_0^1\kappa(s)\,\mathrm{d}s\,,$$

and by identifying $\mathbb{R}^2$ with the complex plane $\mathbb{C}$, we can consider the arc length parametrization

$$\gamma(t)=\int_0^t e^{i(K(s)+K_0)}\,\mathrm{d}s\,,$$

where $K_0\in\mathbb{R}$ is the slope of γ at $t=0$, *i.e.*, $\gamma'(0)=e^{iK_0}$. Such an arc length parametrization automatically satisfies the isometry constraint $|\gamma'(t)|=1$. Furthermore, the potential energy becomes

$$\begin{aligned}\mathcal{E}_{\text{pot}}(K)&=\int_0^1\beta\gamma_2(t)\,\mathrm{d}t=\int_0^1\beta\int_0^t\sin(K(s)+K_0)\,\mathrm{d}s\,\mathrm{d}t\\&=\int_0^1\int_s^1\beta\sin(K(s)+K_0)\,\mathrm{d}t\,\mathrm{d}s=\int_0^1(1-s)\beta\sin(K(s)+K_0)\,\mathrm{d}s\,.\end{aligned}$$

Thus, we can rewrite the above constrained minimization problem in terms of K by minimizing the free energy

$$\mathcal{E}_{\text{free}}(K)=\int_0^1\frac{1}{2}E(t)(K'(t))^2-\beta(1-t)\sin(K(t)+K_0)\,\mathrm{d}t$$

over all phases K in the space

$$X_K:=\left\{K\in W^{1,2}([0,1],[-\pi,\pi)):K(0)=0\right\}\,.$$

Then, we can state the following existence and partial uniqueness result.

Proposition 8.3.1 (Existence and Uniqueness of Minimizer)**.** *Let* $E \in L^\infty((0,1),[E^0,E^1])$ *and let* $\beta < 0$.

1. *There exists a global minimizer of* $\mathcal{E}_{free}$ *within* X_K.
2. *If in addition* $K_0 \in [-\pi,0]$, *there exists a unique global minimizer of* $\mathcal{E}_{free}$ *within* X_K.

Proof. The existence of a minimizer follows by the direct method in the calculus of variations. The uniqueness follows by convexity of the sine function on $[-\pi,0]$ and the fact that a minimizer K of $\mathcal{E}_{\text{free}}$ with $K_0 \in [-\pi,0]$ satisfies $K(t)+K_0 \in [-\pi,0]$ for all $t \in (0,1]$. The second statement follows by restricting to the case $K_0 \in [-\frac{\pi}{2},0]$ (or in analogy to $K_0 \in [-\pi,-\frac{\pi}{2}]$). Assuming that K is a minimizer with $K > 0$ on a subinterval $J \subset [0,1]$ of maximal length, we can define $\tilde{K}(t) := \chi_{[0,1]\setminus J} K$, which contradicts the minimality of K, since

$$0 \leqslant \mathcal{E}_{\text{free}}(\tilde{K}) - \mathcal{E}_{\text{free}}(K) = \int_J -\frac{1}{2}E(t)(K'(t))^2\,\mathrm{d}t - \int_J \beta(1-t)\sin(K_0) - \sin(K(t)+K_0)\,\mathrm{d}t < 0\,.$$

For details, we refer the reader to [HRS19, Proposition 3.5]. □

8.3.2 Shape Optimization for Nonlinear Elastic Beams in 2D

Now, we consider a characteristic function $\chi \in L^\infty((0,1),\{0,1\})$ describing a material distribution by

$$E(\chi) = E^1\chi + E^0(1-\chi)\,.$$

Then we define the free energy in terms of χ and K as

$$\mathcal{E}_{\text{free}}(\chi,K) = \int_0^1 \frac{1}{2}E(\chi)(t)(K'(t))^2 - \beta(1-t)\sin(K(t)+K_0)\,\mathrm{d}t\,.$$

We aim at minimizing a cost functional

$$\mathcal{J}^\alpha_{\text{tot}}(\chi) = \mathcal{J}^\alpha_{\text{expl}}(\chi,K(\chi))$$

over all characteristic functions $\chi \in L^\infty((0,1),\{0,1\})$, where $K(\chi)$ is a stationary point of $\mathcal{E}_{\text{free}}$. The cost functional $\mathcal{J}^\alpha_{\text{expl}}$ explicitly depending on χ and K is defined as

$$\begin{aligned}\mathcal{J}^\alpha_{\text{expl}}(\chi,K) &= \mathcal{E}_{\text{pot}}(\chi,K) + \alpha\mathcal{V}(\chi)\\ &= \int_0^1 -\beta(1-t)\sin(K(t)+K_0)\,\mathrm{d}t + \alpha\int_0^1 \chi\,\mathrm{d}t\,.\end{aligned}$$

Here, $\alpha > 0$ is a parameter penalizing the amount of hard material. Since weak-* limits of characteristic functions in general only belong to the larger space $L^\infty((0,1),[0,1])$, relaxation is required. We apply relaxation by the homogenization method, which is for a one-dimensional family of parameters simply given by the harmonic mean (see Theorem 6.2.3). Thus, we define for $\chi \in L^\infty((0,1),[0,1])$ the homogenized material coefficient

$$E^*(\chi) = \left(\frac{\chi}{E^1} + \frac{1-\chi}{E^0}\right)^{-1}\,.$$

This allows to extend $\mathcal{E}_{\text{free}}$ and thus $\mathcal{J}^\alpha_{\text{expl}}$ for $\chi \in L^\infty((0,1),[0,1])$.

For the simple scenario with initial slope $K_0 \in [-\pi,0]$, where uniqueness of global minimizer of $\mathcal{E}_{\text{free}}$ is guaranteed, the following classification result for optimal designs is established.

Theorem 8.3.2 (Classification of Optimal Designs)**.** *Let* $K_0 \in [-\pi,0]$. *Then the optimal design is classical and ordered. More precisely, if* χ *is a critical point of* $\mathcal{J}^\alpha_{tot}$ *within* $L^\infty((0,1),[0,1])$, *then there exists* $t^* \in (0,1)$ *s.t.* $\chi = 1$ *a.e. on* $(0,t^*)$ *and* $\chi = 0$ *a.e. on* $(t^*,1)$.

Proof. See [HRS19, Theorem 5.8] □

Note that we here consider a penalization of the amount of hard material instead of a constraint on the length. This amount exactly corresponds to the value t^*.

8.3.3 Phase-Field Approximation and Finite Element Discretization

In the following, we present a numerical solution scheme to compute solutions to the state equation in the phase variable K. Proposition 8.3.1 always guarantees the existence of a minimizer, but uniqueness is only provided if $K_0 \in [-\pi, 0]$. Thus, we are especially interested in the case $K_0 > 0$.

Here, we apply a phase-field approach. More precisely, we take into account a phase-field function $v \in W^{1,2}([0,1], \mathbb{R})$ and approximate the material coefficient E in terms of v by the harmonic mean

$$E(v) = 2\left(\frac{1+v}{E^1} + \frac{1-v}{E^0}\right)^{-1} .$$

Moreover, we approximate the length covered by hard material by

$$\mathcal{V}(v) = \int_0^1 \frac{v+1}{2}\,\mathrm{d}t .$$

Note that the classification result in Theorem 8.3.2 is obtained for a cost functional without perimeter regularization. However, for numerical purpose, we use the $1D$ version of the Modica–Mortola functional $\mathcal{A}^\epsilon$ as defined in (6.12) as regularizer to ensure the phase-field function to be smooth and essentially to take values $v \in \{-1, 1\}$. Altogether, this allows defining the augmented compliance functional as

$$\mathcal{J}^{\alpha,\eta}_{\mathrm{expl}}(v, K) = \int_0^1 \beta(1-t)\sin(K(t) + K_0)\,\mathrm{d}t + \alpha\mathcal{V}(v) + \eta\mathcal{A}^\epsilon(v) ,$$

with coefficients $\alpha, \eta > 0$. Thus, the total cost functional in terms of a phase-field function is given by

$$\mathcal{J}^{\alpha,\eta}_{\mathrm{tot}}(v) = \mathcal{J}^{\alpha,\eta}_{\mathrm{expl}}(v, K(v)) ,$$

where $K(v)$ is a solution to the state equation $\partial_K \mathcal{E}_{\mathrm{free}}(v, K)(\widehat{K}) = 0$ for all test functions $\widehat{K} \in X_K$. To compute a local minimizer of $\mathcal{J}^{\alpha,\eta}_{\mathrm{tot}}$, we can apply the same approach as in Section 6.3.2 by solving a corresponding adjoint problem

$$\partial^2_{K,K}\mathcal{E}_{\mathrm{free}}(v, K(v))(\widehat{K})(A) = -\partial_K \mathcal{J}^{\alpha,\eta}_{\mathrm{expl}}(v, K(v))(\widehat{K}) \quad \forall \widehat{K} \in X_K$$

in the adjoint variable $A \in X_K$. Then we obtain the derivative

$$\frac{d}{dv}\mathcal{J}^{\alpha,\eta}_{\mathrm{tot}}(v)(\hat{v}) = \partial_v \mathcal{J}^{\alpha,\eta}_{\mathrm{expl}}(v, K(v))(\hat{v}) + \partial^2_{v,K}\mathcal{E}_{\mathrm{free}}(v, K(v))(A)(\hat{v}) .$$

For the state equation, we apply Newton's method to find local minimizers of the free energy. This requires to compute the first and second derivative $D\mathcal{E}_{\mathrm{free}}(v, K)$ and $D^2\mathcal{E}_{\mathrm{free}}(v, K)$ of the stored energy. For the numerical implementation, we use piecewise affine and continuous finite element functions. More precisely, we consider an equidistant grid with N nodes $t_n = \frac{n}{N-1}$ for $n = 0, \ldots, N-1$ and associated $N-1$ cells $C_n = (t_{n-1}, t_n)$ for $n = 1, \ldots, N-1$. The corresponding grid width is given by $h = \frac{1}{N-1}$. Then we approximate the phase K by a finite element function K_h in the space

$$V^1_h([0,1]) := \left\{K_h \in C^0([0,1]) \ : \ K_h\big|_{C_n} \text{ is affine for all } n = 1, \ldots, N-1\right\} .$$

Moreover, we approximate the phase-field variable v by a finite element function $v_h \in V^1_h([0,1])$. For the numerical integration, we choose a Gaussian quadrature with $Q = 5$ quadrature points per element. Applying this quadrature to the free energy and its derivatives, we get a discrete free energy $\mathcal{E}_{\mathrm{free},h}$ on $V^1_h([0,1]) \times V^1_h([0,1])$ and associated derivatives $D\mathcal{E}_{\mathrm{free},h}$ and $D^2\mathcal{E}_{\mathrm{free},h}$. Finally, for a fixed material distribution $v_h \in V^1_h([0,1])$, we compute a solution to $\partial_{K_h}\mathcal{E}_{\mathrm{free},h} = 0$ with Newton's method. To cope with the nonlinearity, we use a multilevel scheme, by first solving the problem on a coarse grid, prolongate the obtained result onto a finer grid, and proceed iteratively. Here, we take into account a dyadic sequence $N_l = 2^l + 1$ with $l = L_c, \ldots, L_f$, where we use $L_c = 3$ and $L_f = 10$.

Furthermore, using the above discretization we obtain a discrete operator $\mathcal{J}^{\alpha,\eta}_{\mathrm{tot},h}$ to approximate the total cost functional and the corresponding derivative, s.t. we can apply a Quasi-Newton method (BFGS) to compute minimizers of $\mathcal{J}^{\alpha,\eta}_{\mathrm{tot},h}$. We emphasize that we have to impose the Dirichlet boundary condition for the adjoint variable, *i.e.*, $A_h(0) = 0$. Moreover, for a given phase-field function v_h, we note that $K_h(v_h)$ is an element of the set of solutions to the state equation. Thus, starting with some initial phase, the Newton method converges to a state $K_h(v_h)$, which depends upon this initialization.

8.3.4 Numerical Results for Nonlinear Elastic Beams in 2D

Now, we present our numerically computed results. First, we consider solutions to the state equation for a homogeneous material distribution. Then, we compute optimal material distributions, where we initialize our optimization scheme with different solutions to the state equation.

Different Solutions to the State Equation

For a homogeneous material $E = 1$, we experimentally observe essentially three types of stationary points (see Figure 8.9). First, there is of course a simple configuration where the curve is just turning downwards. In fact, this appears to be an approximation of the global minimizer of the energy functional $\mathcal{E}_{\mathrm{free},h}$. Secondly, we get a twisted curve. We observe that these two configurations are stable under a change of material, *i.e.*, taking some simple (resp. twisted) beam as initialization for a different material, the computed discrete solution in our experiments always turned out to be a simple (resp. twisted) beam again. However, there is also a highly unstable configuration in between, where the beam neither decides to fall towards the left side nor towards the right side.

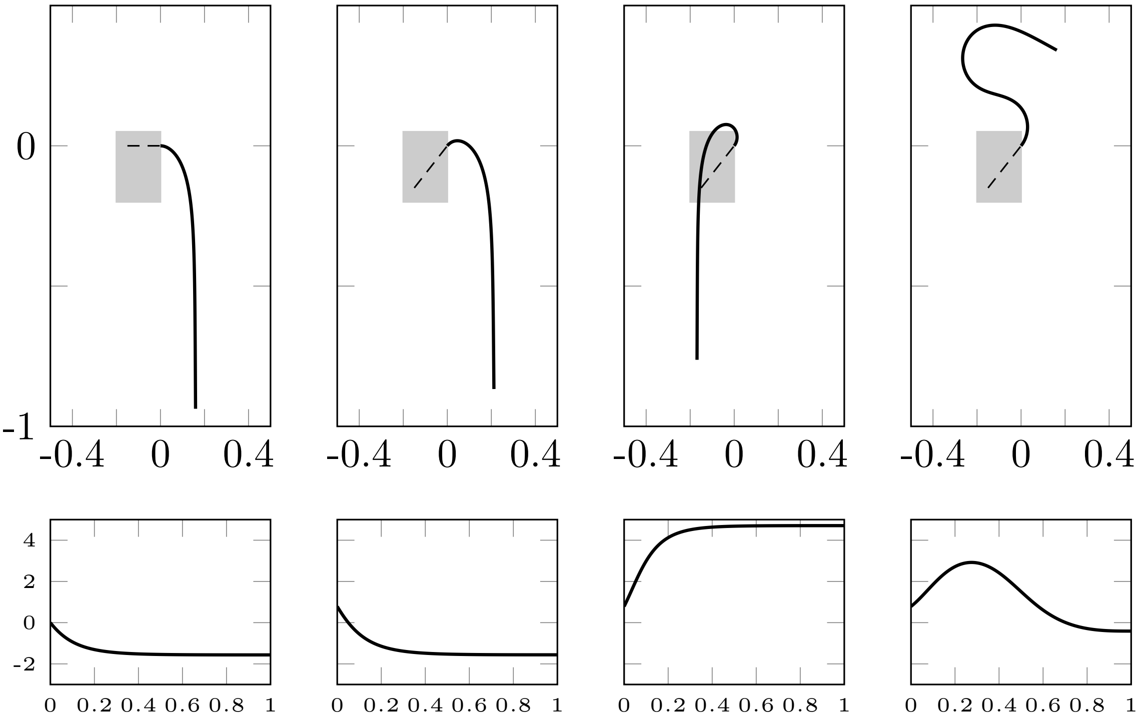


Figure 8.9: Different solutions to the state equation (top row) with corresponding phase variable K (bottom row). From left to right: simple configurations with $K_0 = 0$ and $K_0 = \frac{\pi}{4}$, a twisted beam with $K_0 = \frac{\pi}{4}$, and an S-shaped configuration with $K_0 = \frac{\pi}{4}$. The clamped boundary conditions are indicated by dotted lines. Here, we have chosen $\beta = -100$, $E = 1$.

Optimal Material Distributions for Different Scenarios

In Figure 8.10, we show our numerical results for optimal material distributions under different initial conditions. First, the optimal design for an initial slope K_0 reflects the classification result from Theorem 8.3.2. Furthermore, the optimal material distributions for initial slopes $K_0 \neq 0$ and different solutions to the state equation suggest that the classification result can be extended to more general assumptions. In fact, in our numerical simulations for clamped boundary conditions at $t = 0$, the optimal design is always given by the hard material on the left, *i.e.*, in some interval $[0, t^*]$.

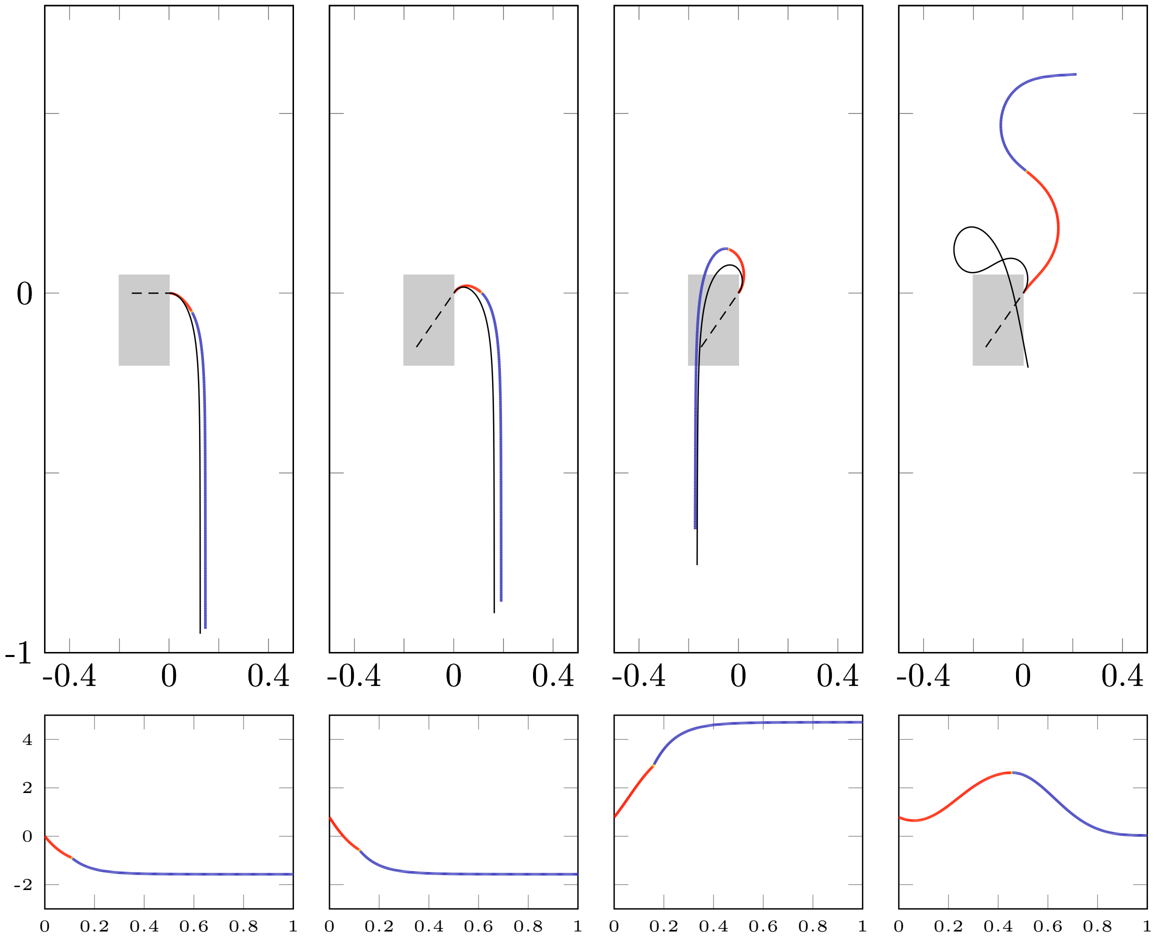


Figure 8.10: Top: Starting from different initializations with $v = 0$ (black), we obtain optimal designs (from left to right) for a simple configuration with $K_0 = 0$ and with $K_0 = \frac{\pi}{4}$, as well as a twisted configuration with $K_0 = \frac{\pi}{4}$, and an S-configuration with $K_0 = \frac{\pi}{4}$. The clamped boundary conditions are indicated by dotted lines. Bottom: We see the corresponding plots of the phase K. Here, we have chosen $E^1 = 1$, $E^0 = 0.5$, $\beta = -100$, $\alpha = 1$, $\eta = 1$, $N = 513$, and $\epsilon = \frac{1}{N-1}$. Both the curve and the phase are colored according to the phase-field variable v, where red denotes hard material $(v = 1)$ and blue denotes soft material $(v = -1)$.

Shape Optimization with Pointwise Conditions

Finally, we implement additional constraints prescribing a set of beam positions on $(0, 1]$. In this case, we obtain that the resulting optimal design is characterized by separated subintervals with hard material. Hence, also in this more general setup, we do not observe microstructures, even for small values of η. In Figure 8.11, we show an instance of these computational results with additional point constraints.

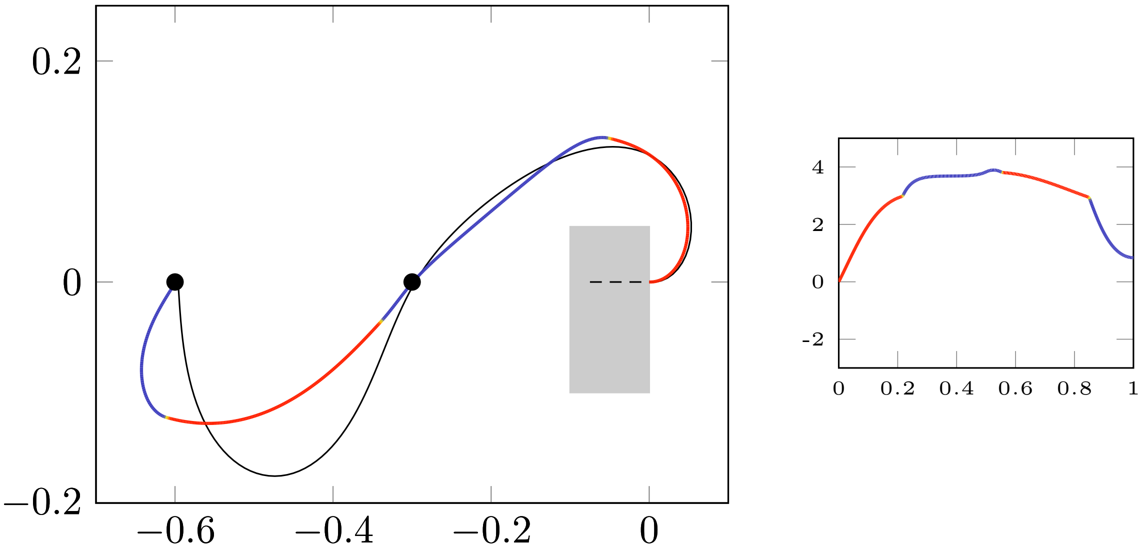


Figure 8.11: Left: Optimal design for a beam under the constraint that two fixed beam positions $(-0.3, 0)$ and $(-0.6, 0)$ are achieved for $t = 0.5, 1$, respectively. Here, the initial slope is given by $K_0 = 0$ and we choose parameters $E^1 = 4.0$, $E^0 = 0.5$, $\beta = -100$, $\alpha = 1$, $\eta = 1$, $N = 513$, and $\epsilon = \frac{1}{N-1}$. Right: The corresponding phase K. Both the curve and the phase are colored according to the phase-field variable v, where red denotes hard material ($v = 1$) and blue denotes soft material ($v = -1$).

8.4 Shape Design for Bending Isometries of Plates

Now, we consider pure bending isometries for two-dimensional plates. In analogy to Section 8.2 and Section 8.3, we investigate the optimal material distribution.

8.4.1 Shape Optimization Problem for Bending Isometries

State Equation. We consider a two-dimensional plate $\omega = \mathcal{M}_A \subset \mathbb{R}^2$. Moreover, a material distribution on ω is described by a phase-field $v \in W^{1,2}(\omega, [-1, 1])$. Then, for a smooth isometry ψ_B and the density function $\mathbb{W}_{\text{ben}}$ as in (8.7), we recall from Proposition 8.1.3 that

$$\mathbb{W}_{\text{ben}}(v, \psi_B) := \frac{E(v)}{24} \| g_A^{-1}(h_B - h_A) \|_F^2 = \frac{E(v)}{24} \| h_B \|_F^2 = \frac{E(v)}{24} |D^2 \psi_B|^2 = \frac{E(v)}{24} |D^2 u|^2 \, .$$

Thus, we can define the stored elastic energy in terms of the displacement $u = \psi_B - \psi_A$ as

$$\mathcal{E}_{\text{stored}}^{\text{biso}}(v, u) \;=\; \begin{cases} \mathcal{E}_{\text{ben}}(v, u) & \text{if } u \in W^{2,2}_{\text{iso}}(\omega, \mathbb{R}^3)\,, \\ \infty & \text{otherwise,} \end{cases} \;=\; \begin{cases} \displaystyle\int_\omega \frac{E(v)}{24} |D^2 u|^2 \, \mathrm{d}\xi & \text{if } u \in W^{2,2}_{\text{iso}}(\omega, \mathbb{R}^3)\,, \\ \infty & \text{otherwise.} \end{cases}$$

Then, for a given force $f_A \colon \omega \to \mathbb{R}^3$ and a fixed phase-field v, we aim to minimize the free energy

$$\mathcal{E}_{\text{free}}(v, u) = \mathcal{E}_{\text{ben}}(v, u) - \mathcal{E}_{\text{pot}}(u) \tag{8.10}$$

over all displacements u in the space

$$W^{2,2}_{\text{iso},\Gamma_\omega}(\omega, \mathbb{R}^3) := \left\{ u \in W^{2,2}(\omega, \mathbb{R}^3) \ : \ g_A^{-1} g_B = \mathbb{1} \, , \ u = 0 \text{ on } \Gamma_\omega \, , \ \nabla u = 0 \text{ on } \Gamma_\omega \right\} \, .$$

As for the elastic beams in Section 8.3, we neglect the thickness δ, since scaling the force f_A would be equivalent.

Cost Functional. In analogy to the shape optimization problem (8.8) for a mixed membrane-bending model, we aim at minimizing a total cost functional

$$\mathcal{J}^{\eta}_{\mathrm{tot}}(v) = \mathcal{J}_{\mathrm{expl}}(v, u(v)) + \eta\mathcal{A}^{\epsilon}(v)\,,$$

over all phase-fields $v \in W^{1,2}(\omega, [-1, 1])$ satisfying the area constraint $\mathcal{V}(v) = V$, where $u(v)$ is the displacement of a pure bending isometry defined as a minimizer of the free energy in (8.10). For the cost functional $\mathcal{J}_{\mathrm{expl}}(v, u) = \mathcal{E}_{\mathrm{pot}}(u)$, we choose the potential energy.

8.4.2 Finite Element Discretization for Bending Isometries

As for the mixed models in Section 8.2, we choose discrete phase-fields $v_h \in V^1_h(\omega)$ and take into account the DKT element to discretize displacements $u_h \in \mathcal{W}_{h,\Gamma_\omega}(\omega)^3$ with clamped boundary conditions. Then, we apply a Gaussian quadrature of degree 6 with $Q = 12$ quadrature points for each triangle element with weights ω, and obtain a discrete bending energy

$$\mathcal{E}_{\mathrm{ben},h}(v_h, u_h) = \sum_{T\in\mathcal{T}_h}\sum_{q=1,\dots,Q}\omega(q)\frac{E(v_h(q))}{24}\|\nabla\nabla_h u_h(q)\|^2\,.$$

Consequently, a discrete free energy is given by $\mathcal{E}_{\mathrm{free},h}(v_h, u_h) = \mathcal{E}_{\mathrm{ben},h}(v_h, u_h) - \mathcal{E}_{\mathrm{pot},h}(u_h)$. Note that in the continuous setup the isometry constraint in terms of a displacement u is pointwise given by

$$0 = g_B - g_A = Du^T D\psi_A + D\psi_A^T Du + Du^T Du\,,$$

which simplifies in the case $\psi_A = \mathrm{id}$ to

$$0 = \begin{pmatrix} 2\partial_1 u_1 + \sum_{i=1}^3(\partial_1 u_i)^2 & \partial_2 u_1 + \partial_1 u_2 + \sum_{i=1}^3 \partial_1 u_i \partial_2 u_i \\ \mathrm{sym} & 2\partial_2 u_2 + \sum_{i=1}^3(\partial_2 u_i)^2 \end{pmatrix}. \tag{8.11}$$

In our numerical method we enforce the isometry constraint nodalwise, which can be easily formulated due to the degrees of freedom for derivative values of u_h. This approach was already applied in [Bar13], where a linearization of the isometry constraint was proposed. Instead, we implement a Newton method for a corresponding Lagrangian. Therefore, we denote by $\mathcal{N}^{\mathrm{int}}_h = \mathcal{N}_h\backslash\Gamma_\omega$ the set of interior nodes and consider for each $z \in \mathcal{N}^{\mathrm{int}}_h$ Lagrange multipliers $\lambda_h(z) = \left(\lambda^1_h(z), \lambda^2_h(z), \lambda^{12}_h(z)\right)$ for each of the three constraints in (8.11), *i.e.*, we define

$$\begin{aligned}
\mathrm{Iso}^1_h(u_h)(z) &= 2\partial_1 u_h(z)\cdot e_1 + \sum_{i=1}^3(\partial_1 u_h(z)\cdot e_i)^2\,,\\
\mathrm{Iso}^2_h(u_h)(z) &= 2\partial_2 u_h(z)\cdot e_2 + \sum_{i=1}^3(\partial_2 u_h(z)\cdot e_i)^2\,,\\
\mathrm{Iso}^{12}_h(u_h)(z) &= \partial_2 u_h(z)\cdot e_1 + \partial_1 u_h(z)\cdot e_2 + \sum_{i=1}^3(\partial_1 u_h(z)\cdot e_i)(\partial_2 u_h(z)\cdot e_i)\,,\\
\mathrm{Iso}_h(u_h, \lambda_h) &= \sum_{z\in\mathcal{N}^{\mathrm{int}}_h}\lambda^1_h(z)\mathrm{Iso}^1_h(u_h)(z) + \lambda^2_h(z)\mathrm{Iso}^2_h(u_h)(z) + \lambda^{12}_h(z)\mathrm{Iso}^{12}_h(u_h)(z)\,.
\end{aligned}$$

Note that all the values $\partial_j u_h(z)_i$ for $j = 1, 2$ and $i = 1, 2, 3$ are explicit degrees of freedom. Finally, the Lagrangian is given by

$$\mathcal{L}_{\mathrm{ben},h}(v_h, u_h, \lambda_h) = \mathcal{E}_{\mathrm{ben},h}(v_h, u_h) - \mathcal{E}_{\mathrm{pot},h}(u_h) + \mathrm{Iso}_h(u_h, \lambda_h)\,.$$

Then, to compute for a fixed material distribution given by v_h solutions to the state equation we apply Newton's method to solve

$$\partial_{(u_h,\lambda_h)}\mathcal{L}_{\mathrm{ben},h}(v_h, u_h, \lambda_h) = 0\,. \tag{8.12}$$

Now, the above finite element discretization allows to define a discrete cost functional $\mathcal{J}^{\eta}_{\mathrm{tot},h}$. Again, we aim at computing minimizer of $\mathcal{J}^{\eta}_{\mathrm{tot},h}$ with a first-order method, which requires to evaluate the derivative

$$\frac{d}{dv_h}\mathcal{J}^{\eta}_{\mathrm{tot},h}(v_h)(\widehat{v_h}) = \partial_{v_h}\mathcal{J}^{\eta}_{\mathrm{expl},h}(v_h, u_h(v_h))(\widehat{v_h}) + \partial_{u_h}\mathcal{J}^{\eta}_{\mathrm{expl},h}(v_h, u_h(v_h))(\partial_{v_h}u_h(v_h)(\widehat{v_h}))\,.$$

Therefore, we want to apply the same approach as described in Section 6.3.2 by solving a suitable adjoint problem. Differently, the state equation is now defined by stationary points of the Lagrangian $\mathcal{L}_{\mathrm{ben},h}$ and we have to incorporate the Lagrange multiplier λ_h in the computation. For fixed v_h, we denote by $(u_h, \lambda_h)(v_h)$ a solution to (8.12). Then, by the inverse function theorem we have

$$\partial_{v_h}\mathcal{L}_{\mathrm{ben},h}\left(v_h, (u_h, \lambda_h)(v_h)\right) = -\left(\partial^2_{(u_h,\lambda_h),(u_h,\lambda_h)}\mathcal{L}_{\mathrm{ben},h}\left(v_h, (u_h, \lambda_h)(v_h)\right)\right)^{-1}\partial^2_{v_h,(u_h,\lambda_h)}\mathcal{L}_{\mathrm{ben},h}\left(v_h, (u_h, \lambda_h)(v_h)\right)\,.$$

Thus, we introduce an adjoint problem for variables $(A_{u_h}, A_{\lambda_h}) \in \mathcal{W}_{h,\Gamma_\omega}(\omega)^3 \times \mathbb{R}^{3|\mathcal{N}_h^{\mathrm{int}}|}$ by solving the linear system

$$\partial^2_{(u_h,\lambda_h),(u_h,\lambda_h)}\mathcal{L}_{\mathrm{ben},h}(v_h, (u_h, \lambda_h)(v_h))(\widehat{u_h}, \widehat{\lambda_h})(A_{u_h}, A_{\lambda_h}) = -\partial_{(u_h,\lambda_h)}\mathcal{J}^{\eta}_{\mathrm{expl},h}(v_h, (u_h, \lambda_h)(v_h))(\widehat{u_h}, \widehat{\lambda_h})$$

for all $(\widehat{u_h}, \widehat{\lambda_h}) \in \mathcal{W}_{h,\Gamma_\omega}(\omega)^3 \times \mathbb{R}^{3|\mathcal{N}_h^{\mathrm{int}}|}$. This allows computing

$$\begin{aligned}\frac{d}{dv_h}\mathcal{J}^{\eta}_{\mathrm{tot},h}(v_h)(\widehat{v_h}) &= \partial_{v_h}\mathcal{J}^{\eta}_{\mathrm{expl},h}(v_h, (u_h, \lambda_h)(v_h))(\widehat{v_h}) + \partial_{u_h}\mathcal{J}^{\eta}_{\mathrm{expl},h}(v_h, (u_h, \lambda_h)(v_h))(\partial_{v_h}(u_h, \lambda_h)(v_h)(\widehat{v_h}))\\ &= \partial_{v_h}\mathcal{J}^{\eta}_{\mathrm{expl},h}(v_h, (u_h, \lambda_h)(v_h))(\widehat{v_h}) + \partial^2_{v_h,(u_h,\lambda_h)}\mathcal{L}_{\mathrm{ben},h}(v_h, (u_h, \lambda_h)(v_h))(A_{u_h}, A_{\lambda_h})(\widehat{v_h})\,.\end{aligned}$$

Since $\partial^2_{v_h,(u_h,\lambda_h)}\mathcal{L}_{\mathrm{ben},h}$ is given by

$$\partial^2_{v_h,(u_h,\lambda_h)}\mathcal{L}_{\mathrm{ben},h}(v_h, u_h, \lambda_h) = \begin{pmatrix}\partial^2_{v_h,u_h}\mathcal{E}_{\mathrm{ben},h}(v_h, u_h) & 0\\ 0 & 0\end{pmatrix}\,,$$

the expression for the shape derivative simplifies to

$$\frac{d}{dv_h}\mathcal{J}^{\eta}_{\mathrm{tot},h}(v_h)(\widehat{v_h}) = \partial_{v_h}\mathcal{J}^{\eta}_{\mathrm{expl},h}(v_h, (u_h, \lambda_h)(v_h))(\widehat{v_h}) + \partial^2_{v_h,u_h}\mathcal{E}_{\mathrm{ben},h}(v_h, u_h(v_h))(A_{u_h})(\widehat{v_h})\,.$$

Then, as for the mixed membrane-bending models in Section 8.2, we apply the `IPOPT` package [WB06] to compute minimizer of the fully discrete cost functional $\mathcal{J}^{\eta}_{\mathrm{tot},h}$ over all $v_h \in V^1_h(\omega, [-1, 1])$ with the area constraint $\mathcal{V}_h(v_h) = V$. In the adaptive refinement scheme, we additionally mark those elements $T \in \mathcal{T}_h$, where the isometry error $\int_T |g_B - g_A|^2\,\mathrm{d}x$ is large. More precisely, we compute this error for all elements and select the largest 25% for the longest edge bisection refinement.

8.4.3 Numerical Results for Bending Isometries

Now, we present our numerical results for the shape optimization problem for bending isometries of plates. In the $1D$ case of elastic beams, for a constant load scenario with a force $f = (0, 0, \beta)$ and clamped boundary on the left side, we recall from Section 8.3 that the optimal material distribution on the interval $[0, 1] \times \{0\}$ is given by an ordered design with the hard phase on the interval $(0, V)$ at the clamped boundary (see Theorem 8.3.2). Here, we consider the same scenario in 2D, *i.e.*, given is a plate $[0, 1]^2$ with clamped boundary at $\{0\} \times [0, 1]$ and a constant load $f = (0, 0, \beta)$ is acting in orthogonal direction. However, there are material distribution on $[0, 1]^2$, which cannot be represented in the $1D$ case, and thus, it is unclear whether the optimal material distribution for the $1D$ case is still optimal for the 2D case. In the following, we first compare three different classical designs w.r.t. the potential energy. Afterwards, we compute optimal designs for small and large forces. As in Section 8.2, the hard material is colored in orange.

Comparison of Different Designs

We define three different material distributions, where, depending on the area V, the subdomain covered with hard material is given by

(I) a layer $[0, V] \times [0, 1]$ at the clamped boundary, *i.e.*, the solution to the $1D$ problem,

(II) a layer $[0, 1] \times [0.5 - 0.5V, 0.5 + 0.5V]$ orthogonal to the clamped boundary, and

(III) a square $[0, \sqrt{V}] \times [0.5 - 0.5\sqrt{V}, 0.5 + 0.5\sqrt{V}]$ centered in the middle of the clamped boundary.

Here, we consider three area fractions $V = 0.25, 0.5, 0.75$ for the amount of hard material. In Figure 8.12, we compare the potential energy of these three designs in dependence of $|f|$. For all computations, we use a mesh of $|\mathcal{N}_h| = 16641$ nodes. We observe that for a large area fraction $V = 0.75$ of the hard material, the $1D$ optimizer (I) is optimal w.r.t. the potential energy independent of $|f|$. For an area fraction $V = 0.5$, and small forces, design (III) is optimal. For an area fraction $V = 0.25$, we even obtain that design (II) is optimal for small forces and design (III) is better on an intermediate range. In any cases, it seems that for large forces design (I) is optimal.

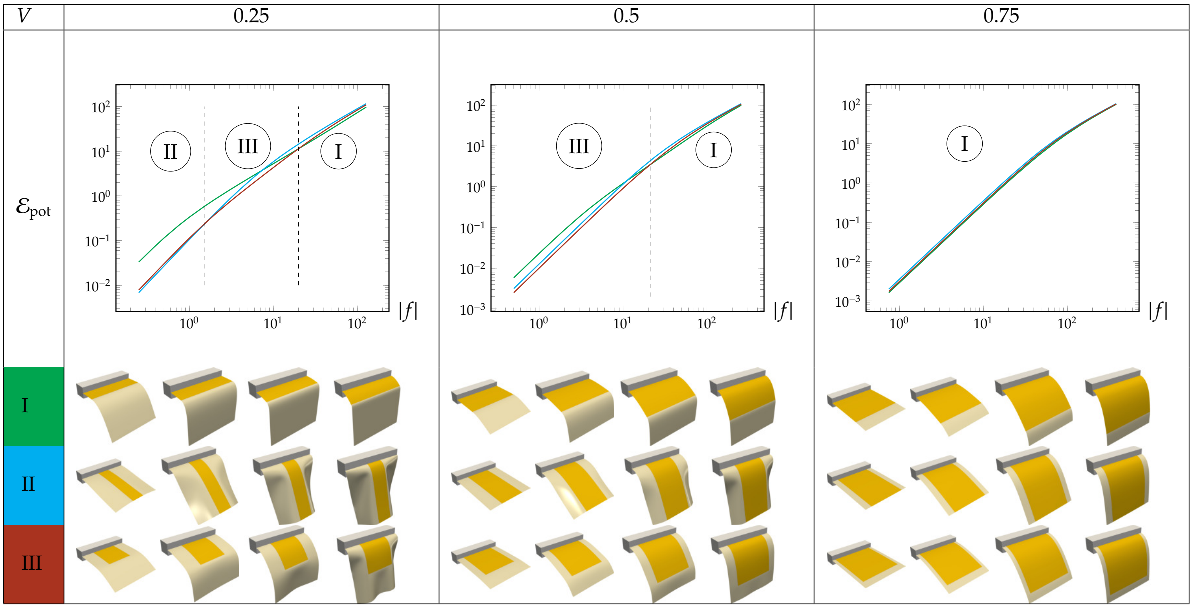


Figure 8.12: Comparison of the potential energy functional in dependence of $|f|$ in a logarithmic scale for three design types and different area fractions $V = 0.25, 0.5, 0.75$ of the hard material. By dotted lines we separate the ranges, where a specific design is optimal w.r.t. the potential energy.

Optimal Designs

Now, we compute the optimal material distribution for the above scenario. The comparison of the different designs as in Figure 8.12 shows that in particular cases depending on the force and the amount of hard material, the optimal solution is different to design (I). Here, we take into account the same area constraints $V = 0.25, 0.5, 0.75$ as above. For all computations, we start with a coarse mesh of $|\mathcal{N}_h| = 289$ nodes and use 8 adaptive refinement steps. In Figure 8.13, we consider large forces with $|f| = 100V$. We observe for a large amount of hard material with $V = 0.75$ that design (I) is optimal. However, for $V = 0.25, 0.5$ we obtain optimal designs with significantly better compliance compared to the above considered designs. Furthermore, in Figure 8.14, we investigate small forces with $|f| = 10V$. Here, for all constraints V, the optimal solutions are different to the above chosen designs, even for an area $V = 0.75$, where design (I) performs better than (II) and (III). However, in all of our computational results, microstructures do not appear.

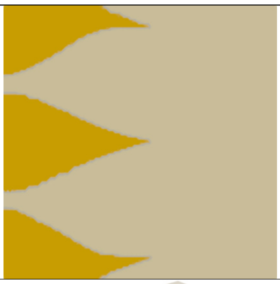
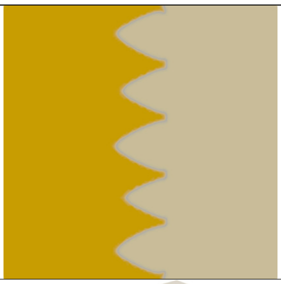
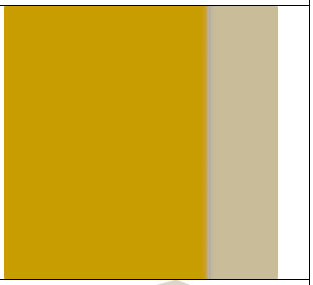
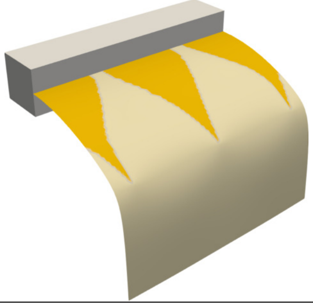
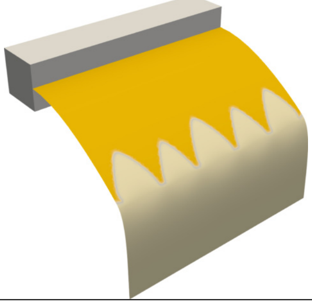
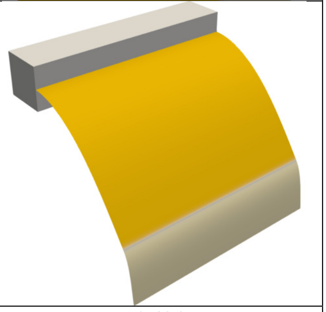
V	0.25	0.5	0.75
$\mathcal{M}_A$			
$\mathcal{M}_B$			
$\lvert\mathcal{N}_h\rvert$	8531	8401	10580
$\mathcal{E}_{\text{pot}}(v_{\text{opt}})$	5.32684	11.6839	21.0868
$\mathcal{E}_{\text{pot}}(v_{\text{I}})$	6.71972	11.7869	21.0853
$\mathcal{E}_{\text{pot}}(v_{\text{II}})$	7.91447	15.7301	23.5953
$\mathcal{E}_{\text{pot}}(v_{\text{III}})$	5.89819	13.6732	22.4527

Figure 8.13: Optimal material distributions for bending isometries of a plate for large forces with $|f| = 100V$.

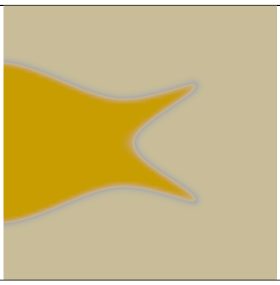
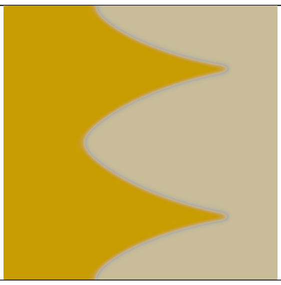
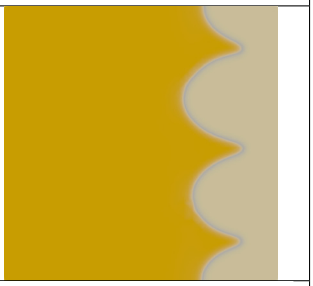
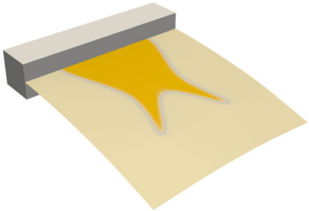
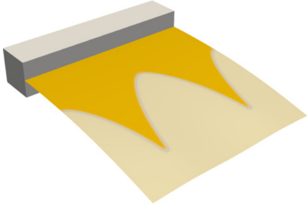
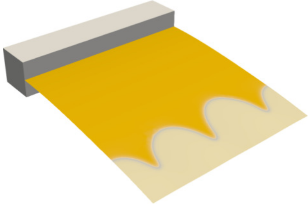
V	0.25	0.5	0.75
$\mathcal{M}_A$			
$\mathcal{M}_B$			
$\lvert\mathcal{N}_h\rvert$	11624	15295	12901
$\mathcal{E}_{\text{pot}}(v_{\text{opt}})$	0.106411	0.184621	0.340829
$\mathcal{E}_{\text{pot}}(v_{\text{I}})$	0.453524	0.422204	0.360064
$\mathcal{E}_{\text{pot}}(v_{\text{II}})$	0.161784	0.303079	0.44151
$\mathcal{E}_{\text{pot}}(v_{\text{III}})$	0.16907	0.239765	0.387821

Figure 8.14: Optimal material distributions for bending isometries of a plate for small forces with $|f| = 10V$.

8.5 Conclusion and Outlook

We have investigated optimal material distributions on thin elastic objects w.r.t. the potential energy. For our numerical discretization, we have made use of the DKT element on parametric surfaces. Depending on the particular force acting on the surface, for mixed membrane-bending models, it has turned out that interesting microstructures appear, where we have observed significant differences between linear and nonlinear elasticity. Furthermore, we have studied the case of pure bending isometries. For a one-dimensional model of elastic beams, our numerical results have confirmed and extended a classification result for the optimal design. In the two-dimensional model for pure bending isometries, it seems that no microstructures appear. Indeed, in all our numerical tests, we have obtained classical designs without microstructures, even for initializations of the phase-field with random values. However, a possible classification result for the optimal design as in the one-dimensional model might require a specific case study, since we have observed different optimal designs depending on the amount of hard material and the force.

Although the DKT element only allows a nonconforming approximation of second derivatives, suitable convergence estimates, *e.g.*, for bending isometries in [Bar13], can be established. In contrast to [Bar13], where a gradient flow with a linearized isometry constraint was proposed, we have implemented a Newton method for a Lagrangian with an exact isometry constraint at nodal positions. Here, we have focused on the material optimization problem. Moreover, we note that our numerical implementation of the DKT element is so far restricted to parametric surfaces, *i.e.*, the midsurfaces of the reference and the deformed shell are obtained as images of vector-valued DKT functions on a common chart domain. An extension to arbitrary shells would be desirable, but this requires an interpretation of the degrees of freedom for deformation gradients. Now, on parametric surfaces having an approximation of the relative shape operator at hand, we could study further mechanical properties of thin elastic objects. In [Bar17], the DKT element has been applied to approximate deformations of plates for a Föppel–von Kármán model, which has been used to verify a break of symmetry on circular cones that has been previously proven in [COT17]. Similar buckling effects on the sphere have been simulated in [VM08, NAL$^+$13] for a different finite element discretization.

Finally, a two-scale optimization of thin elastic objects would be a possible extension of our numerical scheme to explore optimal microstructures. This might have a similar medical application in bone tissue engineering as we have considered in Chapter 7. There, the biologically degradable polymer implants are, *e.g.*, applicable to the tibia bone. However, bone substitutes to fill holes in the skull are comparably thin and have to be curved according to the patient-specific skull.

Bibliography

[AB93] Luigi Ambrosio and Giuseppe Buttazzo. An optimal design problem with perimeter penalization. *Calc. Var. Partial Differential Equations*, 1(1):55–69, 1993.

[ABFJ97] Grégoire Allaire, Eric Bonnetier, Gilles Francfort, and François Jouve. Shape optimization by the homogenization method. *Numer. Math.*, 76(1):27–68, 1997.

[AD14] Grégoire Allaire and Charles Dapogny. A linearized approach to worst-case design in parametric and geometric shape optimization. *Math. Models Methods Appl. Sci.*, 24(11):2199–2257, 2014.

[AFP00] Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[AJM16] G. Allaire, F. Jouve, and G. Michailidis. Thickness control in structural optimization via a level set method. *Struct. Multidiscip. Optim.*, 53(6):1349–1382, 2016.

[AJT04] Grégoire Allaire, François Jouve, and Anca-Maria Toader. Structural optimization using sensitivity analysis and a level-set method. *J. Comput. Phys.*, 194(1):363–393, 2004.

[All02] Grégoire Allaire. Shape optimization by the homogenization method. 2002.

[Alt16] Hans Wilhelm Alt. *Linear functional analysis*. Universitext. Springer-Verlag London, Ltd., London, 2016. An application-oriented introduction, Translated from the German edition by Robert Nürnberg.

[Arn66] V. Arnold. Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. *Ann. Inst. Fourier (Grenoble)*, 16(fasc. 1):319–361, 1966.

[Bab76] Ivo Babuška. Homogenization approach in engineering. In *Computing methods in applied sciences and engineering (Second Internat. Sympos., Versailles, 1975), Part 1*, pages 137–153. Lecture Notes in Econom. and Math. Systems, Vol. 134. 1976.

[Bal77] John M. Ball. Convexity conditions and existence theorems in nonlinear elasticity. *Arch. Rational Mech. Anal.*, 63(4):337–403, 1976/77.

[Bär01] Christian Bär. *Elementare Differentialgeometrie*. de Gruyter Lehrbuch. [de Gruyter Textbook]. Walter de Gruyter & Co., Berlin, 2001.

[Bar13] Sören Bartels. Approximation of large bending isometries with discrete Kirchhoff triangles. *SIAM J. Numer. Anal.*, 51(1):516–525, 2013.

[Bar15] Sören Bartels. *Numerical methods for nonlinear partial differential equations*, volume 47 of *Springer Series in Computational Mathematics*. Springer, Cham, 2015.

[Bar17] Sören Bartels. Numerical solution of a föppl–von kármán model. *SIAM Journal on Numerical Analysis*, 55(3):1505–1524, 2017.

[BB90] Guy Bouchitté and Giuseppe Buttazzo. New lower semicontinuity results for nonconvex functionals defined on measures. *Nonlinear Anal.*, 15(7):679–692, 1990.

[BB92] G. Bouchitté and G. Buttazzo. Integral representation of nonconvex functionals defined on measures. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 9(1):101–117, 1992.

[BB00] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[BBH80] Jean-Louis Batoz, Klaus-Jürgen Bathe, and Lee-Wing Ho. A study of three-node triangular plate bending elements. *International Journal for Numerical Methods in Engineering*, 15(12):1771–1812, 1980.

[BC17] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, second edition, 2017. With a foreword by Hédy Attouch.

[BC18] Kosala Bandara and Fehmi Cirak. Isogeometric shape optimisation of shell structures using multiresolution subdivision surfaces. *Computer-Aided Design*, 95:62–71, 2018.

[BCC08] Adrien Blanchet, Vincent Calvez, and José A. Carrillo. Convergence of the mass-transport steepest descent scheme for the subcritical Patlak-Keller-Segel model. *SIAM J. Numer. Anal.*, 46(2):691–721, 2008.

[BCC$^+$15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.*, 37(2):A1111–A1138, 2015.

[BCO$^+$15] Kosala Bandara, Fehmi Cirak, Günther Of, Olaf Steinbach, and Jan Zapletal. Boundary element based multiresolution shape optimisation in electrostatics. *Journal of computational physics*, 297:584–598, 2015.

[BE88] Dimitri P. Bertsekas and Jonathan Eckstein. Dual coordinate step methods for linear network flow problems. *Math. Programming*, 42(2, (Ser. B)):203–243, 1988.

[Ben89] Martin P Bendsøe. Optimal shape design as a material distribution problem. *Structural optimization*, 1(4):193–202, 1989.

[BER15] Benjamin Berkels, Alexander Effland, and Martin Rumpf. Time discrete geodesic paths in the space of images. *SIAM J. Imaging Sci.*, 8(3):1457–1488, 2015.

[BFO10] Jean-David Benamou, Brittany D. Froese, and Adam M. Oberman. Two numerical methods for the elliptic Monge-Ampère equation. *M2AN Math. Model. Numer. Anal.*, 44(4):737–758, 2010.

[BGHR16] Luise Blank, Harald Garcke, Claudia Hecht, and Christoph Rupprecht. Sharp interface limit for a phase field model in structural optimization. *SIAM Journal on Control and Optimization*, 54(3):1558–1584, 2016.

[BH81] Klaus-Jürgen Bathe and Lee-Wing Ho. A simple and effective element for analysis of general shell structures. *Computers & Structures*, 13(5-6):673–681, 1981.

[BL15] Christoph Brauer and Dirk Lorenz. Cartoon-texture-noise decomposition with transport norms. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 142–153. Springer, 2015.

[Ble14] Kai-Uwe Bletzinger. A consistent frame for sensitivity filtering and the vertex assigned morphing of optimal shape. *Structural and Multidisciplinary Optimization*, 49(6):873–895, 2014.

[BLP78] Alain Bensoussan, Jacques-Louis Lions, and George Papanicolaou. *Asymptotic analysis for periodic structures*, volume 5 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-New York, 1978.

[Bra06] Andrea Braides. A handbook of Γ-convergence. In *Handbook of Differential Equations: stationary partial differential equations*, volume 3, pages 101–213. Elsevier, 2006.

[Bra07] Dietrich Braess. *Finite elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, 2007.

[Bre89] Yann Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *J. Amer. Math. Soc.*, 2(2):225–255, 1989.

[Bre91] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[BS05] Giuseppe Buttazzo and Filippo Santambrogio. A model for the optimal planning of an urban area. *SIAM J. Math. Anal.*, 37(2):514–530, 2005.

[BSW18] David P Bourne, Bernhard Schmitzer, and Benedikt Wirth. Semi-discrete unbalanced optimal transport and quantization. *arXiv preprint arXiv:1808.01962*, 2018.

[BW16] Alessio Brancolini and Benedikt Wirth. Equivalent formulations for the branched transport and urban planning problems. *Journal de Mathématiques Pures et Appliquées*, 106(4):695–724, 2016.

[CGLR17] Sergio Conti, Benedict Geihe, Martin Lenz, and Martin Rumpf. A posteriori modeling error estimates in the optimization of two-scale elastic composite materials. *ESAIM: Mathematical Modelling and Numerical Analysis*, 2017. to appear.

[CGRS14] Sergio Conti, Benedict Geihe, Martin Rumpf, and Rüdiger Schultz. Two-stage stochastic optimization meets two-scale simulation. In Günter Leugering, Peter Benner, Sebastian Engell, Andreas Griewank, Helmut Harbrecht, Michael Hinze, Rolf Rannacher, and Stefan Ulbrich, editors, *Trends in PDE Constrained Optimization*, volume 165 of *International Series of Numerical Mathematics*, pages 193–211. Springer International Publishing, 2014.

[CHP$^+$08] Sergio Conti, Harald Held, Martin Pach, Martin Rumpf, and Rüdiger Schultz. Shape optimization under uncertainty—a stochastic programming perspective. *SIAM J. Optim.*, 19(4):1610–1632, 2008.

[CHP$^+$11] Sergio Conti, Harald Held, Martin Pach, Martin Rumpf, and Rüdiger Schultz. Risk averse shape optimization. *SIAM J. Control Optim.*, 49(3):927–947, 2011.

[Cia88] Philippe G. Ciarlet. *Mathematical elasticity. Vol. I*, volume 20 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1988. Three-dimensional elasticity.

[Cia08] Philippe G. Ciarlet. An introduction to differential geometry in $\mathbb{R}^3$. In *Differential geometry: theory and applications*, volume 9 of *Ser. Contemp. Appl. Math. CAM*, pages 1–93. Higher Ed. Press, Beijing, 2008.

[CM08] Philippe G. Ciarlet and Cristinel Mardare. An introduction to shell theory. In *Differential geometry: theory and applications*, volume 9 of *Ser. Contemp. Appl. Math. CAM*, pages 94–184. Higher Ed. Press, Beijing, 2008.

[CM10] Luis A. Caffarelli and Robert J. McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Ann. of Math. (2)*, 171(2):673–730, 2010.

[COS00] Fehmi Cirak, Michael Ortiz, and Peter Schroder. Subdivision surfaces: a new paradigm for thin-shell finite-element analysis. *International Journal for Numerical Methods in Engineering*, 47(12):2039–2072, 2000.

[COT17] Sergio Conti, Heiner Olbermann, and Ian Tobasco. Symmetry breaking in indented elastic cones. *Mathematical Models and Methods in Applied Sciences*, 27(02):291–321, 2017.

[CP11] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.

[CPSV15] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: geometry and kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

[CPSV18] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: dynamic and Kantorovich formulations. *J. Funct. Anal.*, 274(11):3090–3123, 2018.

[CRST18] Sergio Conti, Martin Rumpf, Rüdiger Schultz, and Sascha Tölkes. Stochastic dominance constraints in elastic shape optimization. *SIAM J. Control Optim.*, 56(4):3021–3034, 2018.

[dC92] Manfredo Perdigão do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser, 1992. Translated from the second Portuguese edition by Francis Flaherty.

[DM93] Gianni Dal Maso. *An introduction to Γ-convergence*, volume 8 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser, 1993.

[DMM86] G Dhatt, L Marcotte, and Y Matte. A new triangular discrete kirchhoff plate/shell element. *International journal for numerical methods in engineering*, 23(3):453–470, 1986.

[DNS09] Jean Dolbeault, Bruno Nazaret, and Giuseppe Savaré. A new class of transport distances between measures. *Calc. Var. Partial Differential Equations*, 34(2):193–231, 2009.

[DPRS19] Patrick Dondl, Patrina S. P. Poh, Martin Rumpf, and Stefan Simon. Simultaneous elastic shape optimization for a domain splitting in bone tissue engineering. *Proc. R. Soc. A*, 475, 2019.

[EB92] Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Programming*, 55(3, Ser. A):293–318, 1992.

[EG15] Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, revised edition, 2015.

[EM12] Matthias Erbar and Jan Maas. Ricci curvature of finite Markov chains via convexity of the entropy. *Arch. Ration. Mech. Anal.*, 206(3):997–1038, 2012.

[EM14] Matthias Erbar and Jan Maas. Gradient flow structures for discrete porous medium equations. *Discrete Contin. Dyn. Syst.*, 34(4):1355–1374, 2014.

[EMW19] Matthias Erbar, Jan Maas, and Melchior Wirth. On the geometry of geodesics in discrete optimal transport. *Calc. Var. Partial Differential Equations*, 58(1):58:19, 2019.

[Erb16] Matthias Erbar. A gradient flow approach to the boltzmann equation. *arXiv preprint arXiv:1603.00540*, 2016.

[ERSS17] Matthias Erbar, Martin Rumpf, Bernhard Schmitzer, and Stefan Simon. Computation of optimal transport on discrete metric measure spaces. *arXiv preprint arXiv:1707.06859*, 2017. to appear in Numer. Math.

[Ess09] Ernie Esser. Applications of lagrangian-based alternating direction methods and connections to split bregman. *CAM report*, 9:31, 2009.

[ET99] Ivar Ekeland and Roger Témam. *Convex analysis and variational problems*, volume 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english edition, 1999. Translated from the French.

[Fig10] Alessio Figalli. The optimal partial transport problem. *Arch. Ration. Mech. Anal.*, 195(2):533–560, 2010.

[FJM02] Gero Friesecke, Richard D. James, and Stefan Müller. A theorem on geometric rigidity and the derivation of nonlinear plate theory from three-dimensional elasticity. *Comm. Pure Appl. Math.*, 55(11):1461–1506, 2002.

[FJMM03] Gero Friesecke, Richard D. James, Maria Giovanna Mora, and Stefan Müller. Derivation of nonlinear bending theory for shells from three-dimensional nonlinear elasticity by Gamma-convergence. *C. R. Math. Acad. Sci. Paris*, 336(8):697–702, 2003.

[FL07] Irene Fonseca and Giovanni Leoni. *Modern methods in the calculus of variations: L^p spaces*. Springer Monographs in Mathematics. Springer, New York, 2007.

[FMP10] A. Figalli, F. Maggi, and A. Pratelli. A mass transportation approach to quantitative isoperimetric inequalities. *Invent. Math.*, 182(1):167–211, 2010.

[Gal16] Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, Princeton, NJ, 2016.

[GHDS03] Eitan Grinspun, Anil N Hirani, Mathieu Desbrun, and Peter Schröder. Discrete shells. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 62–67. Eurographics Association, 2003.

[GHHK15] Harald Garcke, Claudia Hecht, Michael Hinze, and Christian Kahle. Numerical approximation of phase field based shape and topology optimization for fluids. *SIAM J. Sci. Comput.*, 37(4):A1846–A1871, 2015.

[GKM18] Peter Gladbach, Eva Kopfer, and Jan Maas. Scaling limits of discrete optimal transport. *arXiv preprint arXiv:1809.01092*, 2018.

[GM13] Nicola Gigli and Jan Maas. Gromov-Hausdorff convergence of discrete transportation metrics. *SIAM J. Math. Anal.*, 45(2):879–899, 2013.

[GR16] Benedict Geihe and Martin Rumpf. A posteriori error estimates for sequential laminates in shape optimization. *Discrete and Continuous Dynamical Systems - Series S*, 9(5):1377–1392, 2016.

[HCB05] Thomas JR Hughes, John A Cottrell, and Yuri Bazilevs. Isogeometric analysis: Cad, finite elements, nurbs, exact geometry and mesh refinement. *Computer methods in applied mechanics and engineering*, 194(39-41):4135–4195, 2005.

[Hee16] Behrend Heeren. *Numerical Methods in Shape Spaces and Optimal Branching Patterns*. PhD thesis, University of Bonn, 2016.

[HMP15] Romain Hug, Emmanuel Maitre, and Nicolas Papadakis. Multi-physics optimal transportation and image interpolation. *ESAIM Math. Model. Numer. Anal.*, 49(6):1671–1692, 2015.

[HN59] Philip Hartman and Louis Nirenberg. On spherical image maps whose Jacobians do not change sign. *Amer. J. Math.*, 81:901–920, 1959.

[Hor11] Peter Hornung. Approximation of flat $W^{2,2}$ isometric immersions by smooth ones. *Arch. Ration. Mech. Anal.*, 199(3):1015–1067, 2011.

[HPUU08] Michael Hinze, René Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE constraints*, volume 23. Springer Science & Business Media, 2008.

[HRS$^+$14] Behrend Heeren, Martin Rumpf, Peter Schröder, Max Wardetzky, and Benedikt Wirth. Exploring the geometry of the space of shells. *Comput. Graph. Forum*, 33(5):247–256, 2014.

[HRS19] Peter Hornung, Martin Rumpf, and Stefan Simon. Material optimization for nonlinearly elastic planar beams. *ESAIM Control Optim. Calc. Var.*, 25:Art. 11, 19, 2019.

[HRW17] Behrend Heeren, Martin Rumpf, and Benedikt Wirth. Variational time discretization of Riemannian splines. *IMA J. Numer. Anal.*, 2017. accepted.

[HRWW12] Behrend Heeren, Martin Rumpf, Max Wardetzky, and Benedikt Wirth. Time-discrete geodesics in the space of shells. *Comput. Graph. Forum*, 31(5):1755–1764, 2012.

[HX10] X. Huang and Y. M. Xie. Evolutionary topology optimization of continuum structures with an additional displacement constraint. *Struct. Multidiscip. Optim.*, 40(1-6):409–416, 2010.

[HZ14] Xiaocong Han and David W Zingg. An adaptive geometry parametrization for aerodynamic shape optimization. *Optimization and Engineering*, 15(1):69–91, 2014.

[HZTA04] Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60(3):225–240, 2004.

[IBRS13] Jose A Iglesias, Benjamin Berkels, Martin Rumpf, and Otmar Scherzer. A thin shell approach to the registration of implicit surfaces. 2013.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

[JKZ98] Florian Jarre, Michal Kočvara, and Jochem Zowe. Optimal truss design by interior-point methods. *SIAM J. Optim.*, 8(4):1084–1107, 1998.

[Kan42] L Kantorovich. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, pages 227–229, 1942.

[Kan48] Leonid V Kantorovich. On a problem of monge. In *CR (Doklady) Acad. Sci. URSS (NS)*, volume 3, pages 225–226, 1948.

[KHM$^+$11] Sebastian C Kapfer, Stephen T Hyde, Klaus Mecke, Christoph H Arns, and Gerd E Schröder-Turk. Minimal surface scaffold designs for tissue engineering. *Biomaterials*, 32(29):6875–6882, 2011.

[Kno57] Herbert Knothe. Contributions to the theory of convex bodies. *Michigan Math. J.*, 4:39–52, 1957.

[Koi66] W. T. Koiter. On the nonlinear theory of thin elastic shells. I, II, III. *Nederl. Akad. Wetensch. Proc. Ser. B*, 69:1–17, 18–32, 33–54, 1966.

[KPRA18] Laurence Kedward, Alexandre D Payot, T Rendall, and Christian B Allen. Efficient multi-resolution approaches for exploration of external aerodynamic shape and topology. In *2018 Applied Aerodynamics Conference*, page 3952, 2018.

[KSKW15] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[Kui55] Nicolaas H. Kuiper. On C^1-isometric imbeddings. I, II. *Nederl. Akad. Wetensch. Proc. Ser. A.* **58** = *Indag. Math.*, 17:545–556, 683–689, 1955.

[LDR95] Hervé Le Dret and Annie Raoult. The nonlinear membrane model as variational limit of nonlinear three-dimensional elasticity. *J. Math. Pures Appl. (9)*, 74(6):549–578, 1995.

[LDR96] H. Le Dret and A. Raoult. The membrane shell model in nonlinear elasticity: a variational asymptotic derivation. *J. Nonlinear Sci.*, 6(1):59–84, 1996.

[Lév15] Bruno Lévy. A numerical algorithm for l2 semi-discrete optimal transport in 3d. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1693–1715, 2015.

[LMS15] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger-kantorovich distance between positive measures. *arXiv preprint arXiv:1508.07941*, 2015.

[LR05] Grégoire Loeper and Francesca Rapetti. Numerical solution of the monge–ampère equation by a newton's algorithm. *Comptes Rendus Mathematique*, 340(4):319–324, 2005.

[LV09] John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Ann. of Math. (2)*, 169(3):903–991, 2009.

[Maa11] Jan Maas. Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.*, 261(8):2250–2292, 2011.

[Mér11] Quentin Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011.

[MM77] Luciano Modica and Stefano Mortola. Un esempio di Γ^--convergenza. *Boll. Un. Mat. Ital. B (5)*, 14(1):285–299, 1977.

[MM03] Maria Giovanna Mora and Stefan Müller. Derivation of the nonlinear bending-torsion theory for inextensible rods by Γ-convergence. *Calc. Var. Partial Differential Equations*, 18(3):287–305, 2003.

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[MRCS10] Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow type. *Math. Models Methods Appl. Sci.*, 20(10):1787–1821, 2010.

[MRS17] Jan Maas, Martin Rumpf, and Stefan Simon. Transport based image morphing with intensity modulation. In *Scale Space and Variational Methods in Computer Vision*, pages 563–577, Cham, 2017. Springer International Publishing.

[MRSS15] Jan Maas, Martin Rumpf, Carola Schönlieb, and Stefan Simon. A generalized model for optimal transport of images including dissipation and density modulation. *ESAIM Math. Model. Numer. Anal.*, 49(6):1745–1769, 2015.

[MT97] François Murat and Luc Tartar. H-convergence. In *Topics in the mathematical modelling of composite materials*, volume 31 of *Progr. Nonlinear Differential Equations Appl.*, pages 21–43. Birkhäuser Boston, Boston, MA, 1997.

[Mur78] François Murat. Compacité par compensation. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 5(3):489–507, 1978.

[NAL$^+$13] Alice Nasto, Amin Ajdari, Arnaud Lazarus, Ashkan Vaziri, and Pedro M Reis. Localization of deformation in thin shells under indentation. *Soft Matter*, 9(29):6796–6803, 2013.

[Nas54] John Nash. C^1 isometric imbeddings. *Ann. of Math. (2)*, 60:383–396, 1954.

[Nit81] J. A. Nitsche. On Korn's second inequality. *RAIRO Anal. Numér.*, 15(3):237–248, 1981.

[Ott01] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.

[PC17] Gabriel Peyré and Marco Cuturi. Computational optimal transport. Technical report, 2017.

[PCS16] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

[PCW$^+$16] Patrina SP Poh, Mohit P Chhaya, Felix M Wunner, Elena M De-Juan-Pardo, Arndt F Schilling, Jan-Thorsten Schantz, Martijn van Griensven, and Dietmar W Hutmacher. Polylactides in additive biomanufacturing. *Advanced Drug Delivery Reviews*, 107:228–246, 2016.

[Ped00] Niels L Pedersen. Maximization of eigenvalues using topology optimization. *Structural and multidisciplinary optimization*, 20(1):2–11, 2000.

[Pey15] Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM J. Imaging Sci.*, 8(4):2323–2351, 2015.

[PFR12] Gabriel Peyré, Jalal Fadili, and Julien Rabin. Wasserstein active contours. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 2541–2544. IEEE, 2012.

[PPC11] Nicolas Papadakis, Edoardo Provenzi, and Vicent Caselles. A variational model for histogram transfer of color images. *IEEE Transactions on Image Processing*, 20(6):1682–1695, 2011.

[PPO14] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM J. Imaging Sci.*, 7(1):212–238, 2014.

[PR14] Benedetto Piccoli and Francesco Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Arch. Ration. Mech. Anal.*, 211(1):335–358, 2014.

[PR16] Benedetto Piccoli and Francesco Rossi. On properties of the generalized Wasserstein distance. *Arch. Ration. Mech. Anal.*, 222(3):1339–1365, 2016.

[PRW12] Patrick Penzler, Martin Rumpf, and Benedikt Wirth. A phase-field model for compliance shape optimization in nonlinear elasticity. *ESAIM Control Optim. Calc. Var.*, 18(1):229–258, 2012.

[PVB$^+$18] Patrina S. P. Poh, Dvina Valainis, Kaushik Bhattacharya, Martijn van Griensven, and Patrick Dondl. Optimizing bone scaffold porosity distributions, 2018.

[RHSH18] Michael Rabinovich, Tim Hoffmann, and Olga Sorkine-Hornung. Discrete geodesic nets for modeling developable surfaces. *ACM Transactions on Graphics (ToG)*, 37(2):16, 2018.

[RW13] Martin Rumpf and Benedikt Wirth. Discrete geodesic calculus in shape space and applications in the space of viscous fluidic objects. *SIAM J. Imaging Sci.*, 6(4):2581–2602, 2013.

[RW15] Martin Rumpf and Benedikt Wirth. Variational time discretization of geodesic calculus. *IMA J. Numer. Anal.*, 35(3):1011–1046, 2015.

[San15] Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[Sas19] Josua Sassen. Discrete gauss-codazzi equations for efficient triangle mesh processing. Master's thesis, Universität Bonn, 2019.

[Sch16a] Bernhard Schmitzer. A sparse multiscale algorithm for dense optimal transport. *J. Math. Imaging Vision*, 56(2):238–259, 2016.

[Sch16b] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *arXiv preprint arXiv:1610.06519*, 2016.

[Sch18] Almut Scheerer. Numerische approximation des optimalen transports auf graphen, 2018.

[Sch19] Mareike Schmerling. Optimizing printable 3d bone implant microstructure. Master's thesis, Universität Bonn, 2019.

[SHB$^+$18] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[Sil07] Luis Silvestre. A characterization of optimal two-phase multifunctional composite designs. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 463(2086):2543–2556, 2007.

[SL05] Jan Stegmann and Erik Lund. Discrete material optimization of general composite shell structures. *International Journal for Numerical Methods in Engineering*, 62(14):2009–2027, 2005.

[SRGB16] Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Continuous-flow graph transportation distances. *arXiv preprint arXiv:1603.06927*, 2016.

[SS13] Volker Schulz and Claudia Schillings. Optimal aerodynamic design under uncertainty. In *Management and Minimisation of Uncertainties and Errors in Numerical Aerodynamics*, pages 297–338. Springer, 2013.

[SSW15] Volker Schulz, Martin Siebenborn, and Kathrin Welker. Towards a lagrange–newton approach for pde constrained shape optimization. In *New Trends in Shape Optimization*, pages 229–249. Springer, 2015.

[Stu06a] Karl-Theodor Sturm. On the geometry of metric measure spaces. I. *Acta Math.*, 196(1):65–131, 2006.

[Stu06b] Karl-Theodor Sturm. On the geometry of metric measure spaces. II. *Acta Math.*, 196(1):133–177, 2006.

[Tar79] L. Tartar. Compensated compactness and applications to partial differential equations. In *Nonlinear analysis and mechanics: Heriot-Watt Symposium, Vol. IV*, volume 39 of *Res. Notes in Math.*, pages 136–212. Pitman, Boston, Mass.-London, 1979.

[TD04] S. Torquato and A. Donev. Minimal surfaces and multifunctionality. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 460(2047):1849–1856, 2004.

[THD02] S Torquato, S Hyun, and A Donev. Multifunctional composites: optimizing microstructures for simultaneous transport of heat and electricity. *Physical review letters*, 89(26):266601, 2002.

[VHWP12] Etienne Vouga, Mathias Höbinger, Johannes Wallner, and Helmut Pottmann. Design of self-supporting surfaces. *ACM Transactions on Graphics (TOG)*, 31(4):87, 2012.

[Vil03] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[Vil09] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[VM08] Ashkan Vaziri and L Mahadevan. Localized and extended deformations of elastic shells. *Proceedings of the National Academy of Sciences*, 105(23):7913–7918, 2008.

[War08] Max Wardetzky. Convergence of the cotangent formula: An overview. In *Discrete differential geometry*, pages 275–286. Springer, 2008.

[WB06] Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.

[WBH$^+$07] Max Wardetzky, Miklós Bergou, David Harmon, Denis Zorin, and Eitan Grinspun. Discrete quadratic curvature energies. *Computer Aided Geometric Design*, 24(8-9):499–518, 2007.

[WDAH10] Tim Winkler, Jens Drieseberg, Marc Alexa, and Kai Hormann. Multi-scale geometry interpolation. In *Computer graphics forum*, volume 29, pages 309–318. Wiley Online Library, 2010.

[XS93] Yi M Xie and Grant P Steven. A simple evolutionary procedure for structural optimization. *Computers & structures*, 49(5):885–896, 1993.

[ZW07] Shiwei Zhou and Michael Yu Wang. Multimaterial structural topology optimization with a generalized Cahn-Hilliard model of multiphase transition. *Struct. Multidiscip. Optim.*, 33(2):89–111, 2007.