Fach: Mathematik

Numerical methods for transportation networks

Inaugural Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften
– Dr. rer. nat. –
im Fachbereich Mathematik und Informatik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

eingereicht von

Carolin Maria Dirks geb. Rossmanith
aus Steinfurt

– 2019 –

Dekan:	Prof. Dr. Xiaoyi Jiang
Erster Gutachter:	Prof. Dr. Benedikt Wirth (Universität Münster)
Zweiter Gutachter:	Prof. Dr. Édouard Oudet (Université Grenoble Alpes)
Tag der mündlichen Prüfung:	24.10.2019
Tag der Promotion:	24.10.2019

Abstract

The optimal transportation network problem consists in constructing a pathway interconnecting two measures μ_+, μ_- to the lowest possible costs. The structure of the optimal network depends on the definition of the cost functional. Within this work, we want to consider functionals which promote branching structures in the desired solution, where the grade of ramification is controlled by a branching parameter $\varepsilon > 0$ and whose most famous representatives are the branched transport and the urban planning problem. Typically, the corresponding energy landscape is highly non-convex and as a consequence, the identification and construction of a global minimizer is a challenging task.

Despite these difficulties, the aim for numerical optimization methods for branched transportation problems is beyond all question, since there exists a variety of interesting applications such as the development of a blood vessel system or the construction of a public transportation network. In recent years, researchers came up with several ideas successfully tackling the problem of numerically finding an optimal transportation path, however, most of these methods suffer from certain drawbacks of practical or theoretical nature.

In this work, we present two novel numerical treatments based on different formulations of the branched transport and urban planning problem. The first one is based on a convex relaxation achieved via an image-based Mumford–Shah-type reformulation and subsequent functional lifting of the energy. The resulting convex optimization problem can be solved efficiently via an adaptive finite element approach, where a specific class of finite elements is designed to efficiently handle the particular problem structure. The second approach exploits the ideas of Ambrosio and Tortorelli to formulate a phase field approximation of the generalized urban planning model featuring multiple phase fields and a possible diffuse component allowing additional transport outside of the desired network.

This thesis deals with the numerical treatment of the previously mentioned relaxed energy functionals, discusses the numerical challenges, designs an appropriate discretization framework and presents simulation results.

Keywords: Optimal transport, branched transport, urban planning, Mumford–Shah, functional lifting, adaptive finite elements, phase field, Ambrosio–Tortorelli

Acknowledgements

Before we start to delve into the mathematical results, I would like to express my sincere gratitude to everyone who contributed in any conceivable way to the development of this thesis. Although the list of supporting colleagues and friends is too long to address everyone personally, there are a few people I want to explicitly acknowledge at this point for their outstanding contribution, namely

- Benedikt Wirth for introducing me to this fascinating topic and for being an excellent supervisor, constantly providing me with sophisticated advice and assistance, patiently answering every question and repeatedly taking a lot of his time for endless discussions,

- Édouard Oudet for agreeing to become my second reviewer, for kindly hosting me at his institute in Grenoble during my PhD and for lots of interesting discussions, new ideas and helpful hints,

- Ulrich Böttcher for being the best office co-worker during the last years one can imagine, for being a great colleage and friend at the same time and for all his outstanding advice in personal, mathematical and computer-related issues,

- my work group and all former members for lots of both funny and useful discussions during the weekly group meeting,

- my proof-readers Liesel Sommer, Annika Bach, Bernhard Schmitzer and Hendrik Dirks, for pointing out lots of typos, errors and inconsistencies,

- Mira Schedensack for her expertise concerning adaptive finite elements,

- all my colleages and good friends at the institute, in particular Eva-Maria Brinkmann, Julian Rasch, Janic Föcke, Fjedor Gaede, Ina Humpert, Meike Kinzel, Ramona Sasse, Juliane Braunsmann, Bernhard Schmitzer, Frank Wübbeling, Liesel Sommer, Lena Frerking and Stefan Wierling, for lots of coffee, conferences and for creating the most enjoyable working atmosphere,

- the secretary, Carolin Gietz, Claudia Giesbert and Silvia Wernke, for severely lightening everyone's administrative workload,

- my friends outside of the institute for supporting me with a lot of patience, understanding and lots of wonderful and distracting activities,

- my parents for their constant, sympathetical and never pressurizing support, for helping me up again when I struggled and fell and for always having an open ear for any kind of problem,

- my husband Hendrik Dirks, without whom the path up to this point would have been disproportionately harder, for always being there and solidly believing in me, regardless of my doubts and fears.

Contents

List of Figures

1 Introduction

The concept of optimal transport dates back to the very beginnings of the human civilization. In every phase of historical development, people were confronted with various kinds of transportation issues, whether in travelling, trade or construction, just to name some examples. Here, the term *transport* can be related to different questions, such as an optimal assignment of measurable objects or the design of an optimal travelling path. Moreover, the term *optimal* refers to a specific definition of the transport costs, which is usually related to the amount of transported mass as well as the length of the transport path.

In this thesis, we aim at investigating a specific class of transportation problems, where the object of interest is a transportation network interconnecting two prescribed measures under the assumption that mass is preferably transported in bulks instead of each particle travelling individually, causing the occurrence of a branching structure. The cost function of this network penalizes the length of each network segment as well as the amount of mass flowing along this segment. There exists a variety of practical examples where branching networks occur. Plenty of natural transportation paths such as the blood vessel distribution or the water supply system in plants admit a ramified structure. A public transportation system is typically designed to interconnect most city districts while being as short as possible to reduce the maintenance costs.

In mathematical terms, one way to describe such a network in a discrete setting is via a weighted directed graph G consisting of a set of vertices $V(G)$ and edges $E(G)$. Denoting by $l(e)$ the length of an edge $e \in E(G)$ and by $w(e)$ the amount of mass flowing through e, the cost of the graph is then defined as

$$\mathcal{E}(G) = \sum_{e \in E(G)} l(e)c(w(e)),$$

which is to be minimized over graphs which connect a given initial and final measure. The function $c : [0, \infty) \to [0, \infty)$ is continuous, non-decreasing, concave and satisfies $c(0) = 0$. The essential property demanded in this work is the concavity, which is the very factor enforcing a branching structure in the desired network. For a mass $w \in \mathbb{R}$, the function c typically satisfies $c(2w) < 2c(w)$, which essentially means that transporting twice as much mass along one network edge is cheaper than transporting two times the mass w along two different edges.

Since the optimal network problem has a quite practical background, the aim for suitable numerical optimization approaches stands to reason. Unfortunately, the problem typically does not clearly suggest a straightforward discretization, such as a restriction to any *simple* finite-dimensional function space describing the network. In addition, the energy functional turns out to be highly non-convex, possibly comprising several local minima which cause any numerical treatment to be a challenging task. This makes it commonly impossible to derive the truly optimal network, but limits most simulations to the construction of an *almost* optimal path.

Despite the mentioned difficulties, there exist several numerical optimization approaches in the literature, surmounting the obstacles in quite different ways. While some of them restrict themselves to a kind of manual construction of an almost optimal transportation path in a discrete setting, others aim at relaxing the problem by representing the network by a smooth function in the manner of the Ambrosio–Tortorelli [4] or Modica–Mortola [47] approximation of the Mumford–Shah functional [50]. Although all approaches admit interesting and well applicable features and lead to nice approximation results, the research on numerical methods for transportation network is still far from being complete. On the one hand, many numerical solutions cannot be rigorously proven to represent a global optimum due to the lack of convexity. On the other hand, some variants of the problem above, such as the urban planning problem, where particles are allowed to travel outside of the network as well, have never been treated numerically to the best of our knowledge.

This work aims at providing some solutions to the problems mentioned in the last paragraph and is focussing on the production of satisfactory numerical simulation results. In this sense, the main contributions of this thesis are

- two different numerical discretization approaches for a convex image-based reformulation of the branched transport and urban planning energy introduced by [19], where the first one is based on a simple finite difference scheme and the second one consists of a more efficient adaptive finite element implementation specifically designed for functional lifting problems including non-local constraints,
- a novel numerical optimization strategy for a phase field approximation of the generalized urban planning energy functional introduced by [33], including a diffuse component corresponding to transport outside of the network.

The rest of this work is organized as follows. In Chapter 2, we start with some basic notation and a selection of relevant definitions and mathematical concepts. This includes a short review of functions of bounded variations, the main idea of Γ-convergence and a brief review of the Mumford–Shah image segmentation problem, which is related in slightly different ways to both main chapters of this thesis. Furthermore, we present a construction of a suitable finite element space for three-dimensional imaging problems arising from a functional lifting approach of the latter. In Chapter 3, we provide an overview of the relevant models and concepts of optimal transport and transport network problems. We also review some of the existing numerical methods introduced in the literature. The following two chapters contain the main contributions as stated above. Chapter 4 starts with a description of an image-based reformulation of the branched transport and urban planning energy and a convex relaxation of the latter. After an investigation of the differences between the original energy and its relaxed counterpart, we describe in detail two different optimization approaches, solve them with a suitable algorithmic framework and present some simulation results. For the second approach based on adaptive finite elements, we discuss the challenges arising from the involved non-local constraint set, perform some runtime efficiency tests to prove the benefit of adaptivity and compare different refinement strategies. Thereafter, Chapter 5 addresses a phase field approximation of the generalized urban planning energy. Starting with a short model description, we cite some analytical results such as existence of a solution and Γ-convergence of the relaxation, followed by a presentation of the numerical optimization strategy and several simulation results. We complete this work by a short summary of the main results and an outlook to possible future projects in Chapter 6.

2 Mathematical preliminaries

Before we present the results of this work, we want to introduce some basic notation and review some mathematical concepts the following chapters are based on. Starting with a review of the space of functions of bounded variation, we will establish the notion of Γ-convergence and equi-coercivity. Furthermore, we will explain and investigate the Mumford–Shah image-segmentation functional as a representative of an interesting class of problems which are going to play a major role in this thesis. Finally, we review the concepts of the finite element method and present a novel class of custom-designed finite elements for a special type of discretization problems.

2.1. Basic notation

In the following, if not specified otherwise, let $n, N \in \mathbb{N}$ be positive integers, $N \geq 1$ and $\Omega \subset \mathbb{R}^n$ an open bounded subset. Throughout this thesis, we make use of the following standard notation.

- **Euclidean norm and scalar product.** For $x, y \in \mathbb{R}^n$, we denote by $|x|$ the standard Euclidean norm and by $\langle x, y\rangle$ the Euclidean scalar product.
- **Scalar product on Hilbert space.** For a Hilbert space X, we denote for $x, y \in X$ the scalar product of x and y by $\langle x, y\rangle_X$.
- **Borel subsets.** Let $\mathcal{B}(\Omega)$ be the family of all Borel subsets of Ω.
- **Finite Radon measures.** We denote by $\mathcal{M}(\Omega, \mathbb{R}^N)$ the space of finite $\mathbb{R}^N$-valued Radon measures on Ω. For $N = 1$, we write $\mathcal{M}(\Omega)$ and define $\mathcal{M}_+(\Omega)$ as the set of non-negative finite Radon measures on Ω. For a measure $\mu \in \mathcal{M}(\Omega, \mathbb{R}^N)$, the corresponding total-variation norm is denoted by $|\mu|$.

- **Lebesgue, Hausdorff and Dirac measure.** By $\mathcal{L}^n$ we denote the Lebesgue measure in $\mathbb{R}^n$ and by $\mathcal{H}^k$ the k-dimensional Hausdorff measure for $k \in \mathbb{N}$. Additionally, we define
$$\delta_x(B) := \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{otherwise} \end{cases}$$
for all $B \in \mathcal{B}(\Omega)$ as the Dirac measure for a point $x \in \Omega$.

- **Discrete measure.** A measure μ is called discrete, if μ is a (possibly infinite) sum of Dirac measures, i.e. there exists a sequence of points $(x_i) \subset \mathbb{R}^n$ and weights $(a_i) \subset \mathbb{R}$ such that $\mu = \sum_i a_i \delta_{x_i}$.

- **Support of a measure.** The support of a measure $\mu \in \mathcal{M}(\Omega)$ is defined as $\operatorname{spt} \mu := \{x \in \Omega \ : \ \mu(B) > 0$ for every open neighbourhood $B \in \mathcal{B}(\Omega)$ of $x\}$.

- **Spaces of continuously differentiable functions.** We set $C(\Omega, \mathbb{R}^N)$ as the space of $\mathbb{R}^N$-valued continuous functions and $C^k(\Omega, \mathbb{R}^N)$ as the space of $\mathbb{R}^N$-valued continuous functions which are k-times continuously differentiable for $k \in \mathbb{N}$. For $N = 1$, we write $C(\Omega)$ and $C^k(\Omega)$. By $C_0^k(\Omega)$, we denote all functions in $C^k(\Omega)$ with compact support in Ω.

- **L^p spaces.** By $L^p(\Omega, \mathbb{R}^N)$, we denote the space of $\mathbb{R}^N$-valued p-integrable functions with respect to the Lebesgue measure for $p \in \mathbb{N}$. For $N = 1$, we write $L^p(\Omega)$. $L^p(\Omega)$ is a Banach space and we denote the corresponding L^p-norm by $\|\cdot\|_{L^p}$. $L^2(\Omega)$ is a Hilbert space with the scalar product denoted by $\langle \cdot, \cdot \rangle_{L^2}$. In the special case of $p = \infty$, we define $\|f\|_{L^\infty} := \operatorname*{ess\,sup}_{x \in \Omega} |f(x)| < \infty$ for $f \in L^\infty(\Omega, \mathbb{R}^N)$. By $L^p_{loc}(\Omega, \mathbb{R}^N)$ we denote the space of Lebesgue-measurable functions f such that for every compact $V \subset \Omega$, $f \in L^p(V, \mathbb{R}^N)$.

- **Sobolev spaces.** We define $W^{k,p}(\Omega)$ as the space of functions in $L^p(\Omega)$ with p-integrable weak derivative up to order k for $k, p \in \mathbb{N}$. For $p = 2$, $W^{k,2}(\Omega)$ is a Hilbert space with the scalar product denoted by $\langle \cdot, \cdot \rangle_{W^{k,2}}$. By $W^{k,p}_{loc}(\Omega)$ we denote the space of functions $f \in L^p(\Omega)$ such that for every compact $V \subset \Omega$, $f \in W^{k,p}(V)$. By $W_0^{k,p}(\Omega)$, we denote all functions in $W^{k,p}(\Omega)$ with compact support in Ω.

- **Lipschitz-continuous functions.** We define the space of Lipschitz-continuous functions on Ω as
$$C^{0,1}(\Omega) := \{u \in C(\Omega) \ : \ \exists\, L \geq 0 \text{ s.t. } |u(x_1) - u(x_2)| \leq L|x_1 - x_2| \ \forall\, x_1, x_2 \in \Omega\}.$$

- **Unit ball.** We define $B_r(x_0) := \{x \in \mathbb{R}^n \ : \ |x - x_0| < r\}$ as the open unit ball with radius r and midpoint $x_0 \in \mathbb{R}^n$.

- **Unit sphere.** We define $S^{n-1} := \{x \in \mathbb{R}^n \ : \ |x| = 1\}$ as the unit sphere in $\mathbb{R}^n$.

- **Characteristic/indicator function.** For a set $A \subset \Omega$, we define

$$\chi_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise,} \end{cases} \qquad \iota_A(x) := \begin{cases} 0 & \text{if } x \in A, \\ \infty & \text{otherwise.} \end{cases}$$

  χ_A is called the characteristic function of the set A, ι_A is called indicator function of A. Note that if A is a convex set, ι_A is a convex function.

- **Restriction of a measure.** For a measure space $(X, \mathcal{A}, \mu)$ and some $Y \subset X$ with $Y \in \mathcal{A}$, the restriction of the measure μ onto Y is a measure defined as $\mu \llcorner Y(A) := \mu(A \cap Y)$ for every $A \in \mathcal{A}$.

- **Pushforward measure.** For a measure space $(X, \mathcal{A}, \mu)$, a measurable space $(Y, \mathcal{N})$ and a mapping $T : X \to Y$, the pushforward measure of μ is defined as $T\#\mu(B) := \mu(T^{-1}(B))$ for all $B \in \mathcal{N}$.

- **Weak-* convergence.** The weak-* convergence on the space $\mathcal{M}(\Omega, \mathbb{R}^N)$ is denoted by $\rightharpoonup^*$.

- **Subdifferential.** For a convex function $f : \Omega \to \mathbb{R}$, the subdifferential of f in a point $x_0 \in \Omega$ is defined as the set $\partial f(x_0) := \{g \in \mathbb{R}^n \ : \ f(x) \geq f(x_0) + \langle g, x - x_0 \rangle\}$.

- **Orthogonal projection.** For a convex set $C \subset \mathbb{R}^n$ and $x \in \mathbb{R}^n$ we define the orthogonal projection of x onto C as $\mathcal{P}_C(x) := \operatorname*{argmin}_{y \in C} |x - y|$.

- **Volume of a set.** For a set $A \in \mathcal{B}(\Omega)$, we denote the n-dimensional volume of A by $|A| := \mathcal{L}^n(A)$.

- **Convex hull.** For a finite set $X = \{x_1, \dots, x_m\} \subset \Omega$, $m \in \mathbb{N}$, we define the convex hull of X as

$$\operatorname{conv} X := \Big\{ \sum_{i=1}^m \alpha_i x_i \ \Big| \ \sum_{i=1}^m \alpha_i = 1, \ \alpha_i \geq 0 \ \forall \ i = 1, \dots, m \Big\}.$$

2.2. Functions of bounded variation

In this section, we will introduce the basic concepts of functions of bounded variations and state some important properties. We start with the definition of the space $BV(\Omega)$ for an open bounded set $\Omega \subset \mathbb{R}^n$ based on the Radon measure representation of the distributional derivative. For more details, we refer the reader to [3].

Definition 2.2.1 (The space $BV(\Omega)$)**.** A function $u \in L^1(\Omega)$ is called a *function of bounded variation*, if its distributional derivative is a finite vector-valued Radon measure, i.e. there exists $Du = (D_1 u, \dots, D_n u)^T \in \mathcal{M}(\Omega, \mathbb{R}^n)$ such that

$$\int_\Omega u \frac{\partial \phi}{\partial x_i} \mathrm{d}x = -\int_\Omega \phi \; \mathrm{d}D_i u \;\forall\; \phi \in C_0^\infty(\Omega),\; i = 1, \dots, n.$$

The space of all functions of bounded variation on Ω is denoted by $BV(\Omega)$.

$BV(\Omega)$ equipped with the norm

$$\|u\|_{BV} := \|u\|_{L^1} + |Du|(\Omega)$$

is a Banach space. Furthermore, the following standard result shows the density of smooth functions in $BV(\Omega)$ ([3], Theorem 3.9).

Theorem 2.2.2 (Density of smooth functions)**.** *Let $u \in L^1(\Omega)$. Then, $u \in BV(\Omega)$ if and only if there exists a sequence $(u_k) \subset C^\infty(\Omega)$ with $u_k \to u$ in $L^1(\Omega)$ and $\lim\limits_{k\to\infty} \|\nabla u_k\|_{L^1} < \infty$.*

In particular, this means that any function of bounded variation can be approximated by a sequence of smooth functions whose gradient is uniformly bounded in L^1.
An interesting property of BV-functions is the characterization of their distributional derivative. Let us first investigate some properties and introduce some notation. For more details, we refer the reader to [3], Chapter 3.

Definition 2.2.3 (Approximate discontinuity set)**.** A function $u \in L^1_{loc}(\Omega)$ is *approximately continuous* in a point $x \in \Omega$, if there exists $z \in \mathbb{R}$ such that

$$\lim_{\varepsilon\to 0} \frac{1}{|B_\varepsilon(x)|} \int_{B_\varepsilon(x)} |u(y) - z| \mathrm{d}y = 0. \tag{2.1}$$

We define the set S_u of all points where u is not approximately continuous as the *approximate discontinuity set.*

The value $z \in \mathbb{R}$ from the previous definition is uniquely determined by equation (2.1), thus z will be denoted by $\tilde{u}(x)$ and called the approximate limit of u in x. For a function $u \in L^1_{loc}(\Omega)$, one can show (see for instance [3], Proposition 3.64) that $\mathcal{L}^n(S_u) = 0$, i.e. S_u is a Lebesgue-negligible Borel set and thus, u is approximately continuous in $\mathcal{L}^n$-a.e. $x \in \Omega$. Those points belonging to S_u can be further distinguished by their affiliation to the set of approximate jump points defined in the following.

Definition 2.2.4 (Approximate jump set)**.** Let $u \in L^1_{loc}(\Omega)$. A point $x \in \Omega$ is an *approximate jump point* of u if there exist $a, b \in \mathbb{R}, a \neq b$ and $\nu \in S^{n-1}$ such that

$$\lim_{\varepsilon\to 0} \frac{1}{|B^+_\varepsilon(x,\nu)|} \int_{B^+_\varepsilon(x,\nu)} |u(y) - a| \mathrm{d}y = 0,\; \lim_{\varepsilon\to 0} \frac{1}{|B^-_\varepsilon(x,\nu)|} \int_{B^-_\varepsilon(x,\nu)} |u(y) - b| \mathrm{d}y = 0,$$

where $B^\pm_\varepsilon := \{y \in B_\varepsilon(x) \;:\; \langle y - x, \nu\rangle \gtrless 0\}$. The set of approximate jump points of u is denoted by J_u.

Obviously, if $x \in \Omega$ is an approximate jump point, x is also an approximate discontinuity point, hence we have $J_u \subset S_u$ and J_u is a Borel set. Moreover, the triplet (a, b, ν) is uniquely determined for every x by the definition (up to a permutation of a and b and the sign of ν) and admits a quite graphical intuition: If $x \in J_u$, then u has a jump in function value from a to b in the direction of ν, which can be shown to equal the direction of the unit normal on S_u in x. Therefore, for $x \in J_u$, we denote the triplet (a, b, ν) by $(u^+(x), u^-(x), \nu_u(x))$.
Outside of the set S_u, we can further define the set of approximate differentiability as in [3], Definition 3.70.

Definition 2.2.5 (Approximate differentiability)**.** Let $u \in L^1_{loc}(\Omega)$. u is called *approximately differentiable* at a point $x \in \Omega \setminus S_u$, if there exists a vector $L \in \mathbb{R}^n$ such that

$$\lim_{\varepsilon \to 0} \frac{1}{|B_\varepsilon(x)|} \int_{B_\varepsilon(x)} \frac{|u(y) - \tilde{u}(x) - \langle L, y - x \rangle|}{\varepsilon} \mathrm{d}y = 0. \tag{2.2}$$

If u is approximately differentiable at x, the vector L is uniquely determined by (2.2) and denoted by $\nabla u(x)$, which is called the approximate gradient of u in x.

Now we can employ the previous definitions in order to characterize the distributional gradient of a BV-function. From the Radon–Nikodym theorem (see for instance [3], Theorem 1.28), we obtain that Du can be decomposed into an absolutely continuous part $v\mathcal{L}^n$ and a singular part $D^s u$ with respect to the Lebesgue measure such that

$$Du = v\mathcal{L}^n + D^s u,$$

where $v \in L^1(\Omega, \mathbb{R}^n)$ is the density of Du with respect to $\mathcal{L}^n$. By the Calderón–Zygmund theorem (see [3], Theorem 3.83), $u \in BV(\Omega)$ is approximately differentiable almost everywhere in Ω and $v = \nabla u$. Furthermore, the remaining part $D^s u$ can be decomposed into a so-called *jump part* $D^j u$ and a *Cantor part* $D^c u$, which are defined as (see [3], Definition 3.91)

$$D^j u := D^s u \llcorner J_u, \;\; D^c u := D^s u \llcorner (\Omega \setminus S_u).$$

Let us have a closer look at the first part. By the Federer–Vol'pert theorem (see [3], Theorem 3.78), J_u is a countably $\mathcal{H}^{n-1}$-rectifiable set, which yields by [3], Theorem 3.77, that

$$D^j u(B) = \int_{B \cap J_u} (u^+(x) - u^-(x)) \nu_u(x) \; \mathrm{d}\mathcal{H}^{n-1} = \int_{B \cap S_u} (u^+(x) - u^-(x)) \nu_u(x) \; \mathrm{d}\mathcal{H}^{n-1}$$

for all $B \in \mathcal{B}(\Omega)$, where the second equality comes from the fact that $\mathcal{H}^{n-1}(S_u \setminus J_u) = 0$ for $u \in BV(\Omega)$. In other words, we have $D^j u = (u^+ - u^-)\nu_u \mathcal{H}^{n-1} \llcorner S_u$. Altogether, we obtain the decomposition

$$Du = \nabla u \mathcal{L}^n + (u^+ - u^-)\nu_u \mathcal{H}^{n-1} \llcorner S_u + D^c u.$$

The space of functions of bounded variation seems to be practical for investigating problems involving piecewise differentiable functions with a certain discontinuity set such as the Mumford–Shah problem, which will be further introduced in Section 2.4. Unfortunately, the distributional derivative still involves the less intuitive Cantor part $D^c u$. To overcome this problem, De Giorgi and Ambrosio introduced the space of so-called *special functions of bounded variation.*

Definition 2.2.6 (Special functions of bounded variation)**.** A function $u \in BV(\Omega)$ is a *special function of bounded variation*, if the Cantor part $D^c u$ of the derivative vanishes, i.e.

$$Du = \nabla u \mathcal{L}^n + (u^+ - u^-)\nu_u \mathcal{H}^{n-1} \llcorner S_u$$

by the above decomposition. The space of special functions of bounded variation is denoted by $SBV(\Omega)$.

2.3. Γ-convergence and equi-coercivity

For several numerical or analytical purposes, it might be of interest to consider a relaxation of a certain energy guided by a parameter ε in order to obtain more practical properties such as smoothness or convexity. Instead of an energy functional F, which might be somehow difficult to handle, one can introduce a family of relaxations F_ε to perform simulations or investigate features of the problem. A famous example is the Ambrosio–Tortorelli approximation of the Mumford–Shah functional, which will be further introduced in Section 2.4. Mathematically, one is interested in analysing the limit behaviour of F_ε for $\varepsilon \to 0$ in order to verify that F_ε is in some sense a good approximation of the functional F. Such a verification is provided by the notion of Γ-convergence as first introduced in [37]. In the following, we want to formulate the definition of Γ-convergence and state some fundamental properties. For more details, we refer the reader to [16] or [43].

Definition 2.3.1 (Γ-convergence)**.** Let (X, d) be a metric space. A family of functions $F_\varepsilon : X \to \mathbb{R} \cup \{-\infty, \infty\}$ for a parameter $\varepsilon > 0$ Γ-converges with respect to d to a functional $F : X \to \mathbb{R} \cup \{-\infty, \infty\}$, if for every $u \in X$ the following two properties are satisfied:

(1) (Γ-lim inf inequality.) For every sequence $(u_\varepsilon) \subset X$ with $d(u_\varepsilon, u) \to 0$ if $\varepsilon \to 0$,

$$F(u) \leq \liminf_{\varepsilon \to 0} F_\varepsilon(u_\varepsilon).$$

(2) (Γ-lim sup inequality.) There exists a sequence $(u_\varepsilon) \subset X$ with $d(u_\varepsilon, u) \to 0$ if $\varepsilon \to 0$ such that

$$\limsup_{\varepsilon \to 0} F_\varepsilon(u_\varepsilon) \leq F(u).$$

We then write $F_\varepsilon \xrightarrow{\Gamma} F$ and F is called Γ-limit of F_ε.

With the definition of Γ-convergence, we obtain a useful tool for approximating functionals. A point which has not been taken care of so far is the question whether minimizers of F_ε also approximate minimizers of F. Suppose that $(u_\varepsilon) \subset X$ is a minimizing sequence for F_ε in the sense that

$$\lim_{\varepsilon \to 0} \Big(F_\varepsilon(u_\varepsilon) - \inf_{u \in X} F_\varepsilon(u) \Big) = 0.$$

If u_ε converges to some $\bar{u} \in X$, then we easily see that

$$\limsup_{\varepsilon \to 0} \inf_{u \in X} F_\varepsilon(u) \leq \inf_{u \in X} F(u) \leq F(\bar{u}) \leq \liminf_{\varepsilon \to 0} F_\varepsilon(u_\varepsilon) \leq \liminf_{\varepsilon \to 0} \inf_{u \in X} F_\varepsilon(u),$$

thus it follows that

$$F(\bar{u}) = \min_{u \in X} F(u) = \lim_{\varepsilon \to 0} \inf_{u \in X} F_\varepsilon(u).$$

The crucial part is the verification that the minimizing sequence (u_ε) converges in X (at least up to subsequences). This requires some compactness of the functionals F_ε, which directly leads to the definition of equi-coercivity.

Definition 2.3.2 (Equi-coercivity)**.** A sequence of functionals (F_ε) is called equi-coercive, if for every $t \in \mathbb{R}$, there exists a compact K_t such that $\{F_\varepsilon < t\} \subset K_t$ for every ε.

The equi-coercivity of (F_ε) leads to the fundamental property of Γ-convergence, as stated in [16], Theorem 1.21.

Theorem 2.3.3. *Let (X, d) be a metric space, (F_ε) with $F_\varepsilon : X \to \mathbb{R} \cup \{-\infty, \infty\}$ a sequence of equi-coercive functionals and $F_\varepsilon \xrightarrow{\Gamma} F$ for a functional $F : X \to \mathbb{R} \cup \{-\infty, \infty\}$. Then*

$$\min_{u \in X} F(u) = \lim_{\varepsilon \to 0} \inf_{u \in X} F_\varepsilon(u)$$

and if $(u_\varepsilon) \subset X$ is a minimizing sequence for F_ε, then u_ε converges (up to subsequences) to a minimizer of F.

2.4. The Mumford–Shah image segmentation problem

An interesting problem in the field of mathematical imaging and a famous example for a free-discontinuity problem is the so-called Mumford–Shah image segmentation problem, presented by D. Mumford and J. Shah in [50]. Given an image $f \in L^\infty(\Omega)$ on some image domain $\Omega \subset \mathbb{R}^n$, one aims at finding a piecewise smooth approximation u of f, which involves an unknown discontinuity set K. The Mumford–Shah energy functional $\widetilde{MS} : \mathcal{D} \to [0, \infty]$ is defined by

$$\widetilde{MS}(K, u) = \int_{\Omega \setminus K} \alpha(u - f)^2 + |\nabla u|^2 \mathrm{d}x + \beta \mathcal{H}^{n-1}(K \cap \Omega), \tag{2.3}$$

with $\mathcal{D} = \{(K, u) : \ K \subset \overline{\Omega} \text{ closed}, \ u \in W^{1,2}_{loc}(\Omega \setminus K)\}$ and parameters $\alpha, \beta > 0$. The first term of the functional guarantees that u is close to the original input f and smooth outside of K by the second term, while the length of the discontinuity set K is penalized on the other hand.

Existence of a minimizer was first shown in [36]. The proof requires some work since the direct method of calculus of variations fails due to the fact that the mapping $K \mapsto \mathcal{H}^{n-1}(K)$ in general is not lower semi-continuous with respect to the Hausdorff metric

$$d(K_1, K_2) = \inf\{r > 0 : \ K_1 \subset B_r(K_2), \ K_2 \subset B_r(K_1)\}.$$

Instead, the authors choose a relaxation of the original energy, which is also referred to as the Mumford–Shah problem in the literature, by defining a functional $MS : SBV(\Omega) \to [0, \infty]$,

$$MS(u) = \int_\Omega \alpha(u - f)^2 + |\nabla u|^2 \mathrm{d}x + \beta \mathcal{H}^{n-1}(S_u).$$

The existence result is then obtained by showing that $\inf\limits_{u \in SBV(\Omega)} MS(u) = \inf\limits_{(K,u) \in \mathcal{D}} \widetilde{MS}(K, u)$ (see [36] or [3], Chapter 6, for a detailed proof), since MS has a minimizer in $SBV(\Omega)$ thanks to a general compactness result in $SBV(\Omega)$ [3].

Note that this relaxation approach gives a natural justification for the definition of the space of *special* functions of bounded variation. Intuitively, one would look for a minimizer in $BV(\Omega)$, but this space turns out to be too large: By defining the set of Cantor-like BV-functions $BV^c(\Omega) := \{u \in BV(\Omega) : \ Du = D^c u\}$, meaning those functions whose gradient only consists of the Cantor part, one can see that

$$\inf_{u \in BV(\Omega)} MS(u) \leq \inf_{u \in BV^c(\Omega)} MS(u) = \inf_{u \in BV^c(\Omega)} \alpha \int_\Omega (u - f)^2 \mathrm{d}x = 0,$$

where the last equation comes from the fact that $BV^c(\Omega)$ is dense in $L^2(\Omega)$.

Although the existence of minimizers of MS is assured, *identifying* them remains the more complicated task due to the non-convexity of the energy [2] (except some situations where the problem can be reduced to one space dimension [3]). This is especially a crucial issue for all numerical purposes and automatically raises the question for any kind of relaxation, which in the optimal case should yield a convex reformulation.

The Mumford–Shah functional defined on SBV-functions is often seen as a model for a more general class of free-discontinuity problems. Since other functionals belonging to this group share the same behaviour in some aspects, it can be useful to investigate so-called *Mumford–Shah-type functionals.*

Definition 2.4.1 (Mumford–Shah-type functionals)**.** Let $\Omega \subset \mathbb{R}^n$, $g : \Omega \times \mathbb{R} \times \mathbb{R}^n \to [0, \infty]$, $h : \Omega \times \mathbb{R} \times \mathbb{R} \times S^{n-1} \to [0, \infty]$. The functional $F : SBV(\Omega) \to [0, \infty]$ defined as

$$F(u) = \int_\Omega g\left(x, u(x), \nabla u(x)\right) \mathrm{d}x + \int_{S_u} h\left(x, u^+, u^-, \nu_u\right) \mathrm{d}\mathcal{H}^{n-1}(x)$$

is called *Mumford–Shah-type functional*, where u^+, u^-, ν_u are defined as in Definition 2.2.4 and the subsequent paragraph.

One can easily see that MS is a special case of F with $g(x, u(x), \nabla u(x)) = \alpha(u(x) - f(x))^2 + |\nabla u(x)|^2$ and $h(x, u^+, u^-, \nu_u) = \beta$.

2.4.1. Phase field approximation of the Mumford–Shah functional

Since the difficulty of computing an optimal pair (K, u) of the original Mumford–Shah functional (2.3) lies mainly in the non-regularity of the discontinuity term, a natural idea is to approximate K by something more regular. This idea has been implemented by the Modica–Mortola theorem [47]: Given a set $E \subset \Omega$, the perimeter of E can be approximated in the sense of Γ-convergence via the sequence of functionals $MM_\varepsilon : L^2(\Omega) \to [0, \infty]$,

$$MM_\varepsilon(v) = \begin{cases} \int_\Omega \varepsilon|\nabla v|^2 + \frac{1}{\varepsilon}W(v) \,\mathrm{d}x & \text{if } v \in W^{1,2}(\Omega), \\ \infty & \text{if } v \in L^2(\Omega) \setminus W^{1,2}(\Omega), \end{cases} \tag{2.4}$$

where W is a so-called double well potential, a continuous non-negative function vanishing only at two points, such as $W(t) = t^2(t-1)^2$ for instance. The result of Modica–Mortola is stated in the following theorem [47].

Theorem 2.4.2 (Modica–Mortola)**.** *The sequence MM_ε in* (2.4) *Γ-converges in $L^2(\Omega)$ to the functional*

$$F(v) = \begin{cases} cPer(E, \Omega) & \textit{if } v = \chi_E \textit{ for some } E \in \mathcal{B}(\Omega), \\ \infty & \textit{otherwise,} \end{cases}$$

where $c = 2\int_0^1 \sqrt{W(s)}\mathrm{d}s$ and $Per(E, \Omega)$ is the perimeter of the set E in Ω.

Proof. See for instance [17], Theorem 7.3. □

Graphically, v provides for some kind of phase transition: While the first term in MM_ε guarantees the smoothness of v, the second term causes v to preferentially stay in its two "phases" specified by the minima of the double well. However, without an additional constraint, the minimizer of the functional (2.4) obviously satisfies $v \equiv 0$ or $v \equiv 1$ almost everywhere. The connection to the discontinuity set of a piecewise smooth image is drawn by some additional linker terms between the phase field v and the image u by the so-called *Ambrosio–Tortorelli functional* $AT_\varepsilon : L^2(\Omega) \times L^2(\Omega) \to [0, \infty]$ defined as

$$AT_\varepsilon(u, v) = \begin{cases} \int_\Omega \alpha(u-g)^2 + v^2|\nabla u|^2 + \beta/2\Big(\varepsilon|\nabla v|^2 + \frac{1}{\varepsilon}(v-1)^2\Big) \,\mathrm{d}x & \text{if } (u, v) \in \mathcal{D}, \\ \infty & \text{otherwise.} \end{cases} \tag{2.5}$$

where $\mathcal{D} = \{(u,v) \,:\, u,v \in W^{1,2}(\Omega),\ 0 \leq v \leq 1\}$. Note that the double well potential W in the Modica–Mortola functionals has been replaced by a single well $(v-1)^2$, however, the role of the second well is assumed by the term $v^2|\nabla u|^2$. The following Γ-convergence result was proved by Ambrosio and Tortorelli in [4] (see also [3]).

Theorem 2.4.3 (Ambrosio–Tortorelli)**.** *The sequence of functionals* (2.5) Γ*-converges in* $L^2(\Omega) \times L^2(\Omega)$ *to the functional*

$$F(u,v) = \begin{cases} \int_\Omega \alpha(u-g)^2 + |\nabla u|^2 \mathrm{d}x + \beta \mathcal{H}^{n-1}(S_u) & \text{if } u \in SBV(\Omega),\ v \equiv 1 \text{ a.e.}, \\ \infty & \text{otherwise.} \end{cases}$$

Proof. See [4]. □

Via a heuristic explanation of the terms in (2.5), it becomes clear that v can indeed be interpreted as a smooth representation of the discontinuity set S_u: The factor v^2 in front of the gradient term of u must tend to zero very close to the discontinuity set of u to ensure the boundedness of the terms. On the other hand, if $\varepsilon \to 0$, the factor $\frac{1}{\varepsilon}$ in front of the term $(v-1)^2$ becomes very large, thus v has to go to 1, the only point where $(v-1)^2$ vanishes. As a consequence, v has to admit some kind of phase transition, which becomes sharper for $\varepsilon \to 0$. Figure 2.1 shows an example for a piecewise smooth approximation of an image obtained via minimization of the Ambrosio–Tortorelli functional.

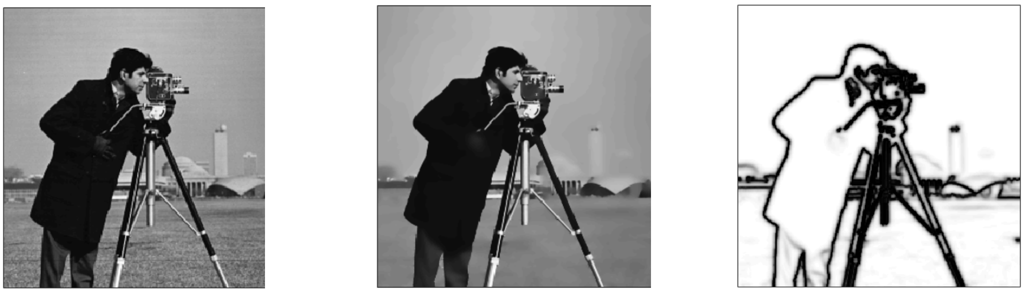


Figure 2.1.: Numerical simulation of the Ambrosio–Tortorelli approximation of the Mumford–Shah functional. Left: Original image f. Middle: Piecewise smooth approximation u of f. Right: Approximation v of the discontinuity set S_u.

The Ambrosio–Tortorelli functional naturally finds sustainable use in numerical applications. From Theorem 2.4.3, for small ε, one can expect a minimizer u_ε of AT_ε to yield a good approximation of the minimizer u of the Mumford–Shah functional. Additionally, one can prove that under certain conditions on the lattice size of a discretization, the Γ-convergence result still holds on a discrete version of the functionals (see for instance [8] for a finite element approximation or [6] for finite differences). However, the Ambrosio–Tortorelli functional remains non-convex in the pair (u,v), which raises the problem of finding a suitable minimization algorithm.

2.4.2. Functional lifting of Mumford–Shah-type problems

In a series of articles [1],[2] the authors introduced a new representation for the non-convex functional

$$J(u) = \int_\Omega |\nabla u|^2 \mathrm{d}x + \beta \mathcal{H}^{n-1}\left(S_u\right).$$

By defining the characteristic function of the subgraph of $u \in SBV(\Omega)$,

$$1_u : \Omega \times \mathbb{R} \to \{0,1\},\ 1_u(x,t) := \begin{cases} 1 & \text{if } u(x) > t, \\ 0 & \text{otherwise,} \end{cases}$$

they show that for every $u \in SBV(\Omega)$,

$$J(u) = \sup_{\phi \in K} \int_{\Omega \times \mathbb{R}} \phi \cdot \mathrm{d}D1_u$$

with

$$K = \Big\{ \phi = (\phi^x, \phi^s) \in C_0^\infty(\Omega \times \mathbb{R}, \mathbb{R}^n \times \mathbb{R}) \ : \\ |\phi^x(x,s)|^2 \le 4\phi^s(x,s)\ \ \forall (x,s) \in \Omega \times \mathbb{R}, \left| \int_{s_1}^{s_2} \phi^x(x,s)\,\mathrm{d}s \right| \le 1\ \forall x \in \Omega,\ s_1, s_2 \in \mathbb{R} \Big\}.$$

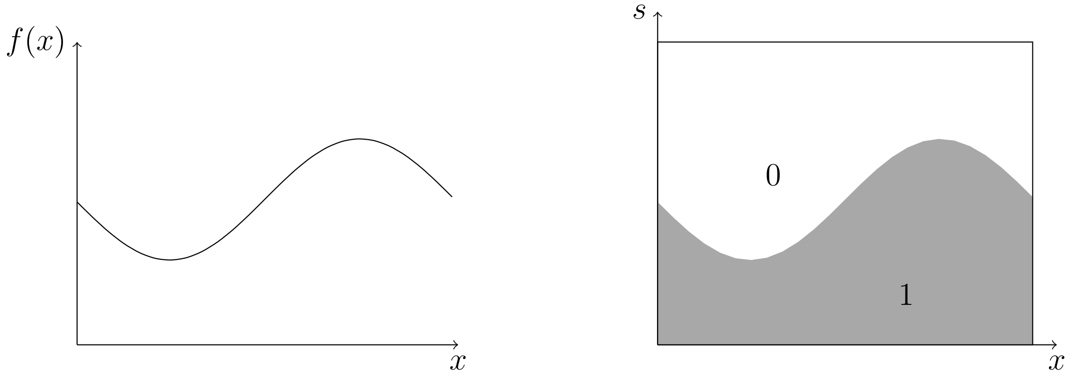


Figure 2.2.: Illustration of functional lifting from a scalar to a two-dimensional function. Left: Graph of a function u, right: Lifted function 1_u.

This idea can be extended to more general functionals of the form

$$F(u) = \int_\Omega g(x, u(x), \nabla u(x))\,\mathrm{d}x + \int_{S_u} h\left(x, u^+, u^-, \nu\right) \mathrm{d}\mathcal{H}^{n-1}(x) \tag{2.6}$$

as defined in Definition 2.4.1 for $u \in SBV(\Omega)$. The main result from [1],[2] is stated in the following theorem. We will give a sketch of the proof here, for more details we refer the reader to [2].

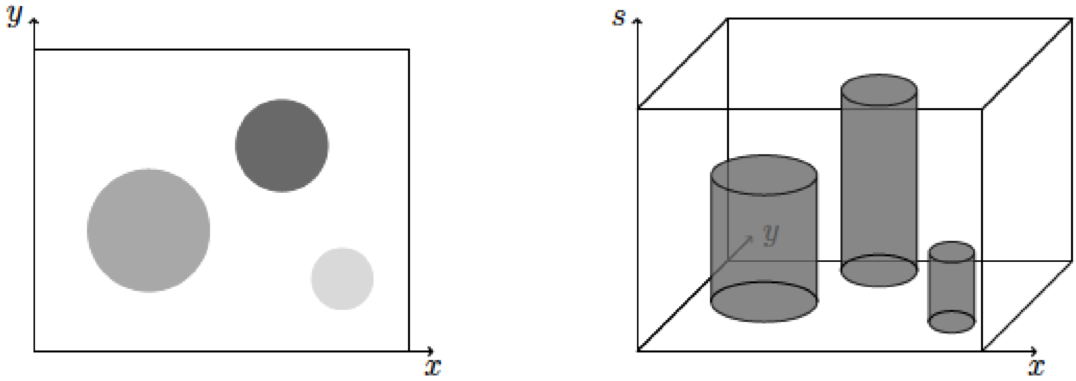


Figure 2.3.: Illustration of functional lifting from a two-dimensional to a three-dimensional function. Left: Image u, right: Lifted function 1_u. The height of the lifted function 1_u in a point $x \in \Omega$ corresponds to the value $u(x)$.

Theorem 2.4.4. *Let $\Omega \subset \mathbb{R}^n$, $F : SBV(\Omega) \to [0, \infty]$ and S_u, u^+, u^- be defined as above. Then*

$$F(u) \geq \sup_{\phi \in \mathcal{K}} \int_{\Omega \times \mathbb{R}} \phi \cdot \mathrm{d}D1_u \tag{2.7}$$

with

$$\begin{aligned}\mathcal{K} = \Big\{ &\phi = (\phi^x, \phi^s) \in C_0^\infty(\Omega \times \mathbb{R}; \mathbb{R}^n \times \mathbb{R}) \;: \\ &\phi^s(x,s) \geq g^*(x, s, \phi^x(x,s)) \;\; \forall (x,s) \in \Omega \times \mathbb{R}, \\ &\left| \int_{s_1}^{s_2} \phi^x(x,s)\,\mathrm{d}s \right| \leq h(x, s_1, s_2, \nu) \;\; \forall x \in \Omega, s_1 < s_2, \nu \in S^{n-1} \Big\}. \end{aligned} \tag{2.8}$$

Proof. Let Γ_u be the graph of u, i.e. the singular set of the lifted functional 1_u (Γ_u is well-defined since the subgraph of u has finite perimeter in $\Omega \times \mathbb{R}$, cf. [2], Definition 2.7). Since the characteristic function 1_u is piecewise constant for u in $SBV(\Omega)$, its distributional gradient $D1_u$ has no Lebesgue and Cantor part, hence

$$D1_u = \nu_{\Gamma_u} \cdot \mathcal{H}^n \llcorner \Gamma_u,$$

where

$$\nu_{\Gamma_u}(x,s) = \begin{cases} \frac{1}{\sqrt{|\nabla u(x)|^2 + 1}} (\nabla u(x), -1)^T & \text{for } x \in \Omega \setminus S_u, \\ (\nu_u(x), 0)^T & \text{for } x \in S_u \end{cases}$$

is the inner unit normal of the subgraph of u (extended to $\Omega \times \mathbb{R}$) and ν_u is the unit normal on S_u pointing from u^- to u^+. As a consequence, we can write the right-hand side of the inequality (2.7) as

$$\begin{aligned}
&\int_{\Omega\times\mathbb{R}} \phi\left(x,s\right)\cdot \mathrm{d}D1_u\left(x,s\right) = \int_{\Gamma_u} \phi\left(x,u(x)\right)\cdot \nu_{\Gamma_u}\left(x\right)\mathrm{d}\mathcal{H}^n\left(x\right)\\
&= \int_{\Omega\setminus S_u} \phi\left(x,u\left(x\right)\right)\cdot \nu_{\Gamma_u}\left(x\right)\mathrm{d}x + \int_{S_u}\left(\int_{u^-(x)}^{u^+(x)} \phi^x\left(x,s\right)\mathrm{d}s\right)\cdot \nu_u\left(x\right)\mathrm{d}\mathcal{H}^{n-1}\left(x\right)\\
&= \int_{\Omega\setminus S_u} \phi^x\left(x,u\left(x\right)\right)\cdot \nabla u\left(x\right) - \phi^s\left(x,u\left(x\right)\right)\mathrm{d}x + \int_{S_u}\left(\int_{u^-(x)}^{u^+(x)} \phi^x\left(x,s\right)ds\right)\cdot \nu_u\left(x\right)\mathrm{d}\mathcal{H}^{n-1}\left(x\right)
\end{aligned}$$

Now let us regard the two parts separately and compare with the original functional J. From the constraint set $\mathcal{K}$ we obtain

$$\begin{aligned}
&\phi^s\left(x,s\right) \geq g^*\left(x,s,\phi^x\left(x,s\right)\right) = \sup_{\psi\in\mathbb{R}^n}\left(\phi^x\left(x,s\right)\cdot\psi - g\left(x,s,\psi\right)\right)\ \forall\ \left(x,s\right)\in\Omega\times\mathbb{R}\\
\Rightarrow\ &\phi^s\left(x,u\left(x\right)\right) \geq \phi^x\left(x,u\left(x\right)\right)\cdot\nabla u\left(x\right) - g\left(x,u\left(x\right),\nabla u\left(x\right)\right)\\
\Leftrightarrow\ &\phi^x\left(x,u\left(x\right)\right)\cdot\nabla u\left(x\right) - \phi^s\left(x,u\left(x\right)\right) \leq g\left(x,u\left(x\right),\nabla u\left(x\right)\right). \qquad (2.9)
\end{aligned}$$

For the second part we have

$$\begin{aligned}
&\left|\int_{s_1}^{s_2}\phi^x\left(x,s\right)\mathrm{d}s\right| \leq h\left(x,s_1,s_2,\nu\right)\ \forall\ x\in\Omega,\nu\in S^{n-1},s_1<s_2\\
\Rightarrow\ &h\left(x,u^+,u^-,\nu_u\right) \geq \left|\int_{u^-}^{u^+}\phi^x\left(x,s\right)\mathrm{d}s\right| \geq \int_{u^-}^{u^+}\phi^x\left(x,s\right)\mathrm{d}s\cdot\nu_u. \qquad (2.10)
\end{aligned}$$

The inequalities (2.9) and (2.10) together result in (2.7). □

For a given $u\in SBV(\Omega)$, one can easily derive conditions for equality in (2.7) (cf. [2]).

Corollary 2.4.5. *For a given $u\in SBV(\Omega)$, let $\phi=(\phi^x,\phi^s)\in C_0^\infty(\Omega\times\mathbb{R},\mathbb{R}^n\times\mathbb{R})$ be a vector field which satisfies the following assumptions:*

- $\phi^s(x,u(x)) = g^*(x,u(x),\phi^x(x,u(x)))\ \forall\ x\in\Omega,$
- $\phi^x(x,u(x))\in\partial g(x,u(x),\nabla u(x))\ \forall\ x\in\Omega,$ *where ∂g denotes the subdifferential of g with respect to the last variable,*
- $\left(\int_{u^-(x)}^{u^+(x)}\phi^x(x,s)\mathrm{d}s\right)\cdot\nu_u = h(x,u^-(x),u^+(x),\nu_u(x))\ \forall\ x\in S_u.$

Then we have

$$F(u) = \int_{\Omega\times\mathbb{R}}\phi\cdot\mathrm{d}D1_u.$$

Remark 2.4.6. The existence of such a vector field satisfying the conditions in Corollary 2.4.5 is a crucial issue in the general case. In [2], Remark 3.3, the authors state that for the Mumford–Shah functional, one can construct a vector field ϕ such that (2.9) and (2.10) are almost equalities up to an arbitrarily small error, such that one obtains equality in (2.7).

Based on this idea, in [54] and [55] the authors develop an algorithmic framework to handle the non-convexity of the Mumford–Shah functional. Although the lifted problem (2.7) is still non-convex due to the binary function 1_u, it is straightforward to find a convexification by substituting 1_u by a function $v : \Omega \times \mathbb{R} \to [0,1]$. To this end, we introduce the set

$$\mathcal{C} = \{v \in SBV\left(\Omega \times \mathbb{R}, [0,1]\right) \; : \; \lim_{t\to-\infty} v\left(x,t\right) = 1, \; \lim_{t\to\infty} v\left(x,t\right) = 0\}$$

and define the functional $\mathcal{F} : SBV(\Omega \times \mathbb{R}) \to [0,\infty]$ as

$$\mathcal{F}(v) = \sup_{\phi\in\mathcal{K}} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}Dv.$$

The set $\mathcal{C}$ can be seen as an extension of the set containing all binary functions $v = 1_u$ for some $u \in SBV\left(\Omega\right)$. With this relaxation, the original problem can be rewritten as

$$\inf_{u\in SBV(\Omega)} F(u) \geq \inf_{u\in SBV(\Omega)} \sup_{\phi\in\mathcal{K}} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}D1_u \geq \inf_{v\in\mathcal{C}} \mathcal{F}(v), \tag{2.11}$$

where the right-hand side is a convex problem in v.

Definition 2.4.7 (Functional lifting problem)**.** The saddle point problem

$$\inf_{v\in\mathcal{C}} \sup_{\phi\in\mathcal{K}} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}Dv$$

with $\mathcal{C}, \mathcal{K}$ defined above is referred to as a *functional lifting problem.*

The inequality in (2.11) states that the convex relaxation might come along with some "loss" of accuracy concerning the original problem of minimizing F. This raises the question whether under certain conditions the inequality can be turned into an equality such that the minima of F and $\mathcal{F}$ coincide. As shown by the following result, this question is closely related to the existence of a divergence-free vector field ϕ realizing the supremum in (2.7).

Theorem 2.4.8 (Equality of the minima)**.** *Let $\hat{u} \in SBV(\Omega)$ be a minimizer of F. If there exists a divergence-free vector field $\hat{\phi} \in \mathcal{K}$ such that*

$$F(\hat{u}) = \sup_{\phi\in\mathcal{K}} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}D1_{\hat{u}} = \int_{\Omega\times\mathbb{R}} \hat{\phi} \cdot \mathrm{d}D1_{\hat{u}},$$

then we have

$$\min_{u\in SBV(\Omega)} F(u) = \inf_{v\in\widetilde{\mathcal{C}}} \mathcal{F}(v),$$

where

$$\widetilde{\mathcal{C}} = \{v \in SBV(\Omega \times \mathbb{R}, [0,1]) : \lim_{t\to-\infty} v\left(x,t\right) = 1, \; \lim_{t\to\infty} v\left(x,t\right) = 0, \; v = 1_{\hat{u}} \textit{ on } \partial\Omega \times \mathbb{R}\}.$$

Proof. From $1_{\hat{u}} \in \widetilde{\mathcal{C}}$, we obtain

$$\begin{aligned}\mathcal{F}(1_{\hat{u}}) &= \sup_{\phi\in\mathcal{K}} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}D1_{\hat{u}} = \int_{\Omega\times\mathbb{R}} \hat{\phi} \cdot \mathrm{d}D1_{\hat{u}} = F(\hat{u}) = \min_{u\in SBV(\Omega)} F(u) \geq \inf_{v\in\widetilde{\mathcal{C}}} \mathcal{F}(v) \\ &= \inf_{v\in\widetilde{\mathcal{C}}} \sup_{\phi\in\mathcal{K}} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}Dv \geq \sup_{\phi\in\mathcal{K}} \inf_{v\in\widetilde{\mathcal{C}}} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}Dv \geq \inf_{v\in\widetilde{\mathcal{C}}} \int_{\Omega\times\mathbb{R}} \hat{\phi} \cdot \mathrm{d}Dv. \qquad (2.12)\end{aligned}$$

By the divergence theorem and due to the fact that $\hat{\phi}$ is divergence-free, we obtain

$$\int_{\Omega\times\mathbb{R}} \hat{\phi} \cdot \mathrm{d}Dv = \int_{\partial\Omega\times\mathbb{R}} v\ \hat{\phi} \cdot \nu\ \mathrm{d}\mathcal{H}^{n-1} - \underbrace{\int_{\Omega\times\mathbb{R}} v\ \mathrm{div}(\hat{\phi})\ \mathrm{d}x\mathrm{d}s}_{=0} = \int_{\partial\Omega\times\mathbb{R}} v\ \hat{\phi} \cdot \nu\ \mathrm{d}\mathcal{H}^{n-1}$$

where ν is the outer unit normal to $\partial\Omega \times \mathbb{R}$. Since every $v \in \widetilde{\mathcal{C}}$ coincides with $1_{\hat{u}}$ on $\partial\Omega \times \mathbb{R}$, we have

$$\inf_{v\in\widetilde{\mathcal{C}}} \int_{\Omega\times\mathbb{R}} \hat{\phi} \cdot \mathrm{d}Dv = \inf_{v\in\widetilde{\mathcal{C}}} \int_{\partial\Omega\times\mathbb{R}} 1_{\hat{u}}\ \hat{\phi} \cdot \nu\ \mathrm{d}\mathcal{H}^{n-1} = \mathcal{F}(1_{\hat{u}}). \qquad (2.13)$$

Hence by combining (2.12) and (2.13), we conclude

$$\min_{u\in SBV(\Omega)} F(u) = \inf_{v\in\widetilde{\mathcal{C}}} \mathcal{F}(v).$$

□

2.5. Adaptive finite elements for functional lifting problems

In this section, we want to introduce some basic definitions and concepts of adaptive finite elements for problems arising from functional lifting, i.e. on a three-dimensional domain $\Omega \times \mathbb{R}$, where the additional dimension potentially requires a special treatment. We will restrict ourselves for the sake of simplicity to $[0,1]^2 = \Omega \subset \mathbb{R}^2$ and a three-dimensional domain $G = [0,1]^3$.

In the past years, adaptive grid refinement of simplicial n-dimensional grids has been extensively studied (e.g. [64], [62], [45], [63]). In three space dimensions, a finite element grid typically consists of tetrahedrons, which can be locally divided into two subelements by bisection. These grids form a suitable basis for numerically solving partial differential equations which require a different resolution in different parts of the domain. Tetrahedron elements are practical due to several reasons. On the one hand, one can easily define piecewise polynomial, for instance piecewise linear, basis functions which are globally continuous and only have very local support. On the other hand, simplicial grids allow for local refinement without necessarily creating so-called *hanging nodes*.

However, a severe disadvantage that does not play a role in most applications, but does in case of a problem domain created by functional lifting, is the fact that all spatial dimensions are treated in a similar way. In other words, thinking of a two-dimensional image domain lifted to the three-dimensional space, a projection of an adaptive grid onto the xy-plane would in general not provide a two-dimensional simplicial grid, which makes the process of functional lifting and *delifting* more difficult. From Theorem 2.4.4 we obtain that a method dealing with the lifted saddle point problem needs to handle line integrals along the lifted dimension, which requires simple communication between consecutive points in this direction. Summarizing, a proper finite element grid for functional lifting problems should provide *two* discretizations for the original image domain Ω as well as for the lifted domain $\Omega \times \mathbb{R}$ together with a straightforward way of communication between both.

2.5.1. Triangular prism finite elements

In order to include the desired properties as described above, we propose a novel discretization class consisting of a simplicial grid for the image domain $\Omega = [0,1]^2$ coupled with an additional partition of the image range $[0,1]$. The resulting three-dimensional grid contains triangular prism-shaped elements and admits a suitable refinement technique which allows for inheritance of some useful properties. For a more detailed understanding, we want to introduce the basic concepts of triangular prism finite elements within this section. Throughout the rest of this thesis, we will make use of a special notation concerning the coordinates related to the three-dimensional domain arising from functional lifting of a two-dimensional function. To emphasize the difference between the original image domain and the image range, for a point $x \in \mathbb{R}^3$, we denote its first two coordinates as xy-coordinates and the third one as s-coordinate with respect to the standard basis of $\mathbb{R}^3$. In the same way, we will use the notation of xy and s when speaking of any relating concept.
We start by recalling the definition of a two-dimensional simplicial grid (see [63]).

Definition 2.5.1 (Simplex, simplicial grid in 2D)**.** A *two-dimensional simplex* (x_0, x_1, x_2) is a 3-tuple with nodes $x_0, x_1, x_2 \in \mathbb{R}^2$, which do not lie on a one-dimensional hyperplane. The convex hull $\operatorname{conv}\{x_0, x_1, x_2\}$ is also denoted as a simplex. A *two-dimensional simplicial grid* is a connected set of two-dimensional simplices with pairwise disjoint interior.

Based on a two-dimensional simplicial grid for the image domain Ω, we define a *lifted counterpart* consisting of triangular prism-shaped elements.

Definition 2.5.2 (Triangular prism element)**.** A *triangular prism element* $(x_0, \dots, x_5)$ is a 6-tuple with nodes $x_0, \dots, x_5 \in \mathbb{R}^3$, such that $(\mathcal{P}_{H_{xy}}(x_0), \mathcal{P}_{H_{xy}}(x_1), \mathcal{P}_{H_{xy}}(x_2)) = (\mathcal{P}_{H_{xy}}(x_3), \mathcal{P}_{H_{xy}}(x_4), \mathcal{P}_{H_{xy}}(x_5))$ is a two-dimensional simplex and $h(x_0) = h(x_1) = h(x_2)$, $h(x_3) = h(x_4) = h(x_5)$, $h(x_0) < h(x_3)$, where H_{xy} denotes the set of all points in $\mathbb{R}^3$ whose s-coordinate equals zero and $h(x_i)$ denotes the s-coordinate of a point $x_i \in \mathbb{R}^3$. For a triangular prism element T, we define

- $\mathcal{N}(T) := \{x_0, \dots, x_5\}$ is the set of nodes
- $\mathcal{E}_h(T) := \{\text{conv}\{x_0, x_1\}, \text{conv}\{x_1, x_2\}, \text{conv}\{x_0, x_2\}, \text{conv}\{x_3, x_4\}, \text{conv}\{x_4, x_5\}, \text{conv}\{x_3, x_5\}\}$ is the set of horizontal edges
- $\mathcal{E}_v(T) := \{\text{conv}\{x_0, x_3\}, \text{conv}\{x_1, x_4\}, \text{conv}\{x_2, x_5\}\}$ is the set of vertical edges
- $\mathcal{E}(T) := \mathcal{E}_h(T) \cup \mathcal{E}_v(T)$ is the set of edges
- $\mathcal{F}_h(T) := \{\text{conv}\{x_0, x_1, x_2\}, \text{conv}\{x_3, x_4, x_5\}\}$ is the set of horizontal faces
- $\mathcal{F}_v(T) := \{\text{conv}\{x_0, x_1, x_3, x_4\}, \text{conv}\{x_1, x_2, x_4, x_5\}, \text{conv}\{x_0, x_2, x_3, x_5\}\}$ is the set of vertical faces
- $\mathcal{F}(T) := \mathcal{F}_h(T) \cup \mathcal{F}_v(T)$ is the set of faces.

The domain of a triangular prism element $T = (x_0, \dots, x_5)$ is also denoted as T, as far as ambiguity is beyond question. T can be written as $T = T_{xy} \times T_s$ for a two-dimensional simplex $T_{xy} = \text{conv}\{x_0, x_1, x_2\}$ and an interval $T_s = [h(x_0), h(x_3)]$.

Proposition 2.5.3. *The nodes $x_0, \dots, x_5$ of a triangular prism element T are listed in an ascending order by their s-coordinate first (where the secondary ordering does not play a role). Each pair (x_0, x_3), (x_1, x_4) and (x_2, x_5) has the same xy-coordinates. $h(T) := |x_3 - x_0|$ ($= |x_4 - x_1| = |x_5 - x_2|$) defines the height of an element.*

Next, we define a regular grid partition of the three-dimensional domain consisting of triangular prism elements.

Definition 2.5.4 (Regular triangular prism grid)**.** A *regular triangular prism grid $\mathcal{T}$* is a set of triangular prism elements such that $G \subseteq \bigcup\{T : \ T \in \mathcal{T}\}$ and for every pair of elements $T, S \in \mathcal{T}$, one of the following relations holds:

- $T \cap S = \emptyset$
- $T \cap S \in \mathcal{N}(T)$ and $T \cap S \in \mathcal{N}(S)$
- $T \cap S \in \mathcal{E}(T)$ and $T \cap S \in \mathcal{E}(S)$
- $T \cap S \in \mathcal{F}(T)$ and $T \cap S \in \mathcal{F}(S)$
- $T = S$.

The conditions above are denoted as *regularity conditions*. The set of nodes of $\mathcal{T}$ is denoted by

$$\mathcal{N}(\mathcal{T}) = \bigcup_{T \in \mathcal{T}} \mathcal{N}(T).$$

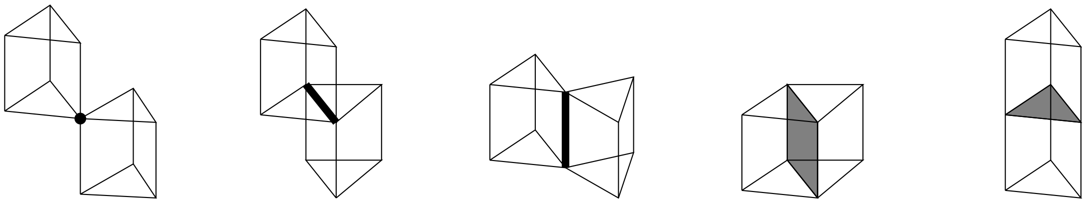


Figure 2.4.: Possible neighbouring relations of two elements in a regular triangular prism grid.

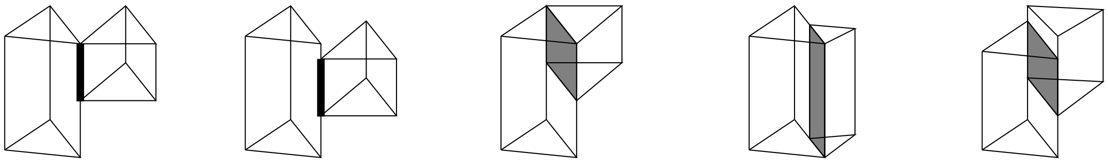


Figure 2.5.: Examples for neighbouring relations that are not allowed in a regular triangular prism grid.

Definition 2.5.4 introduces conditions about the neighbouring relations of two elements in a regular grid. In particular, these conditions prohibit elements with *partially* shared edges or faces (see Figure 2.4 and 2.5) and thus prevent the occurrence of so-called *hanging nodes*.

Definition 2.5.5 (Hanging node)**.** For an element $T \in \mathcal{T}$, a node $N \in \mathcal{N}(T)$ is called *hanging node*, if there exists an element $S \in \mathcal{T}$ with $N \in S$ and $N \notin \mathcal{N}(S)$. A hanging node N is called *xy-hanging*, if $N \in \mathcal{E}_h(S)$, and *s-hanging*, if $N \in \mathcal{E}_v(S)$. The set of hanging nodes of $\mathcal{T}$ is denoted by $\mathcal{H}(\mathcal{T})$.

Note that one can easily show that a regular triangular prism grid cannot contain any hanging node, as stated by the following lemma.

Lemma 2.5.6. *A regular triangular prism grid $\mathcal{T}$ does not contain any hanging node.*

Proof. Let $T, S \in \mathcal{T}$ with $N \in \mathcal{N}(T)$, $N \in S$ and $N \notin \mathcal{N}(S)$. Then $T \cap S \neq \emptyset$, $T \neq S$ and obviously $T \cap S \notin \mathcal{N}(S)$. Additionally, if $T \cap S \in \mathcal{E}(T)$, then there exist $x_{i_1}, x_{i_2} \in \mathcal{N}(T)$ such that $T \cap S = \text{conv}\{x_{i_1}, x_{i_2}\}$, hence $N = x_{i_1}$ or $N = x_{i_2}$. If also $T \cap S \in \mathcal{E}(S)$, with the same argument there exist $x_{j_1}, x_{j_2} \in \mathcal{N}(S)$ such that $T \cap S = \text{conv}\{x_{j_1}, x_{j_2}\} = \text{conv}\{x_{i_1}, x_{i_2}\}$. It follows that w.l.o.g. $x_{i_1} = x_{j_1}$ and $x_{i_2} = x_{j_2}$, thus $N \in \mathcal{N}(S)$, which gives a contradiction. The statement follows with the same argument for the fourth condition in Definition 2.5.4. □

We now want to introduce an appropriate refinement routine preserving the regularity of a triangular prism grid. The routine can be derived as a direct consequence from the standard bisection method of a two-dimensional simplicial grid (see for instance [63]) with some additional steps concerning the third dimension.

Proposition 2.5.7 (Local refinement). *Let $T = (x_0, \dots, x_5) \in \mathcal{T}$ and let w.l.o.g. $E = \operatorname{conv}\{x_1, x_2\} \in \mathcal{E}_h(T)$ be the longest horizontal edge. Then T can be subdivided in xy-direction (also called xy-refined) along the edge E, called the refinement edge, into two subelements*

$$T_1^{xy} := (x_0, x_1, \frac{x_1 + x_2}{2}, x_3, x_4, \frac{x_4 + x_5}{2}),$$
$$T_2^{xy} := (x_0, x_2, \frac{x_1 + x_2}{2}, x_3, x_5, \frac{x_4 + x_5}{2}).$$

Furthermore, T can be subdivided in s-direction (also called s-refined) into two subelements

$$T_1^{s} := (x_0, x_1, x_2, \frac{x_0 + x_3}{2}, \frac{x_1 + x_4}{2}, \frac{x_2 + x_5}{2}),$$
$$T_2^{s} := (\frac{x_0 + x_3}{2}, \frac{x_1 + x_4}{2}, \frac{x_2 + x_5}{2}, x_3, x_4, x_5).$$

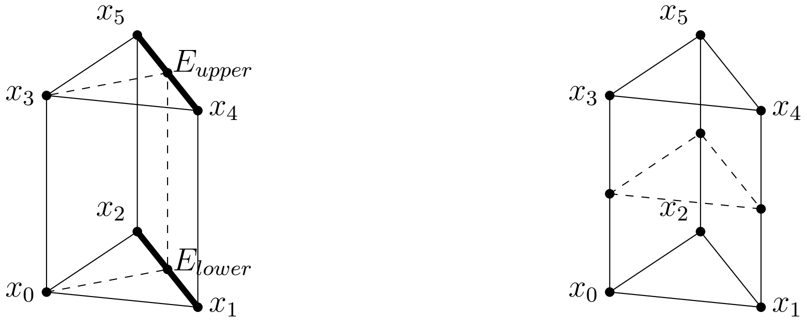


Figure 2.6.: Left: Subdivision of an element T in xy-direction into T_1^{xy} and T_2^{xy} (xy-refinement). The refinement edge E_{lower} (E_{upper} respectively) equals the longest horizontal edge. Right: Subdivision of an element T in s-direction into T_1^s and T_2^s (s-refinement).

In order to prevent degenerating prism elements, the refinement edge for xy-refinement is fixed as the longest horizontal edge of an element. Note that if T is subdivided along an edge $E = \operatorname{conv}\{x_1, x_2\} \in \mathcal{E}_h(T)$, T is also subdivided along the corresponding *upper* edge $\widetilde{E} = \operatorname{conv}\{x_4, x_5\}$. Thus, there exist actually two interrelated refinement edges $E = E_{\text{lower}}$ and $\widetilde{E} = E_{\text{upper}}$ with the same length (cf. Figure 2.6).
Before we introduce the proposed refinement routine, we need to give a definition of the neighbours of an element.

Definition 2.5.8 (Neighbours of an element). Let $\mathcal{T}$ be a regular triangular prism grid. An element $S \in \mathcal{T}$ is an *edge neighbour* of another element $t \in \mathcal{T}$ along an edge $E \in \mathcal{E}(T)$, if $E \in \mathcal{E}(S)$. S is called *face neighbour* of T along a face $F \in \mathcal{F}(T)$, if $F \in \mathcal{F}(T)$.

We now have the tools to introduce the proposed refinement routine. The routine refines a number of elements without violating the regularity conditions from Definition 2.5.4, which can be achieved by refinement of the corresponding neighbour elements.

Algorithm 1 Local refinement routine (xy)

```
function REFINE_XY($T$)
    Find longest edges $E_{\text{lower}}, E_{\text{upper}} \in \mathcal{E}_h(T)$
    Subdivide $T$ along edges $E_{\text{lower}}$ and $E_{\text{upper}}$
    for all neighbours $S \in \mathcal{T}$ along edge $E_{\text{lower}}$ and $E_{\text{upper}}$ do
        Find longest edges $\widetilde{E}_{\text{lower}}, \widetilde{E}_{\text{upper}} \in \mathcal{E}_h(S)$
        if $\widetilde{E}_{\text{lower}} = E_{\text{lower}}$ then
            REFINE_XY($S$)
        else
            REFINE_XY($S$)
            Find subelement $S_i$ of $S$ that contains $E_{\text{lower}}$ or $E_{\text{upper}}$
            REFINE_XY($S_i$)
        end if
    end for
end function
```

Algorithm 2 Local refinement routine (s) for a regular triangular prism grid

```
function REFINE_S($T$)
    Subdivide $T$ in s-direction
    for all neighbours $S \in \mathcal{T}$ sharing a vertical face with $T$ do
        REFINE_S($S$)
    end for
end function
```

Theorem 2.5.9. *Let $\mathcal{T}_0$ be a regular triangular prism grid, and $T \in \mathcal{T}_0$. Then REFINE_XY(T) yields the minimal regular triangular prism grid, where T is refined in xy-direction. The same holds for REFINE_S.*

Proof. The minimality follows directly from the definition of the refinement routine, since only those elements are refined, which share the refinement edge with T. Thus, it is sufficient to show that each pair of elements $T, S \in \mathcal{T}_1$ satisfies the regularity conditions. We distinguish between the following cases:

(1) $T, S \in \mathcal{T}_0$ (nothing to show).

(2) $T \in \mathcal{T}_0$, $S \in \mathcal{T}_1 \setminus \mathcal{T}_0$: There exists $\hat{S} \in \mathcal{T}_0$ such that $S \subset \hat{S}$, i.e. S is a child of $\hat{S}$. We assume w.l.o.g. that $T \cap S \neq \emptyset$, thus $T \cap \hat{S} \neq \emptyset$. For $T \cap \hat{S}$ we have the following possibilities:

(a) If $T \cap \hat{S} \in \mathcal{N}(T)$ and $T \cap \hat{S} \in \mathcal{N}(S)$, there is nothing to show.

(b) If $T \cap \hat{S} \in \mathcal{E}(T)$ and $T \cap \hat{S} \in \mathcal{E}(S)$, then T and $\hat{S}$ can share a vertical or a horizontal edge. If T and $\hat{S}$ share a vertical edge, so do T and S. If T and $\hat{S}$ share a horizontal edge, then this cannot be the refinement edge (since otherwise, T would have been refined), thus T and S share the same edge as T and $\hat{S}$.

(c) The case $T \cap \hat{S} \in \mathcal{F}(T)$ $T \cap \hat{S} \in \mathcal{F}(S)$ follows with the same argument.

(3) $T, S \in \mathcal{T}_1 \setminus \mathcal{T}_0$: There exist $\hat{T}, \hat{S} \in \mathcal{T}_0$ such that $T \subset \hat{T}$ and $S \subset \hat{S}$. The proof then follows with the argument by regarding all possible relations for $\hat{T} \cap \hat{S}$.

The result for REFINE_S can be obtained with a similar argumentation. □

Graphically, Algorithm 1 simply subdivides all neighbours of an element T which share the refinement edge E (or its lower or upper counterpart). Consequently, some elements are refined twice due to the fact that the shared edge is not necessarily the longest edge for all neighbouring elements. Note moreover that all elements which share a horizontal face with T stringently contain the lower or upper refinement edge and thus need to be refined subsequently.

While the xy-refinement satisfies the requirements to an adaptive refinement routine by only locally increasing the grid resolution, the s-refinement needs to be performed *globally* (meaning that every other element with the same s-coordinates needs to be refined) to preserve regularity. However, in some cases it might be more practical to relax the regularity conditions to be able to perform local s-refinement, accepting possible s-hanging nodes. To this end, we introduce the concept of a *semi-regular* triangular prism grid.

Definition 2.5.10 (Semi-regular triangular prism grid)**.** A *semi-regular triangular prism grid* $\mathcal{T}$ is a set of triangular prism elements such that $G \subseteq \bigcup\{T : T \in \mathcal{T}\}$ and for every pair of elements $T, S \in \mathcal{T}$ with $T = (x_0, \dots, x_5)$, $S = (y_0, \dots, y_5)$, one of the following relations holds:

- $T \cap S$ satisfies the regularity conditions
- $\exists\ i \in \{0, 1, 2\}$ s.t. $T \cap S = \operatorname{conv}\{x_i, \frac{x_i + x_{i+3}}{2}\} \in \mathcal{E}_v(S)$ or $T \cap S = \operatorname{conv}\{\frac{x_i + x_{i+3}}{2}, x_{i+3}\} \in \mathcal{E}_v(S)$ (or with exchanged x, y, respectively)
- $\exists\ i_1, i_2 \in \{0, 1, 2\}$, $i_1 \neq i_2$ s.t. $T \cap S = \operatorname{conv}\{x_{i_1}, \frac{x_{i_1} + x_{i_1+3}}{2}, x_{i_2}, \frac{x_{i_2} + x_{i_2+3}}{2}\} \in \mathcal{F}_v(S)$ or $T \cap S = \operatorname{conv}\{\frac{x_{i_1} + x_{i_1+3}}{2}, x_{i_1+3}, \frac{x_{i_2} + x_{i_2+3}}{2}, x_{i_2+3}\} \in \mathcal{F}_v(S)$ (or with exchanged x, y, respectively).

It is straightforward to check that every regular triangular prism grid is also semi-regular. The definition allows, in addition to the regular neighbouring relations, that the intersection of two elements consist of a *half-edge* or *half-face* of one element in vertical direction (see Figure 2.7). As a consequence, the occurrence of one s-hanging node per vertical edge is possible, while xy-hanging nodes are still permitted. Note that the limitation of s-hanging nodes to one per edge is a natural convention to prevent too many successive hanging nodes, which are typically not treated as degrees of freedom (see Section 2.5.2).

Lemma 2.5.11. *A semi-regular triangular prism grid $\mathcal{T}$ does not contain xy-hanging nodes.*

Proof. Let $T, S \in \mathcal{T}$ with $N \in \mathcal{N}(T)$, $N \in S$, $N \notin \mathcal{N}(S)$. In addition, assume that there exists $E \in \mathcal{E}_h(S)$ with $N \in E$, i.e. N is a xy-hanging node. From Lemma 2.5.6 we obtain that $T \cap S$ does not satisfy the regularity conditions, so one of the additional conditions has to hold.

(1) Assume that there exists $i \in \{0,1,2\}$ such that $T \cap S = \operatorname{conv}\{x_i, \frac{x_i+x_{i+3}}{2}\} \in \mathcal{E}_v(S)$. Then $x_i, \frac{x_i+x_{i+3}}{2} \in \mathcal{N}(S)$, and since $N \in T \cap S$ and $N \in \mathcal{N}(T)$, it follows that $N = x_i$, which gives a contradiction.

(2) Assume that there exists $j \in \{0,1,2\}$ such that $T \cap S = \operatorname{conv}\{y_j, \frac{y_j+y_{j+3}}{2}\} \in \mathcal{E}_v(T)$. Since $N \in \mathcal{N}(T)$ and $N \in T \cap S$, it follows $N = \frac{y_j+y_{j+3}}{2}$ ($N = y_j$ would be a contradiction to $N \notin \mathcal{N}(S)$). Consequently, there is no $E \in \mathcal{E}_h(S)$ with $N \in E$, which is a contradiction to the initial assumption.

All other semi-regularity conditions can be handled in a similar way. □

If an element is refined in s-direction, only neighbours which are *larger* in the sense of a their height (the difference between the s-coordinates of upper and lower nodes) need to be refined subsequently in order to prevent more than one hanging node per edge.

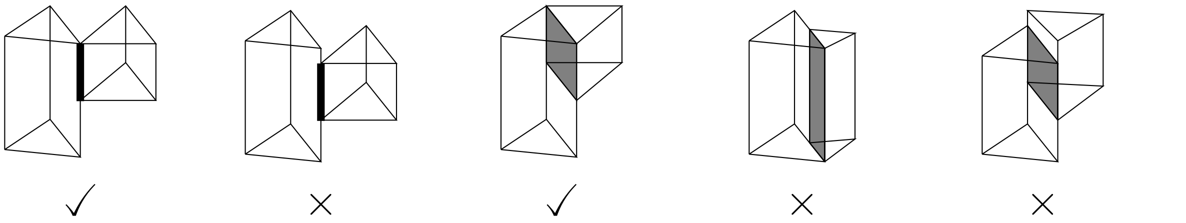


Figure 2.7.: Examples for allowed and permitted neighbouring relations in a semi-regular triangular prism grid.

As in case of a semi-regular triangular prism grid, the semi-regularity is preserved by the refinement routine, which is stated by the following result.

Theorem 2.5.12. *Let $\mathcal{T}_0$ be a semi-regular triangular prism grid and $T \in \mathcal{T}_0$. Then REFINE_XY(T) yields the minimal semi-regular triangular prism grid, where T is refined in xy-direction. The same holds for REFINE_S_{loc}(T).*

Proof. The proof follows the same strategy as for the regular case, therefore we will not provide any details here. □

Algorithm 3 Local refinement routine (s) for a semi-regular triangular prism grid

```
function REFINE_S$_{loc}(T)$
    Subdivide $T$ in s-direction
    for all neighbours $S \in \mathcal{T}$ with $h(S) > h(T)$ do
        REFINE_S($S$)
    end for
end function
```

Remark 2.5.13. The xy-refinement routine for a regular and a semi-regular grid is basically the same method. The only difference lies in the execution of the neighbour refinement. While in a regular grid, only one element is allowed to share a vertical face with the element to be refined, there can be two smaller elements (in the sense of their height) on top of each other, which both will be refined consecutively by REFINE_XY.

Finally, we note that the projection of a semi-regular triangular prism grid onto the xy-hyperplane H_{xy} naturally yields a two-dimensional simplicial grid by construction, so does every horizontal slice of the grid.

2.5.2. Finite element function spaces

In the previous section, we have introduced a suitable spatial discretization for the three-dimensional domain $G = [0,1]^3$. In order to obtain a finite element formulation of a functional lifting problem as defined in Definition 2.4.7, we can now think of a discretization of the underlying function spaces. A typical and practical choice on a two-dimensional simplicial grid is the space of piecewise linear functions, where the degrees of freedom correspond to the set of nodes of each simplex and thus the dimension of the discrete function space equals the number of grid nodes. We take up this idea and adapt it to the case of a triangular prism grid $\mathcal{T}$ by introducing the spaces

$$\begin{aligned} S^1(\mathcal{T}) := \{ w \in C(G) : \; & w|_T(\cdot,\cdot,s) \in \mathbb{P}^1(T_{xy}) \;\forall\; s \in T_s, \\ & w|_T(x,y,\cdot) \in \mathbb{P}^1(T_s) \;\forall\; (x,y) \in T_{xy} \;\forall\; T = T_{xy} \times T_s \in \mathcal{T} \}. \end{aligned}$$

$S^1(\mathcal{T})$ is the space of bilinear polynomials on each prism element $T \in \mathcal{T}$ with $\dim(S^1(\mathcal{T})) = |\mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})|$. Thus, the degrees of freedom in the finite element formulation correspond to the set of all non-hanging nodes of each prism element.

For some problems, it might be practical to further simplify the discrete function space by considering piecewise constant functions in the lifted dimension. To this end, we introduce the space

$$\begin{aligned} S^{1,0}(\mathcal{T}) := \{ w : G \to \mathbb{R} : \; & w|_T(\cdot,\cdot,s) \in \mathbb{P}^1(T_{xy}) \;\forall\; s \in T_s, \\ & w|_T(x,y,\cdot) \in \mathbb{P}^0(T_s) \;\forall\; (x,y) \in T_{xy} \;\forall\; T = T_{xy} \times T_s \in \mathcal{T}, \\ & w(\cdot,\cdot,s) \in C([0,1]^2) \;\forall\; s \in [0,1] \} \end{aligned}$$

of functions which are piecewise linear in (x, y) and piecewise constant in s.

Remark 2.5.14. In order to guarantee well-posedness of the definition of $S^{1,0}(\mathcal{T})$, one would have to define every $T \in \mathcal{T}$ as half-open in s-direction, otherwise every $w \in S^{1,0}(\mathcal{T})$ must automatically be constant along s. However, since any two-dimensional hyperplane is a null set with respect to the three-dimensional Lebesgue measure, we neglect this problem here and associate the lower three nodes x_0, x_1, x_2 with the local degrees of freedom of an element $T = (x_0, \dots, x_5)$ in case of functions defined on $S^{1,0}(\mathcal{T})$ (cf. Proposition 2.5.15).

The constancy assumption in the context of functional lifting problems basically has two reasons: On the one hand side, piecewise constant (and especially not overall continuous) functions in s-direction are a natural choice to represent the binary piecewise constant characteristic functions of the subgraph of an image. Although in our case, the basis functions are still continuous in xy-direction in order to allow for differentiability in the classical sense, the s-constancy makes the representation more natural and the lack of differentiability can be tackled by a finite difference interpretation. On the other hand side, any discretization should preferably reduce the number of constraints in the set $\mathcal{K}$ as defined in (2.8) to a finite count, namely those where the values s_1, s_2 correspond to the degrees of freedom of the system. For some particular cases, one can easily construct an example with piecewise linear basis functions which violates this desirable property (Remark 4.2.4), while piecewise constant functions still work (Theorem 4.2.3).
While for $S^1(\mathcal{T})$, the dimension equals the total number of non-hanging nodes in the grid $\mathcal{T}$, for $S^{1,0}(\mathcal{T})$ we have $\dim(S^{1,0}(\mathcal{T})) < |\mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T})|$. The following proposition fixes the degrees of freedom to equal the lower nodes of an element $T \in \mathcal{T}$.

Proposition 2.5.15 (Degrees of freedom for $S^{1,0}(\mathcal{T})$)**.** *The local degrees of freedom of an element $T = (x_0, \dots, x_5) \in \mathcal{T}$ for a triangular prism grid $\mathcal{T}$ correspond to the lower nodes x_0, x_1, x_2. Hence, the number of degrees of freedom (i.e. the dimension of $S^{1,0}(\mathcal{T})$) equals the number of lower nodes in $\mathcal{T}$ minus the number of hanging nodes. The set of degrees of freedom associated with a grid $\mathcal{T}$ is denoted as $\mathcal{D}(\mathcal{T})$. Note that $\mathcal{D}(\mathcal{T}) \subset (\mathcal{N}(\mathcal{T}) \setminus \mathcal{H}(\mathcal{T}))$.*

A simple set of basis functions for the space $S^{1,0}(\mathcal{T})$ can be constructed in a similar way as for the standard Lagrange finite element case: For each $N_i \in \mathcal{D}(\mathcal{T})$, we define a function $\psi_i \in S^{1,0}(\mathcal{T})$ as

$$\psi_i(P) := \begin{cases} 1 & \text{if } P = N_i, \\ 0 & \text{otherwise} \end{cases}$$

for all $P \in \mathcal{D}(\mathcal{T})$. Then, $(\psi_1, \dots, \psi_q)$ for $q = |\mathcal{D}(\mathcal{T})|$ being the total number of degrees of freedom is obviously a basis of the space $S^{1,0}(\mathcal{T})$ and $\dim(S^{1,0}(\mathcal{T})) = q$.
The concept of a (semi-)regular triangular prism grid $\mathcal{T}$ in combination with $S^{1,0}(\mathcal{T})$ functions admits another useful property in the context of functional lifting problems. Any numerical treatment of the underlying saddle point problem (2.4.7) involves some kind of constraint handling concerning the set $\mathcal{K}$ (cf. (2.8)), consisting of a set of integral inequalities for every point $x \in [0, 1]^2$. In the absence of xy-hanging nodes, the sets for

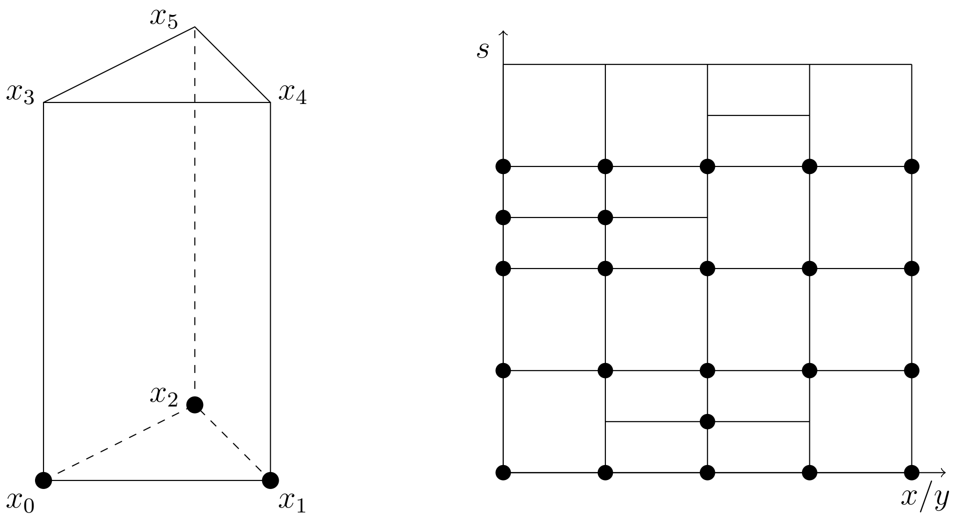


Figure 2.8.: Left: Local degrees of freedom (marked by •) for an element T in the space $S^{1,0}(\mathcal{T})$. Right: Global degrees of freedom for a two-dimensional grid assuming piecewise constant basis functions in s-direction.

different x can be treated independently, which simplifies a constraint projection and increases its efficiency. More precisely, we fix the following definition.

Definition 2.5.16 (Ground node, s-line)**.** A node $N = (N_1, N_2, N_3) \in \mathcal{N}(\mathcal{T})$ for a triangular prism grid $\mathcal{T}$ is called *ground node*, if $N_3 = 0$. For a ground node $N \in \mathcal{N}(\mathcal{T})$, define

$$L_N := \{P = (P_1, P_2, P_3) \in \mathcal{N}(\mathcal{T}) \ : \ P_1 = N_1, \ P_2 = N_2\}.$$

L_N is denoted as *s-line.*

Remark 2.5.17. One can easily show that for every node $P \in \mathcal{N}(\mathcal{T})$ in a semi-regular grid $\mathcal{T}$, there exists a ground node N such that $P \in L_N$, thus the set of all s-lines contains every node in the grid.

The main property of a semi-regular grid concerning s-lines is their independence in the sense of horizontal interpolation. Since L_N does not contain any xy-hanging node, changing the values of nodes (including hanging nodes) within one s-line does not directly affect the neighbouring s-lines. For some special examples of constraint sets $\mathcal{K}$ (as those regarded in the later course of this work, see Chapter 4), this property is essential for an improved efficiency of the projection onto $\mathcal{K}$ (cf. Section 4.4).

3 Review of methods for transportation network problems

In order to understand the tasks and challenges concerning a numerical treatment of transportation problems, in this chapter we would like to introduce the concepts of optimal transport and optimal network problems. Starting with the original famous model formulations by Monge and Kantorovich, we lead over to the fundamental definitions of branched transport, urban planning and some related versions, concluding with a brief summary of the existing numerical approaches.

3.1. Optimal transport and transport networks - A brief overview

Optimal transport or optimal transportation problems arise naturally in several fields of application. What is the most effective way to transport a pile of sand into a hole in the ground? Which path should a travelling merchant follow to deliver his goods? How to establish a public transportation network connecting people's homes with their workplaces? Across all different formulations and practical backgrounds, the main question asked in this context is: How to move mass from some initial distribution to a desired final distribution such that the costs (associated with a certain cost functional) become as low as possible? Here, different choices of the cost functional accompanied by different solution spaces lead to various optimization problems, comprising all prior knowledge about the structure of the desired solution.

The history of mathematical formulations that nowadays are referred to as "optimal transport problems" started in 1781 with a famous paper by Monge [48], [65]. He was motivated by the task of transporting a certain amount of soil to a specified area for construction purposes. The problem can be reformulated in an economic fashion: Given a number of bakeries in the area of Paris that produce a certain amount of bread every day, assume that the bread has to be transported to a number of cafés. If the exact number of bread consumed in each café is known, how to determine which bread unit should go to which place such that the transport costs become minimal? In a general framework, Monge formulated the problem as follows [66], [60].

Definition 3.1.1 (Monge's problem)**.** Let $\Omega \subset \mathbb{R}^n$, $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ be two non-negative finite Borel measures on Ω, $c : \Omega \times \Omega \to [0, \infty]$ a cost functional and $T : \Omega \to \Omega$ a transport map. Then the Monge problem is defined as

$$\inf_T \int_\Omega c(x, T(x)) \mathrm{d}\mu_+(x) \quad \text{s.t.} \quad T\#\mu_+ = \mu_-.$$

In the bakery setting, Ω represents the area of Paris, μ_+ and μ_- the distribution of bakeries and cafés, respectively. The functional $c(x, y)$ defines the cost of transporting one unit of bread from a bakery at position x to a café at position y. A typical choice of the cost function is $c(x, y) = |x - y|^p$ for some $p \geq 0$, which simply penalizes the distance between the two locations.

Several questions concerning Monge's problem remained unsolved for a long time. It is unclear if a solution exists a priori. Furthermore, a transport map T is not able to model all conceivable situations, in particular, mass cannot be split up. In other words, it is impossible to describe the transport of bread from one bakery to two different cafés (which, in a more general setting, would be desirable). This problem was addressed by Kantorovich in 1942 [39]. He changed the point of view by looking for a transport *plan* on the space $\Omega \times \Omega$ instead of a transport *map* T [60].

Definition 3.1.2 (Kantorovich problem)**.** Define the set of transport plans between μ_+ and μ_- as

$$\Pi(\mu_+, \mu_-) := \{\gamma \in \mathcal{M}_+(\Omega \times \Omega) : \ \pi_0\#\gamma = \mu_+, \ \pi_1\#\gamma = \mu_-\}$$

where $\pi_0, \pi_1 : \Omega \times \Omega \to \Omega$ denote the projection onto the first and, respectively, second component. Then the Kantorovich problem is defined as

$$\inf_{\gamma \in \Pi(\mu_+, \mu_-)} \int_{\Omega \times \Omega} c(x, y) \mathrm{d}\gamma(x, y).$$

In contrast to the Monge problem, the Kantorovich formulation does not ask for a map specifying which particle goes where, but gives a plan γ such that $\gamma(x, y)$ indicates how much mass is transported from x to y. As a consequence, mass can be split up easily and the model is able to display more general mass rearrangements.

Although the Kantorovich model is more flexible than the one introduced by Monge, it still cannot handle transportation problems where branching structures play an important role. The transport map $c(x, y)$ only encodes the travelling distance between two points x and y, thus the Kantorovich cost functional is linear in the amount of transported mass. Consider the situation where a single Dirac measure with mass 1 is transported to two Dirac measures with mass $\frac{1}{2}$ and define the function $c(x, y)$ as the distance of two points x, y along a one-dimensional network Σ (for $x, y \in \Sigma$). Then the Kantorovich costs related to a "V-shaped" network would be cheaper than those related to a "Y-shaped" network, although in many cases, the latter would be desirable. In order to address this problem, researchers introduced the so-called *Branched Transport problem* in 2003: Xia [68] developed a model based on transport of atomic measures (which can also approximate more general measures) via weighted directed graphs. At the same time, a similar model was introduced by Maddalena, Solimini and Morel [42], which follows a Lagrangian formulation where paths are represented by trajectories of particles. Since these models form a part of the basis of this work, we will present both in detail in Section 3.2.

Another model belonging to the family of optimal transportation network problems was first studied in 2005 by Brancolini and Buttazzo [18]. Here, the optimal path has the interpretation of a public transportation network: People can either choose to travel on the network or, for a higher cost, on their own expense outside of the network. This leads to a cost functional whose general structure is comparable to the branched transport case, but where the cost functional takes into account the costs associated with the network as well as the costs of passengers travelling on their own. We will outline the details of this model in Section 3.3.

The main focus of this work lies on the numerical treatment of the branched transport and urban planning problem. Nevertheless, note that both models can be regarded as the most common representatives of a larger family of problems, namely those which aim at minimizing a cost functional which is sublinear in the amount of transported mass. There exist other examples belonging to this general class of problems, for example the classical Steiner tree problem [56], where only the length of the graph is penalized. This can either be seen as a limit case of the branched transport problem (see Section 3.2) or as a special variant of the urban planning model (see Section 3.3). Furthermore, one can easily extend the urban planning costs for transported mass by a more general minimum of piecewise affine functions.

3.2. Branched transport

In this section, we present two different formulations of the branched transport problem. Like in the mathematical characterization of a flow field, the two variants can be regarded as a flux-based *Eulerian* and a particle-based *Lagrangian* formulation [20]. Note that there exist even more ways to describe the problem (also in case of urban planning), such as a formulation via flat 1-chains (see [21]). Throughout the rest of this chapter, we consider a region $\Omega \subset \mathbb{R}^n$ and two measures $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$.

3.2.1. Eulerian formulation

The Eulerian formulation from [68] starts with the definition of a mass flux and the associated cost functional between discrete finite measures. Hence, let μ_+, μ_- be given by

$$\mu_+ = \sum_{i=1}^{k} a_i \delta_{x_i}, \ \mu_- = \sum_{j=1}^{l} b_j \delta_{y_j}$$

for a given set of pairwise distinct points $x_i, y_j \in \mathbb{R}^n$, $a_i, b_j \geq 0$, $i = 1, \ldots, k, j = 1, \ldots, l$, where μ_+ and μ_- have the same mass,

$$\sum_{i=1}^{k} a_i = \sum_{j=1}^{l} b_j.$$

A discrete mass flux between μ_+ and μ_- can be defined as follows.

Definition 3.2.1 (Discrete mass flux between μ_+ and μ_-)**.** A *discrete mass flux* between μ_+ and μ_- is a weighted directed graph G, consisting of a set of vertices $V(G)$, a set of edges $E(G)$ and a weight function $w : E(G) \to [0, \infty)$, satisfying the following mass preserving conditions:

- $a_i = \sum_{e \in E(G), e^- = x_i} w(e) - \sum_{e \in E(G), e^+ = x_i} w(e)$ for $i = 1, \ldots k$
- $b_j = \sum_{e \in E(G), e^+ = y_j} w(e) - \sum_{e \in E(G), e^- = y_j} w(e)$ for $j = 1, \ldots l$
- $0 = \sum_{e \in E(G), e^+ = v} w(e) - \sum_{e \in E(G), e^- = v} w(e)$ for every $v \in V(G) \backslash \{x_1, \ldots, x_k, y_1, \ldots, y_l\}$.

Here, e^- and e^+ denote the initial and final point of the edge $e \in E(G)$.

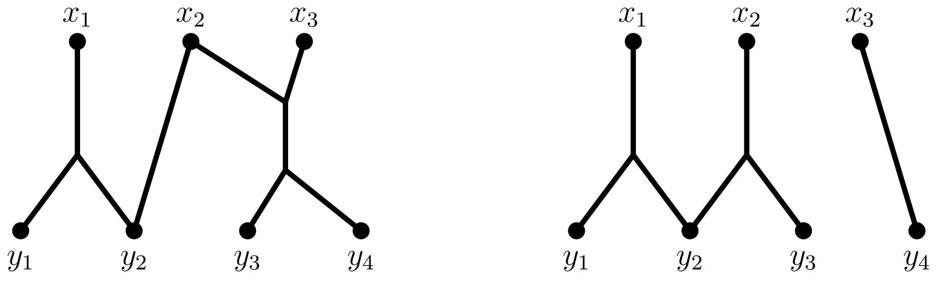


Figure 3.1.: Two possible transport paths between discrete measures $\mu_+ = \sum_{i=1}^{3} a_i \delta_{x_i}$ and $\mu_- = \sum_{j=1}^{4} b_j \delta_{y_j}$.

In other words, the transport path between the two measure can be identified with a set of edges and vertices, where $w(e)$ indicates the amount of mass flowing through an edge $e \in$

$E(G)$ and no mass is created or lost at any interior vertex $v \in V(G)\backslash\{x_1, \dots, x_k, y_1, \dots, y_l\}$. The cost of such a graph associated to branched transport is given by the following definition.

Definition 3.2.2 (Discrete branched transport cost)**.** Given a discrete mass flux G between two discrete measures μ_+ and μ_-, the *discrete branched transport cost* for a given branching parameter $\alpha \in (0, 1)$ is given by

$$\mathcal{M}^\alpha(G) = \sum_{e \in E(G)} w(e)^\alpha l(e),$$

where $l(e)$ denotes the length (associated with the one-dimensional Hausdorff measure $\mathcal{H}^1 \llcorner e$) of the edge e.

The parameter α is responsible for controlling the grade of ramification or branching. The branched transport cost functional directly ties on the fact that a “Y-shaped” or a “V-shaped” path from one Dirac measure to two Dirac measures with half of the mass have the same costs related to the Kantorovich model. As opposed to this, one can easily verify that for a parameter $\alpha = \frac{1}{2}$, the “Y-shaped” graph is preferred by the branched transport cost functional, as shown by the following example.

Example 3.2.3. Let $\Omega \subset \mathbb{R}^2$ and $x_1, y_1, y_2 \in \Omega$ three points with equal distance $d > 0$, $\alpha = \frac{1}{2}$ and $\mu_+ = \delta_{x_1}$, $\mu_- = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$. Let $v = \frac{1}{3}(x_1 + y_1 + y_2)$ be the midpoint of the triangle with vertices x_1, y_1, y_2. We define two graphs G_1 and G_2 by

- $V(G_1) = \{x_1, y_1, y_2\}$, $E(G_1) = \{e_{x_1,y_1}, e_{x_1,y_2}\}$, $w(e_{x_1,y_1}) = w(e_{x_1,y_2}) = \frac{1}{2}$
- $V(G_2) = \{x_1, y_1, y_2, v\}$, $E(G_2) = \{e_{x_1,v}, e_{v,y_1}, e_{v,y_2}\}$, $w(e_{x_1,v}) = 1$, $w(e_{v,y_1}) = w(e_{v,y_2}) = \frac{1}{2}$

where $e_{x,y}$ denotes the edge with starting point x and ending point y. G_1 corresponds to a “V-shaped” and G_2 to a “Y-shaped” graph with a branching point located at the arithmetic mean of the three points (cf. Figure 3.2). Then the costs of the two graphs can be computed as

- $\mathcal{M}^\alpha(G_1) = 2d(\frac{1}{2})^\alpha = \sqrt{2}d$
- $\mathcal{M}^\alpha(G_2) = \frac{\sqrt{3}}{3}d1^\alpha + 2\frac{\sqrt{3}}{3}d(\frac{1}{2})^\alpha = \frac{\sqrt{3}}{3}d(1 + \sqrt{2})$.

One can easily verify that $\mathcal{M}^\alpha(G_2) < \mathcal{M}^\alpha(G_1)$ for the particular choice of α, although the point v was chosen manually and G_2 is not necessarily the minimizer of $\mathcal{M}^\alpha$.

In order to proceed to a more general formulation of the branched transport cost for a general class of measures, let us replace the graph by a vector-valued measure indicating the mass flux.

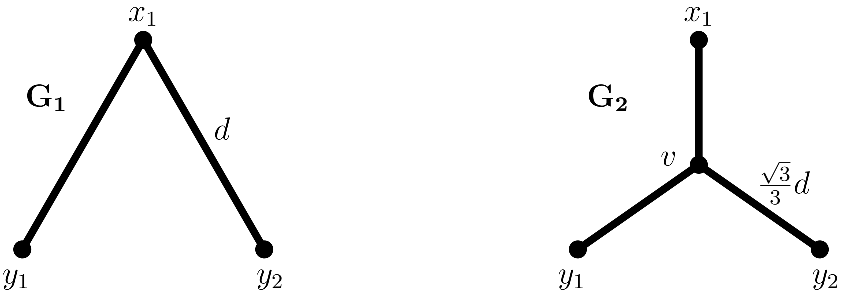


Figure 3.2.: Two graphs representing a possible transport path between $\mu_+ = \delta_{x_1}$, $\mu_- = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$.

Definition 3.2.4 (Mass flux associated with graphs)**.** Let G be a weighted directed graph with a set of vertices $V(G)$ and a set of edges $E(G)$ and let $\hat{e} = \frac{e^+ - e^-}{|e^+ - e^-|}$ denote the direction of the edge $e \in E(G)$. Then the *mass flux associated with the graph G* is a vector-valued measure

$$\mathcal{F}_G = \sum_{e \in E(G)} w(e)(\mathcal{H}^1 \llcorner e)\hat{e}.$$

$\mathcal{F}_G$ is a mass flux between μ_+ and μ_-, if $\mathrm{div}\mathcal{F}_G = \mu_+ - \mu_-$ (in the distributional sense).

Here, all mass-preserving conditions from Definition 3.2.1 summarize to $\mathrm{div}\mathcal{F}_G = \mu_+ - \mu_-$. Using this representation of a mass flux, we can formulate the cost functional in the case of μ_+, μ_- being general (non-discrete) finite Borel measures.

Definition 3.2.5 (Continuous mass flux between μ_+ and μ_-)**.** Let $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ be two non-negative finite Borel measures of equal mass. A vector measure $\mathcal{F} \in \mathcal{M}(\Omega, \mathbb{R}^n)$ is a *mass flux between μ_+ and μ_-*, if there exist two sequences of discrete measures $(\mu_+^k), (\mu_-^k) \subset \mathcal{M}_+(\Omega)$ with

$$\mu_+^k \rightharpoonup^* \mu_+, \ \mu_-^k \rightharpoonup^* \mu_-,$$

and a sequence of discrete mass fluxes $\mathcal{F}_{G_k}$ between μ_+^k and μ_-^k with $\mathrm{div}\mathcal{F}_{G_k} = \mu_+^k - \mu_-^k$ and

$$\mathcal{F}_{G_k} \rightharpoonup^* \mathcal{F}.$$

One can easily verify that indeed, $\mathcal{F}$ satisfies the mass-preserving condition $\mathrm{div}\mathcal{F} = \mu_+ - \mu_-$, following by continuity with respect to the weak-* topology. As a consequence, the cost functional of a mass flux between general measures is defined as follows, finally leading to the definition of the general *branched transport problem.*

Definition 3.2.6 (Continuous branched transport cost)**.** Let $\mathcal{F}$ be a mass flux between two measures $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ with equal mass. Then the *continuous branched transport cost functional* is defined as

$$\mathcal{M}^\alpha(\mathcal{F}) = \inf\Big\{ \liminf_{k \to \infty} \mathcal{M}^\alpha(G_k) : \ (\mu_+^k, \mu_-^k, \mathcal{F}_{G_k}) \rightharpoonup^* (\mu_+, \mu_-, \mathcal{F}) \Big\}$$

with μ_+^k, μ_-^k, G_k as in Definition 3.2.5.

Definition 3.2.7 (Branched transport problem)**.** Let $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ be two measures with equal mass, and set

$$\mathcal{M}^{\alpha,\mu_+,\mu_-}(\mathcal{F}) = \begin{cases} \mathcal{M}^\alpha(\mathcal{F}) & \text{if } \mathrm{div}\mathcal{F} = \mu_+ - \mu_-, \\ \infty & \text{otherwise.} \end{cases}$$

Then the *branched transport problem* is defined as

$$\inf_{\mathcal{F}} \mathcal{M}^{\alpha,\mu_+,\mu_-}(\mathcal{F}).$$

It can be shown that a minimizer of the branched transport cost functional exists [68]. Furthermore, if a mass flux $\mathcal{F}$ has finite costs, it can be written as

$$\mathcal{F} = \widetilde{F}\hat{e}(\mathcal{H}^1 \llcorner \Sigma),$$

where $\Sigma \subset \Omega$ is a rectifiable one-dimensional set, $\widetilde{F} : \Sigma \to [0, \infty)$ a weight function encoding the amount of mass flowing through Σ and $\hat{e} : \Sigma \to \mathbb{R}^n$ with $|\hat{e}| = 1$ the orientation of the network (see for instance [68]). Then the continuous cost functional reads

$$\mathcal{M}^\alpha(\mathcal{F}) = \int_\Sigma \widetilde{F}^\alpha \mathrm{d}\mathcal{H}^1, \tag{3.1}$$

which gives a natural continuous extension to the discrete cost functional as in Definition 3.2.2.

Remark 3.2.8. The term “branched transport” sometimes refers to a more general class of problems in the literature. Instead of fixing the cost per transported mass as $c(m) = m^\alpha$, the branched transport problem is often defined via any cost functional c, where c is a continuous concave function with $c(0) = 0$, leading to a branching structure of the optimal network. Note that the chosen formulation with a specific $c(m) = m^\alpha$ is still practical, since it is a common representative of the more general formulation and also covers some well-known transportation problems as its limit cases. For $\alpha = 1$, the problem becomes linear in the transported mass and therefore equivalent to the Kantorovich problem as in Definition 3.1.2 for a choice of the Kantorovich cost function $c(x, y) = |x - y|$. The problem of minimizing (3.1) for $\alpha = 1$ is also known as *Beckmann's problem* [60]. On the other hand, if $\alpha = 0$, only the length of the network is penalized, with no regard for the amount of transported mass. This problem is referred to as the *Steiner tree problem.*

Remark 3.2.9. The branched transport cost functional in general is highly non-convex. Although a global minimum is known to exist, the minimizer does not even have to be unique (as in case of transportation between two measures $\mu_+ = \frac{1}{2}(\delta_{x_1} + \delta_{x_2})$ and $\mu_- = \frac{1}{2}(\delta_{y_1} + \delta_{y_2})$ for some specific choice of the parameter α, as shown in Section 4.2.1) and any approach to computationally obtain a solution might easily get stuck in local minima. Additionally, the optimal network can involve quite complicated ramification structures. Hence, the numerical treatment of this type of problems requires some careful considerations, which will form the main focus of this work.

3.2.2. Lagrangian formulation

The branched transport problem has been derived from a different point of view in [42] and in a slightly more general setting in [11]. Instead of describing the network via a graph or a corresponding vector measure, this formulation considers trajectories of particles as so-called *irrigation patterns*. The two corresponding cost functionals were shown to be equivalent in [12]. Since we will further focus on the formulation by Xia given in the previous section, we will only briefly describe the idea of the Lagrangian Ansatz at this point.

In the Lagrangian formulation, the network is represented by a measurable function indicating the position of each particle that has to be transported from a measure μ_+ to μ_- at each time point. This function is called an *irrigation pattern*.

Definition 3.2.10 (Irrigation pattern)**.** Let Γ be a separable uncountable metric space together with the Borel σ-algebra $\mathcal{B}(\Gamma)$ and a positive finite Borel measure $P \in \mathcal{M}_+(\Gamma)$ (containing no atoms) and let $I = [0,1]$. Then, a measurable function $\chi : \Gamma \times I \to \mathbb{R}^n$ with $\chi_p = \chi(p,\cdot) : I \to \mathbb{R}^n$ absolutely continuous on I for almost every $p \in \Gamma$ is called an *irrigation pattern*.

The space Γ represents the set of transported particles and χ_p can be seen as the trajectory of the particle p as a function in time. The Borel measure P indicates the amount of transported mass, which becomes clear with the following definition, leading to a formulation of the branched transport cost functional with respect to irrigation patterns, following the notation of [20].

Definition 3.2.11 (Branched transport cost (Lagrangian formulation))**.** Let χ be an irrigation pattern. For any point $x \in \mathbb{R}^n$ we define

$$[x]_\chi := \{q \in \Gamma :\ x \in \chi_q(I)\}$$

as the set of all particles flowing through x. The total mass of all those particles is denoted by $m_\chi(x) = P([x]_\chi)$. Let $\alpha \in (0,1)$. Then the *Lagrangian formulation of the branched transport cost functional* is defined as

$$\mathcal{M}^\alpha(\chi) = \int_{\Gamma \times I} s_\alpha^\chi(\chi_p(t))|\dot{\chi}_p(t)|\mathrm{d}P(p)\mathrm{d}t,$$

where $s_\alpha^\chi(x) = [m_\chi(x)]^{\alpha-1}$ and $s_\alpha^\chi(x) = \infty$ if $m_\chi(x) = 0$.

The branched transport problem in its Lagrangian setting is then defined as follows.

Definition 3.2.12 (Branched transport problem (Lagrangian formulation))**.** Let $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ and χ be an irrigation pattern. Define $i_0, i_1 : \Gamma \to \mathbb{R}^n$ as $i_0(p) = \chi_p(0)$ and $i_1(p) = \chi_p(1)$. Then the *Lagrangian formulation of the branched transport problem* is defined as

$$\inf\Big\{\mathcal{M}^\alpha(\chi) : \; i_0\#P = \mu_+, \; i_1\#P = \mu_-\Big\}.$$

The existence of a solution of this problem was shown in [41], while a proof of the equivalence to the previously introduced Eulerian formulation as stated in the following result was provided by [12].

Theorem 3.2.13 (Equivalence of the minimization problems)**.** *Set*

$$\mathcal{M}^{\alpha,\mu_+,\mu_-}(\chi) = \begin{cases} \mathcal{M}^\alpha(\chi) & \textit{if } i_0\#P = \mu_+, \; i_1\#P = \mu_-, \\ \infty & \textit{otherwise.} \end{cases} \tag{3.2}$$

Then the Eulerian and Lagrangian minimization problems are equivalent in the sense that

$$\underset{\mathcal{F}}{\mathit{min}}\, \mathcal{M}^{\alpha,\mu_+,\mu_-}(\mathcal{F}) = \underset{\chi}{\mathit{min}}\, \mathcal{M}^{\alpha,\mu_+,\mu_-}(\chi).$$

Proof. The proof has been given in [12]. □

3.3. Urban planning

In this section, we want to provide more details about the urban planning problem, which was first introduced in a rather general setting in [18]. Similar to the branched transport case, the urban planning problem aims at finding an optimal path transporting mass from μ_+ to μ_-. The major difference lies in the fact that the urban planning cost is divided into two components: Particles are allowed to travel along a network as well as outside of the network at higher costs. As a consequence, the minimizer has the common interpretation as a public transportation network in a city area, where people can choose to travel on their own or on the network to reach their destination. We will start with a general introduction to the urban planning problem, which was first formulated via a Wasserstein distance [18]. We will discuss this formulation in more detail in Section 3.3.1 and afterwards show that the functional admits a flux-based (Section 3.3.2) and a pattern-based formulation (Section 3.3.3) similar to the branched transport problem [20].

3.3.1. Wasserstein formulation

Let $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ be two non-negative finite Borel measures representing the population density and the workplace density, respectively, and let the transport network be given by a one-dimensional set $\Sigma \subset \Omega$ with finite length. Now each person can choose to travel outside of the network Σ at a cost $a > 0$ per travelling distance, or on the network at a cost $b > 0$ with $b < a$, such that the latter is cheaper. If a person's travelling path is denoted by a curve $\gamma : [0,1] \to \mathbb{R}^n$, the costs of γ consist of the two parts

$$a\mathcal{H}^1(\gamma \setminus \Sigma) + b\mathcal{H}^1(\gamma \cap \Sigma).$$

This definition induces a metric (associated with a given transport network Σ) describing the minimal transportation costs to travel from a point $x \in \Omega$ to $y \in \Omega$ via

$$d_\Sigma(x,y) = \inf\{a\mathcal{H}^1(\gamma \setminus \Sigma) + b\mathcal{H}^1(\gamma \cap \Sigma) \;:\; \gamma \in C_{x,y}\},$$

where

$$C_{x,y} = \Big\{\gamma : [0,1] \to \mathbb{R}^n \;:\; \gamma \text{ is Lipschitz, } \gamma(0) = x,\ \gamma(1) = y\Big\}$$

represents the set of admissible travelling paths. In order to reduce the number of parameters, one typically chooses $b = 1$ and introduces the additional requirement $a > 1$. The metric d_Σ induces a Wasserstein distance between the measures μ_+ and μ_-,

$$W_{d_\Sigma}(\mu_+, \mu_-) = \inf_{\mu \in \Pi(\mu_+,\mu_-)} \int_{\mathbb{R}^n \times \mathbb{R}^n} d_\Sigma(x,y)\mathrm{d}\mu(x,y),$$

where the infimum is taken over all transport plans connecting μ_+ and μ_- given by

$$\Pi(\mu_+, \mu_-) = \Big\{\mu \in \mathcal{M}_+(\Omega \times \Omega) \;:\; \pi_0\#\mu = \mu_+,\ \pi_1\#\mu = \mu_-\Big\},$$

where $\pi_i : \Omega \times \Omega \to \Omega$ denotes the projection onto the i-th component. We now define the *urban planning problem* in its formulation given by [18], which involves the Wasserstein distance induced by the metric d_Σ as well as an additional penalization of the total network length, associated with some maintenance cost $\varepsilon > 0$.

Definition 3.3.1 (Urban planning cost (Wasserstein formulation))**.** Let $\Sigma \subset \Omega$ be a one-dimensional subset, $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ be two measures with equal mass. Then the *urban planning cost functional* is defined as

$$\mathcal{E}^{a,\varepsilon}(\Sigma) = W_{d_\Sigma}(\mu_+, \mu_-) + \varepsilon\mathcal{H}^1(\Sigma).$$

Definition 3.3.2 (Urban planning problem (Wasserstein formulation))**.** Let $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ be two measures with equal mass, the *urban planning problem* is defined as

$$\inf\Big\{\mathcal{E}^{a,\varepsilon}(\Sigma) \;: \Sigma \text{ admissible network }\Big\}.$$

Σ is called *admissible network*, if $\Sigma \subset \Omega$ is a closed one-dimensional subset with finite length.

Remark 3.3.3. Definition 3.3.2 does not require an admissible network Σ to be *connected.*

Two proofs of existence of a minimizer have been stated in [22], where only one requires connectedness of the network.

Similar to the case of branched transport, the urban planning cost can be reformulated in an Eulerian and a Lagrangian fashion equivalent to the previous definition. In order to obtain a unified setting for this work, we will again focus on the flux-based Eulerian framework in Section 3.3.2, only briefly describing the idea of the pattern-based approach (see Section 3.3.3).

3.3.2. Eulerian formulation

As for the branched transport model, we start with μ_+, μ_- being discrete measures and formulate the discrete urban planning problem with respect to graphs, following the course of [20]. Let G be a discrete mass flux between μ_+ and μ_- (cf. Definition 3.2.1) and $\Sigma \subset G$ a subgraph. Since travelling outside of the network Σ is allowed, the discrete urban planning cost functional is defined with respect to both G and Σ, with $G \setminus \Sigma$ representing the travelling path outside of the network.

Definition 3.3.4 (Discrete urban planning cost (1))**.** Given a discrete mass flux G between two discrete measures μ_+, μ_- and a subgraph $\Sigma \subset G$, the *discrete urban planning cost* for given parameters $a > 1$ and $\varepsilon > 0$ is given by

$$\mathcal{E}^{a,\varepsilon}(G, \Sigma) = \sum_{e \in E(G) \setminus E(\Sigma)} aw(e)l(e) + \sum_{e \in E(\Sigma)} (w(e) + \varepsilon)l(e),$$

where $l(e)$ denotes the length (associated with the one-dimensional Hausdorff measure $\mathcal{H}^1 \llcorner e$) of the edge e.

One can easily express the above cost functional with respect to the discrete mass flux G only. Since a is associated with the cost for travelling outside of the network (on $G \setminus \Sigma$) and $b = 1$ on Σ, any optimal pair (G, Σ) must satisfy

$$\begin{aligned} aw(e) &\leq w(e) + \varepsilon \text{ if } e \in E(G) \setminus E(\Sigma), \\ aw(e) &\geq w(e) + \varepsilon \text{ if } e \in E(\Sigma), \end{aligned}$$

since otherwise one can find another pair which has lower costs than (G, Σ). Consequently, the cost for an edge $e \in E(G)$ summarizes to $c^{a,\varepsilon}(w(e)) = \min\{aw(e), w(e) + \varepsilon\}$, leading to a reduced definition of the discrete urban planning cost.

Definition 3.3.5 (Discrete urban planning cost (2))**.** For G, a, ε as before, the *discrete urban planning cost* with respect to G is given by

$$\mathcal{E}^{a,\varepsilon}(G) = \sum_{e \in E(G)} \min\{aw(e), w(e) + \varepsilon\} l(e).$$

Note that for a finite optimal mass flux G, one can recover the edges belonging to the public transportation network Σ as the set of edges whose cost is associated with $w(e)+\varepsilon$, that is

$$E(\Sigma)=\Big\{e\in E(G)\ :\ aw(e)>w(e)+\varepsilon\Big\}.$$

Furthermore, the function $c^{a,\varepsilon}$ is subadditive in the transported mass $w(e)$, similar to the function $c(m)=m^{\alpha}$ in the branched transport case. Thus the urban planning model promotes branching structures in the optimal network as well, though in a slightly weaker form.

Example 3.3.6. The relation between the parameters a and ε determines the branching structure as well as the range of the corresponding network. Let $\Omega\subset\mathbb{R}^2$ and $x_1,x_2,y_1,y_2\in\Omega$ be four points defined as the vertices of a rectangle with side lengths 1 and d, i.e.

$$x_1=(0,d),\ x_2=(1,d),\ y_1=(0,0),\ y_2=(1,0),$$

and let $\mu_+=\frac{1}{2}\delta_{x_1}+\frac{1}{2}\delta_{x_2}$ and $\mu_-=\frac{1}{2}\delta_{y_1}+\frac{1}{2}\delta_{y_2}$. We fix $a=2$ and compare the optimal network structure for different values of ε and d. For every edge $e\in E(G)$ of a minimizing graph G we have $w(e)\in\{\frac{1}{2},1\}$. Hence, G can either transport each particle independently (case A) or gather all the mass by admitting two branching points (case B). The transportation network Σ can adopt the following configurations, depending on the general structure of the minimizer G:

- Case A: $\Sigma=\emptyset$ (case A_1) or $\Sigma=G$ (case A_2),
- Case B: $\Sigma\subsetneq G$ (case B_1) or $\Sigma=G$ (case B_2).

Figure 3.3 shows the possible network structures depending on the choice of ε and d for fixed $a=2$.

In order to obtain a definition for general (non-discrete) measures μ_+,μ_-, we follow the same pathway as in the branched transport case, defining a mass flux as a vector-valued measure $\mathcal{F}$ associated with a graph G (cf. Definition 3.2.4) and approximating a finite Borel measure by a sequence of discrete measures. For the sake of completeness, the corresponding definitions are given in the following.

Definition 3.3.7 (Continuous urban planning cost)**.** Let $\mathcal{F}$ be a mass flux between two finite Borel measures $\mu_+,\mu_-\in\mathcal{M}_+(\Omega)$ with equal mass. Then the *continuous urban planning cost functional* is defined as

$$\mathcal{E}^{a,\varepsilon}(\mathcal{F})=\inf\Big\{\liminf_{k\to\infty}\mathcal{E}^{a,\varepsilon}(G_k)\ :(\mu_+^k,\mu_-^k,\mathcal{F}_{G_k})\rightharpoonup^*(\mu_+,\mu_-,\mathcal{F})\Big\}.$$

with $\mu_+^k,\mu_-^k,G_k,\mathcal{F}_{G_k}$ as in Definition 3.2.5.

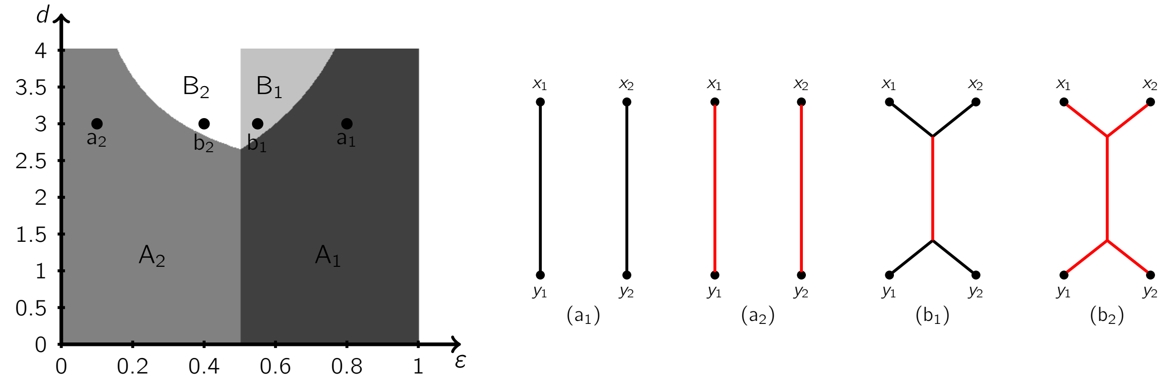


Figure 3.3.: Optimal network structures for the urban planning problem defined in Example 3.3.6 for $a = 2$. Left: Graphical illustration of the optimal network structures depending on the choice of ε (x-axis) and d (y-axis). Right: Structure of the optimal graph G for $d = 3$ and $\varepsilon = 0.8$ (a_1), $\varepsilon = 0.1$ (a_2), $\varepsilon = 0.55$ (b_1), $\varepsilon = 0.4$ (b_2). The network Σ is displayed in red.

Definition 3.3.8 (Urban planning problem)**.** Let $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ be two measures with equal mass, and set

$$\mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\mathcal{F}) = \begin{cases} \mathcal{E}^{a,\varepsilon}(\mathcal{F}) & \text{if } \operatorname{div}\mathcal{F} = \mu_+ - \mu_-, \\ \infty & \text{otherwise.} \end{cases}$$

Then the *urban planning problem* is defined as

$$\inf_{\mathcal{F}} \mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\mathcal{F}).$$

Existence of a solution of the urban planning problem has been shown in [20]. As in case of branched transport, it is known that an optimal mass flux decomposes to

$$\mathcal{F} = F\hat{e}(\mathcal{H}^1 \llcorner \Sigma) + \mathcal{F}^\perp$$

for a weight function $\widetilde{F} : \Sigma \to [0, \infty)$, an orientation $\hat{e} : \Sigma \to \mathbb{R}^n$ and a $\mathcal{H}^1$-diffuse part $\mathcal{F}^\perp$, which consists of a Lebesgue-continuous and a Cantor part. Then the urban planning cost functional can be written as

$$\mathcal{E}^{a,\varepsilon}(\mathcal{F}) = \int_\Sigma \min\{a\widetilde{F}, \widetilde{F} + \varepsilon\} \mathrm{d}\mathcal{H}^1 + a|\mathcal{F}^\perp|(\overline{\Omega}).$$

Remark 3.3.9. Since a part of $\mathcal{F}$ might correspond to a mass flux outside of a transportation network, a minimizer $\mathcal{F}$ can be locally continuous with respect to the Lebesgue measure $\mathcal{L}^n$ and thus does not necessarily coincide with a finite graph. As an example, consider $\mu_+ = \mathcal{L}^1 \llcorner [0,1] \times \{0\}$ and $\mu_- = \mathcal{L}^1 \llcorner [0,1] \times \{1\}$ in a two-dimensional space. Then, for a suitable choice of the parameters a and ε such that $1+\varepsilon < a$ (in other words, transport of mass 1 is cheaper on the network instead of outside), the minimizer is expected to be locally Lebesgue-continuous near the source and sink terms and contain a network of finite length in between (Figure 3.4, cf. [20]).

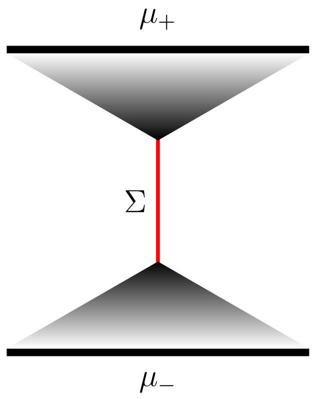


Figure 3.4.: Mass flux between two Lebesgue-continuous measures μ_+ and μ_-.

3.3.3. Lagrangian formulation

We want to briefly introduce the urban planning problem in its pattern-based notation. We make use of the same principles as in the corresponding branched transport formulation with respect to an irrigation pattern $\chi : \Gamma \times I \to \mathbb{R}^n$ (see Definition 3.2.10).

Definition 3.3.10 (Urban planning cost (Lagrangian formulation))**.** Let χ be an irrigation pattern. The *Lagrangian formulation of the urban planning cost functional* is defined as

$$\mathcal{E}^{a,\varepsilon}(\chi) = \int_{\Gamma\times I} r^{\chi}_{a,\varepsilon}(\chi_p(t))|\dot{\chi}_p(t)|\mathrm{d}P(p)\mathrm{d}t,$$

where

$$r^{\chi}_{a,\varepsilon}(x) = \begin{cases} \min\{1+\frac{\varepsilon}{m_\chi(x)}, a\} & \text{if } m_\chi(x) > 0,\\ a & \text{if } m_\chi(x) = 0. \end{cases}$$

Definition 3.3.11 (Urban planning problem (Lagrangian formulation))**.** Let $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ and χ be an irrigation pattern. Then the *Lagrangian formulation of the urban planning problem* is defined as

$$\inf\Big\{\mathcal{E}^{a,\varepsilon}(\chi) \ : \ i_0\#P = \mu_+, \ i_1\#P = \mu_-\Big\}.$$

For completeness, we state the equivalence between Wasserstein, the flux-based and the pattern-based formulations in the following theorem.

Theorem 3.3.12 (Equivalence of the minimization problems)**.** *Set*

$$\mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\Sigma) = \begin{cases} \mathcal{E}^{a,\varepsilon}(\Sigma) & \textit{if } \Sigma \textit{ admissible network,} \\ \infty & \textit{otherwise,} \end{cases}$$

$$\mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\chi) = \begin{cases} \mathcal{E}^{a,\varepsilon}(\chi) & \textit{if } i_0\#P = \mu_+,\ i_1\#P = \mu_-, \\ \infty & \textit{otherwise.} \end{cases}$$

Then the Wasserstein, the Eulerian and the Lagrangian minimization problems are equivalent in the sense that for measures μ_+, μ_- of equal mass and with bounded support we have

$$\min_{\Sigma} \mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\Sigma) = \min_{\mathcal{F}} \mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\mathcal{F}) = \min_{\chi} \mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\chi).$$

Proof. A proof can be found in [20]. □

3.3.4. Generalized urban planning

The urban planning problem is commonly associated with a cost function $c : [0,\infty) \to [0,\infty)$ for the amount of transported mass given by $c(w) = \min\{aw, w+\varepsilon\}$ for parameters $a > 1$, $\varepsilon > 0$. However, one can extend this cost functional in a straightforward way to a more general class of functions $c(w) = \min\{a_0 w, a_1 w + b_1, \dots, a_N w + b_N\}$ for values $a_0, a_i, b_i > 0$ for $i = 1, \dots, N$. This covers the urban planning problem as a special case and moreover entails another subdivision of the optimal network Σ into parts $\Sigma_1, \dots, \Sigma_N$ corresponding to the single terms of the cost function c. Since we will study numerical methods for this generalized urban planning problem within this work, we want to state the discrete and continuous *generalized urban planning functional* (in its Eulerian formulation) at this point.

Definition 3.3.13 (Discrete generalized urban planning cost)**.** For a discrete mass flux G between two discrete measures μ_+ and μ_- and parameters $a_0, a_i, b_i > 0$ for $i = 1, \dots, N$, the *discrete generalized urban planning cost* is given by

$$\mathcal{E}_g^{a,b}(G) = \sum_{e \in E(G)} \min\{a_0 w(e), a_1 w(e) + b_1, \dots, a_N w(e) + b_N\} l(e).$$

Definition 3.3.14 (Continuous generalized urban planning cost)**.** Let $\mathcal{F}$ be a mass flux between two finite Borel measures $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ with equal mass. Then the *continuous generalized urban planning cost functional* is defined as

$$\mathcal{E}_g^{a,b}(\mathcal{F}) = \inf\left\{ \liminf_{k\to\infty} \mathcal{E}_g^{a,b}(G_k) \ : (\mu_+^k, \mu_-^k, \mathcal{F}_{G_k}) \rightharpoonup^* (\mu_+, \mu_-, \mathcal{F}) \right\}$$

with $\mu_+^k, \mu_-^k, G_k, \mathcal{F}_{G_k}$ as in Definition 3.2.5.

As before, we obtain with $\mathcal{F} = \widetilde{F}\hat{e}(\mathcal{H}^1 \llcorner \Sigma) + \mathcal{F}^\perp$

$$\mathcal{E}_g^{a,\varepsilon}(\mathcal{F}) = \int_\Sigma c(\widetilde{F})\mathrm{d}\mathcal{H}^1 + c'(0)|\mathcal{F}^\perp|(\overline{\Omega}),$$

where $c'(0) = a_0$. For completeness, the *generalized urban planning problem* is defined in the following.

Definition 3.3.15 (Generalized urban planning problem)**.** Let $\mu_+, \mu_- \in \mathcal{M}_+(\Omega)$ be two measures with equal mass, and set

$$\mathcal{E}_g^{a,b,\mu_+,\mu_-}(\mathcal{F}) = \begin{cases} \mathcal{E}_g^{a,b}(\mathcal{F}) & \text{if } \mathrm{div}\mathcal{F} = \mu_+ - \mu_-, \\ \infty & \text{otherwise.} \end{cases}$$

Then the *generalized urban planning problem* is defined as

$$\inf_{\mathcal{F}} \mathcal{E}_g^{a,b,\mu_+,\mu_-}(\mathcal{F}).$$

3.4. Existence and properties of minimizers

The branched transport and urban planning model are common representatives of a wide class of problems involving a cost functional for the amount of transported mass of the form $\tau : [0, \infty) \to [0, \infty)$ such that $\tau(0) = 0$, τ is non-decreasing, subadditive and lower semi-continuous. Given these requirements, in [21] the authors proved that for μ_+, μ_- with compact support, either a minimizer of the corresponding problem exists or the cost functional is infinite everywhere ([21], Theorem 2.10). In addition, every discrete mass flux has to satisfy some interesting properties:

- **Acyclicity of discrete mass fluxes:** For any discrete mass flux G there exists a discrete mass flux $\bar{G}$ which contains no cycles and has lower or equal costs ([21], Lemma 2.5).
- **Tree structure of discrete mass fluxes:** For any discrete mass flux G, there exists a discrete mass flux $\bar{G}$ which admits a tree structure and has lower or equal costs ([21], Lemma 2.6).
- **Boundedness of transported mass:** Any acyclic discrete mass flux G satisfies $w(e) \leq \mu_+(\Omega) = \mu_-(\Omega)$ for all edges $e \in E(G)$ ([21], Lemma 2.9).

Finally, although existence of a minimizing transportation network is guaranteed, uniqueness cannot be achieved in the general setting due to a lack of convexity of the minimization problem. To the contrary, it is straightforward to construct examples where a problem admits several minimizers with a completely different topology but with exactly the same costs. Among others, this aspect causes the numerical treatment of optimal network problems to become a challenging task.

3.5. Numerical approaches

The main focus of this thesis will be a detailed study of numerical approaches suitable for the branched transport and urban planning problem introduced in Sections 3.2 and 3.3. Since both models are originally motivated by several applications, it is obvious that one is interested in solving the problems in some numerical framework in order to obtain transport paths for specific examples. Therefore, the models have not only been studied theoretically, but there also exist some considerations about how to computationally address the problem of finding an optimal path.

Every numerical treatment needs to face some technical difficulties arising from the non-convexity of the cost functional and the general structure of the problem in its different formulations. The first fundamental question one has to ask in this context is which formulation is suited best for a numerical treatment and how to represent and discretize the problem variables. For instance, taking in account the flux-based formulation, the flux variable could be represented explicitly as a discrete graph in terms of vertices and edges. While in case of small numbers of sources and sinks, the possible topologies of an optimal network for different parameters can be easily constructed by hand, for larger numbers this structure is not clear a priori and neither is the exact number of vertices. Hence, it is hard to verify if an obtained graph yields the global optimum or is only an approximation of unknown precision. Generally, the non-convexity of the cost function makes the application of several standard numerical methods challenging, which naturally brings up the question if there exists a suitable convexification.

In this section, we want to outline some existing numerical approaches to the branched transport and the Steiner tree problem. Note that to the best of our knowledge, we were the first to address the urban planning and generalized urban planning problem numerically in [19] and [33]. These approaches will be presented in Chapter 4 and Chapter 5. There exists a wide variety of numerical approaches to the branched transport problem and the Steiner tree problem in particular. Since a detailed presentation of each of these would be beyond the scope of this thesis, in the following we confine ourselves to a brief overview of the different strategies.

3.5.1. Branching point optimization

In case of discrete measures μ_+, μ_- as a given set of n points in total, the optimal network Σ simply consists of a set of vertices and edges. Thus, a straightforward minimization approach is the optimization of the vertex positions, provided that the topology of the minimizer and thus the number of vertices is known in advance. Since for very small n, the number of possible topologies is reasonable, this approach often leads to the exact global minimum by computing the optimal vertices for each topology independently and choosing the one with the minimal costs afterwards. However, for large n this strategy would involve a large number of non-linear equations for the vertex positions and the number of possible topologies becomes vast.

In the context of presenting the branched transport model in its flux-based formulation, the authors introduced an initial approach for numerically finding an optimal path between two measures [68],[70]. This local optimization technique was extended to a minimization algorithm in [69], which in some numerical examples seem to yield almost optimal networks in case of transport between a single source point and a fixed number of N sinks. The method makes use of the fact that the optimal network can be decomposed into several smaller transport network problems, which can be solved directly by a branching point optimization as described above. It was shown in [69],[70] that, although not necessarily leading to a global minimizer of $\mathcal{M}^\alpha$, this optimization algorithm provides an approximately optimal transport path which comprises a natural structure and is applicable even in case of a large number of sinks ($N \approx 400$). However, the algorithm does not necessarily provide the global optimal network, hence it cannot be shown that the method converges to a minimizer of the cost functional. Two heuristic approaches based on stochastic optimization techniques on graphs were presented in [44] and [53]. As before, these method are capable of providing almost optimal network structures, but cannot guarantee global optimality either.

The limit case of the Steiner tree problem in two space dimensions was treated more extensively in the literature. Due to the independence of the cost functional of the transported mass, there exist very efficient algorithms providing a globally optimal Steiner tree (see for instance the GeoSteiner method [38] or Melzak's full Steiner tree algorithm [46]). These methods are often restricted to the planar case, for a dimension greater than two, there exist fewer approaches which are less efficient. [31] provides an overview of some methods concerning the Steiner tree problem in n dimensions, where the main listed approaches trace back to [35], [61], [40], [34].

Since a more general cost functional such as in the branched transport case is concave in the transported mass, these methods commonly cannot be extended easily.

3.5.2. Phase field approximations

A widely used approach to tackle the branched transport problem numerically was inspired by elliptic approximations of free discontinuity problems. These problems typically consist in finding an optimal function (in the sense of some energy functional) which is smooth outside of a discontinuity set Σ, which is unknown as well. Common examples are the Modica–Mortola approximation of the perimeter of a set or the Ambrosio–Tortorelli model as an approximation of the Mumford–Shah functional (see Section 2.4.1). These methods minimize with respect to a smoothed version of the desired output, where the grade of smoothness is governed by a parameter ε, such that the relaxed functional Γ-converges to the original formulation as $\varepsilon \to 0$.

A similar approach can be applied to the branched transport problem as shown in [51] and [49]. In the following, we briefly repeat the main ideas presented by the authors. Recalling the explicit formulation of the continuous branched transport cost functional, we define

$$\mathcal{M}_0^\alpha(\mathcal{F}) = \begin{cases} \int_\Sigma \widetilde{F}^\alpha \mathrm{d}\mathcal{H}^1 & \text{if } \mathcal{F} = \widetilde{F}\hat{e}(\mathcal{H}^1 \llcorner \Sigma) \text{ and } \mathcal{F} \text{ has finite divergence,} \\ \infty & \text{otherwise} \end{cases}$$

as in equation (3.1). The vector measure $\mathcal{F}$ is concentrated on a one-dimensional rectifiable set $\Sigma \subset \mathbb{R}^n$. Let $n = 2$ and $\Omega \subset \mathbb{R}^2$ be an open set containing the support of μ_+ and μ_- such that $\overline{\Omega}$ is compact. The main results of [51] are given by the following two theorems.

Theorem 3.5.1 (Γ-convergence for branched transport for $\alpha \in (\frac{1}{2}, 1)$)**.** *Let $\alpha \in (\frac{1}{2}, 1)$ and $n = 2$. Define a functional $\mathcal{M}_\varepsilon^\alpha : L^1(\Omega) \to [0, \infty]$ as*

$$\mathcal{M}_\varepsilon^\alpha(u) = \begin{cases} \int_\Omega \varepsilon^{\alpha-1}|u(x)|^\beta + \varepsilon^{\alpha+1}|\nabla u(x)|^2 \mathrm{d}x & \text{if } u \in W^{1,2}(\Omega), \\ \infty & \text{otherwise,} \end{cases}$$

where $\beta = \frac{4\alpha-2}{\alpha+1}$. Then we have

$$\mathcal{M}_\varepsilon^\alpha \xrightarrow{\Gamma} c\mathcal{M}_0^\alpha \quad \text{for} \quad \varepsilon \to 0,$$

with $c = \alpha^{-1}(\frac{4c_0\alpha}{1-\alpha})^{1-\alpha}$, $c_0 = \int_0^1 \sqrt{t^\beta - t}\mathrm{d}t$.

Theorem 3.5.2 (Γ-convergence for branched transport for $\alpha \in (0, \frac{1}{2})$)**.** *Let $\alpha \in (0, \frac{1}{2})$ and $n = 2$. Define a functional $\mathcal{M}_\varepsilon^{\alpha,B} : L^1(\Omega) \to [0, \infty]$ as*

$$\mathcal{M}_\varepsilon^{\alpha,B}(u) = \begin{cases} \int_\Omega \varepsilon^{\alpha-1}B(|u(x)|) + \varepsilon^{\alpha+1}|\nabla u(x)|^2 \mathrm{d}x & \text{if } u \in W_0^{1,2}(\Omega), \\ \infty & \text{otherwise,} \end{cases}$$

where $\beta = \frac{2\alpha-1}{\alpha+1}$ and a continuous function $B : [0, \infty) \to [0, \infty)$ with $B(0) = 0$, $B(x) > 0$ for all $x > 0$, $\lim\limits_{t\to\infty} \frac{B(t)}{t^\beta} = 1$ and $B'(0) > 0$. Then we have

$$\mathcal{M}_\varepsilon^{\alpha,B} \xrightarrow{\Gamma} c\mathcal{M}_0^\alpha \quad \text{for} \quad \varepsilon \to 0,$$

with $c = \alpha^{-1}(\frac{4c_0\alpha}{1-\alpha})^{1-\alpha}$, $c_0 = \int_0^1 \sqrt{t^\beta - t}\mathrm{d}t$.

Proof. The proofs of Theorem 3.5.1 and 3.5.2 can be found in [51]. □

Remark 3.5.3. The difference between Theorem 3.5.1 and 3.5.2 comes from the fact that the construction and proof of the first result only work for $\alpha > 1 - \frac{1}{n}$. If $\alpha < \frac{1}{2}$, the integral involves the term $|u(x)|^\beta$ with $\beta < 0$, which causes the minimization of the functional to become more challenging. Some additional work is required in order to prove the second theorem, where α is allowed to be smaller than $\frac{1}{2}$.

Remark 3.5.4. Theorems 3.5.1 and 3.5.2 do not include the constraint div$\mathcal{F} = \mu_+ - \mu_-$, since $\mathcal{M}_0^\alpha$ only enforces $\mathcal{F}$ to have finite divergence. The extension was provided by [49] via introducing a relaxed constraint version, where the measure $\mu_+ - \mu_-$ is smoothed by convolution with a kernel $\rho_\varepsilon(x) = \varepsilon^{-2\gamma}\rho(\varepsilon^{-\gamma}x)$ for some $\rho \in C_0^1(\Omega, \mathbb{R}^+)$ and $\int_\Omega \rho(x)\mathrm{d}x = 1$ and $\gamma = \frac{\alpha+1}{3}$. Hence, setting $f_\varepsilon = \rho_\varepsilon * (\mu_+ - \mu_-)$, the author shows that

$$\mathcal{M}_\varepsilon^\alpha + \iota_{\mathrm{div}\ u=f_\varepsilon} \xrightarrow{\Gamma} c\mathcal{M}_0^\alpha + \iota_{\mathrm{div}\mathcal{F}=\mu_+-\mu_-} \quad \text{for} \quad \varepsilon \to 0$$

for some constant c.

Remark 3.5.5. The intuition behind Theorem 3.5.1 becomes clearer by considering the idea of the Modica–Mortola approximations [47]: Let us regard the functional

$$F_\varepsilon(u) = \int_\Omega \frac{1}{\varepsilon} W(u(x)) + \varepsilon|\nabla u(x)|^2 \mathrm{d}x$$

for $u \in W^{1,2}(\Omega)$ and W being a double well potential with $W(0) = W(1) = 0$. By minimizing this functional, the first term will enforce u to take either values in $\{0, 1\}$, the so-called two *phases*, while the second term will prefer smooth functions. Together with some constraint forcing u not to be a constant function, the minimizer will most likely admit a phase transition from one phase to the other, where the smoothness of this transition depends on the parameter ε. If ε tends to zero, the transition will become sharper until it converges to a piecewise constant function (cf. Section 2.4.1). Now returning to our functional defined in Theorem 3.5.1, the first term $|u(x)|^\beta$ plays the role of the double well potential with wells at 0 and ∞ (from $\alpha < 1$ we obtain $\beta < 1$, hence the first term is a concave function). Thus, $|u|$ wants to be either equal to 0 or as large as possible, being bounded by the finite divergence constraint.

Similar to the case of Γ-convergence for the Mumford–Shah functional, this result can be exploited for numerical purposes. Instead of minimizing the original branched transport energy functional, one can solve the relaxed functional with a small parameter ε. Roughly spoken, the minimizer now has the structure of a smooth grey value image and is not restricted to a one-dimensional set Σ, but lives on the whole image domain Ω. As a consequence, the functional can be treated with standard numerical differentiation approaches such as finite differences in combination with some minimization algorithm. In [51], the authors also observe that for large ε, the functional is even close to being convex, hence starting with a large value and then slowly decreasing ε, in each iteration starting with the solution from the previous step, reduces the chances of getting stuck in local minima. For small ε, the solution u should then yield a good approximation of the vector measure $\mathcal{F}$.

The phase field approximation presented by [51] and [49] is just one example for a bunch of related problems, such as [52],[13],[14]. An example of special interest is a comparable numerical treatment of the Steiner tree problem (which acts as a limit case of the branched transport problem for $\alpha \to 0$) presented in [24]. On the basis of [14], the authors present a slightly different phase field approach, where the phase field is introduced as an additional variable coexisting with a relaxed version of the vector-valued measure $\mathcal{F}$. More precisely, the authors define a functional $\mathcal{S}^\alpha : \mathcal{M}(\overline{\Omega}, \mathbb{R}^2) \times L^1(\Omega) \to [0, \infty]$ as

$$\mathcal{S}^\alpha(\sigma, \varphi) = \begin{cases} \int_\Sigma (1 + \alpha \widetilde{F}) \mathrm{d}\mathcal{H}^1 & \text{if } \varphi \equiv 1,\ \sigma = \widetilde{F}\hat{e}(\mathcal{H}^1 \llcorner \Sigma),\ \text{div}\, \sigma = \mu_+ - \mu_-, \\ \infty & \text{otherwise} \end{cases}$$

and state their main Γ-convergence result as follows:

Theorem 3.5.6 (Γ-convergence for the Steiner tree problem)**.** *Define a functional* $\mathcal{F}_\varepsilon : \mathcal{M}(\Omega, \mathbb{R}^2) \times L^1(\Omega) \to [0, \infty]$ *as*

$$\mathcal{F}_\varepsilon(\sigma, \varphi) = \begin{cases} \int_\Omega \frac{1}{2\varepsilon}\varphi^2|\sigma|^2 + \frac{\varepsilon}{2}|\nabla\varphi|^2 + \frac{1}{2\varepsilon}(1-\varphi)^2 \mathrm{d}x & \textit{if } (\sigma, \varphi) \in V_\varepsilon(\Omega) \times W_\varepsilon(\Omega), \\ \infty & \textit{otherwise} \end{cases}$$

with

$$\begin{aligned} V_\varepsilon(\Omega) &= \{\sigma \in L^2(\Omega, \mathbb{R}^2) \ : \ \textit{div}\, \sigma = (\mu_+ - \mu_-) * \rho_\varepsilon\}, \\ W_\varepsilon(\Omega) &= \{\varphi \in W^{1,2}(\Omega) \ : \ \eta \le \varphi \le 1 \quad \textit{in}\ \Omega,\ \varphi \equiv 1 \quad \textit{on}\ \partial\Omega\}. \end{aligned}$$

Then we have

$$\mathcal{F}_\varepsilon \xrightarrow{\Gamma} \mathcal{S}^\alpha \quad \textit{on} \quad \mathcal{M}(\Omega, \mathbb{R}^2) \times L^1(\Omega),$$

in the sense of the weak- topology on* $\mathcal{M}(\Omega, \mathbb{R}^2)$ *and the classical strong topology on* $L^1(\Omega)$*.*

Proof. The proof can be found in [24]. □

Remark 3.5.7. Note that the cost functional $\mathcal{S}^\alpha$ is not exactly the Steiner tree problem, but is linear in the transported mass (instead of constant). Numerical simulations in [24] are obtained for a small parameter α, such that the results approximate the Steiner case quite well.

Remark 3.5.8. The approximating functional $\mathcal{F}_\varepsilon$ seems even more intuitive at this point having a closer look at the single terms. The second and third part of the functional correspond to a Modica–Mortola-type component, while the first part acts as a linker between phase field φ and vector field σ. For $\sigma = 0$ (or very small values of σ), φ is drawn to 1, since the first part does not play a role and the second and third part prefer constant functions $\varphi \equiv 1$. For σ larger, hence at those points where the (relaxed) transport network has support, the first part draws φ to 0. The second part provides for a smooth transition between the values 0 and 1 of φ, while the grade of smoothness is governed by the parameter ε: Roughly spoken, for smaller ε, the gradient term becomes less important and thus the transition between the two phases becomes sharper.

We use the approach presented by [24] to develop a phase field approximation of the generalized urban planning model, which covers a larger class of problems including the Steiner tree case. The method will be described in more details in Chapter 5.

3.5.3. Time-dependent PDE-based methods

In [28], the authors propose a PDE-based formulation of the classical Kantorovich problem as in Definition 3.1.2 with c being the Euclidean distance. In some recent works [29], [30], this idea has been seized in order to define a time-dependent system of equations whose long-time solution is conjectured to yield a solution to the Kantorovich problem. The resulting partial differential equations have been further extended by a branching parameter $\alpha > 0$ in [9]. Therein, the authors introduce the system

$$\begin{aligned} -\nabla \cdot (\xi(t,x)\nabla u(t,x)) &= \mu_+(x) - \mu_-(x) \\ \partial_t \xi(t,x) &= (\xi(t,x)|\nabla u(t,x)|)^\alpha - \xi(t,x) \\ \xi(0,x) &= \mu_+(x) \end{aligned} \tag{3.3}$$

equipped with zero Neumann boundary conditions. Here, ξ denotes the density of transported mass and u corresponds to the transport potential, i.e. the optimal u for $\alpha = 1$ maximizes the Kantorovich dual problem [66], [28]. The system of equations (3.3) is then solved in two space dimensions via a finite element triangulation combined with an explicit Euler discretization in time.

By presenting some numerical simulation results, the authors experimentally show that for $\alpha > 1$, the obtained transport density admits a branching structure and thus resembles a solution to the branched transport problem.

3.5.4. Discussion

Although there exists a variety of promising numerical approaches to the branched transport problem and some extensions, the complexity of the methods shows that the problems do not admit a straightforward computational treatment. One reason certainly is the non-convexity of the cost functional (in the general case), which causes many common numerical methods to get stuck in local minima easily. Another reason is the rising complexity of the underlying network topology as the number of sources and sinks (in case of atomic measures) increases. This can be seen exemplarily by studying the method proposed by [69]. Although the algorithm provides a high quality result and skilfully exploits several facts about the network structure, it could not be shown that it converges to a global minimizer. In addition, the method does not work in case of a source consisting of more than one point. As shown by [71], the exact construction method for an optimal network can be extended to this case, but requires some knowledge about the possible network topologies. As a consequence, such methods easily become computationally inefficient.

In order to avoid the incorporation of prior knowledge about the network structure, phase field approximation methods are a promising tool. The constructed energy functionals usually allow a simple numerical treatment and are even close to being convex (as shown by [51], for instance). However, the method still bears some disadvantages, such as a possible strong dependence on the choice of the initial values and the parameters. As shown in [6], the Γ-convergence of a finite difference discretization of the energy functional depends strongly on the choice of the discretization step size.

Since the existing methods, although admitting promising results, do not cover the whole class of problems involving a concave cost functional in a satisfactory way, this work is devoted to the investigation of some novel aspects about the numerical treatment of transportation networks. We introduce a completely new approach making use of the principles of functional lifting in order to obtain a novel numerical framework in Chapter 4. Additionally, based on past works, we extend the phase field approximation approach to the case of a generalized urban planning problem, both methods also admitting one of the first numerical treatments of the urban planning problem to the best of our knowledge (see Chapter 5).

4 Numerical optimization of transportation networks via functional lifting

Although the branched transport and urban planning energy functionals have been extensively studied and analysed, it still remains a challenging task to compute optimal networks numerically. After reviewing some properties such as the non-convexity of the energy (see Chapter 3) as well as some existing numerical approaches (cf. Section 3.5), we have learned that one still lacks a *simple* mathematical model allowing to apply iterative methods leading to a global minimizer.

In this chapter, we present an approach that ascribes the task of finding an optimal transportation network to a mathematical imaging problem, which was first introduced in [19]. The method reformulates the energy functionals in terms of functions of bounded variation, which opens the door for some well-studied mathematical imaging techniques. Precisely, the new energy admits a convex reformulation via so-called *functional lifting* (see Section 2.4.2), where the original model is lifted to a higher-dimensional space. In the course of this chapter, we will derive and analyse the resulting imaging problem in respect of its numerical implementation. We review some straightforward discretization methods and suggest a novel general treatment of problems arising from functional lifting via special adaptive finite elements.

4.1. Model

First, we want to provide a detailed description of the functional lifting approach as introduced in Section 2.4.2 applied to the branched transport and urban planning model and the resulting convex minimization problems.

4.1.1. Reformulation as image inpainting problems in two dimensions

Let us recall the definitions of the branched transport and urban planning energy derived in Chapter 3 with respect to the continuous mass flux $\mathcal{F} \in \mathcal{M}\left(\mathbb{R}^n, \mathbb{R}^n\right)$,

$$\mathcal{M}^{\alpha,\mu_+,\mu_-}(\mathcal{F}) = \begin{cases} \int_\Sigma \widetilde{F}^\alpha \mathrm{d}\mathcal{H}^1 + \iota_0(|\mathcal{F}^\perp|) & \text{if } \mathcal{F} \text{ is a mass flux between } \mu_+ \text{ and } \mu_-, \\ \infty & \text{otherwise,} \end{cases}$$

$$\mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\mathcal{F}) = \begin{cases} \int_\Sigma \min\{a\widetilde{F}, \widetilde{F} + \varepsilon\} \mathrm{d}\mathcal{H}^1 + a|\mathcal{F}^\perp|(\overline{\Omega}) & \text{if } \mathcal{F} \text{ is a mass flux between } \mu_+ \text{ and } \mu_-, \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathcal{F} = \widetilde{F}\hat{e}\,(\mathcal{H}^1 \llcorner \Sigma) + \mathcal{F}^\perp$ with a real multiplicity $\widetilde{F} : \Sigma \to (0, \infty)$, an orientation $\hat{e} : \Sigma \to \mathbb{R}^n$, $|\hat{e}| = 1$, a diffuse part $\mathcal{F}^\perp$ and

$$\iota_0\left(x\right) = \begin{cases} 0 & \text{if } x = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Following the course of [19], Chapter 3, we show that in two dimensions the energy admits a reformulation as a Mumford–Shah-type imaging problem and hence the task of finding an optimal network can be regarded as an image inpainting problem.
In the following, we set $n = 2$ and $\Omega \subset \mathbb{R}^2$, $V \subset \mathbb{R}^2$ be an open bounded convex domain with $\Omega \subset\subset V$. For simplicity, we shall assume that $\Omega = (0, l) \times (0, 1)$, $V = B_1\left(\Omega\right)$ and

$$\mathrm{spt}\mu_+ \subset \partial\Omega,\ \mathrm{spt}\mu_- \subset \partial\Omega.$$

Note that a generalization of the image domain as well as of the restriction of initial and final measure to the region boundaries is possible (see examples in Section 4.3.4), however, for the clearer understanding we restrict ourselves to the simpler case of initial and final measure on the domain boundaries. Let $u \in BV\left(V\right)$ be a grey value image. As seen in Section 2.2, the derivative of u can be decomposed into a Lebesgue-continuous part, a discontinuous jump part on a set S_u and a Cantor part $D^c u$

$$Du = \nabla u \mathcal{L}^2 \llcorner V + [u]\nu\mathcal{H}^1 \llcorner S_u + D^c u,$$

where $[u] = u^+ - u^-$ denotes the height of the jump across S_u. Via this gradient decomposition, we can define a mass flux $\mathcal{F}_u$ associated to u as the rotated gradient of the

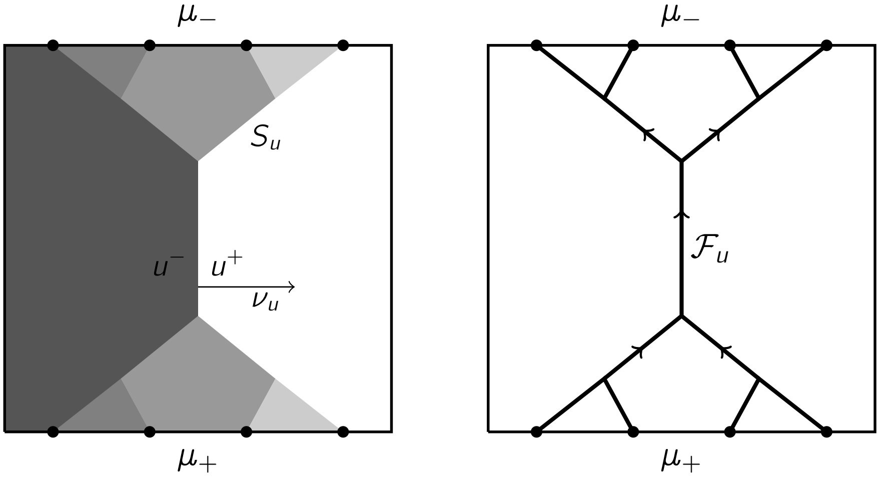


Figure 4.1.: Comparison between a grey value image u (left) and the mass flux $\mathcal{F}_u$ associated with u (right) on a domain $\overline{\Omega}$. The discontinuity set S_u can be regarded as the transportation network, the jump in function value across S_u encodes the amount of mass flowing through the corresponding network segment.

image, which then points to the direction of the network orientation (cf. [19], Definition 3.1.1):

Definition 4.1.1 (Mass flux associated with an image)**.** For an image $u \in BV(V)$, we define the *mass flux associated with* u $\mathcal{F}_u \in \mathcal{M}(V, \mathbb{R}^2)$ as

$$\mathcal{F}_u = Du^\perp = \nabla u^\perp \mathcal{L}^2 \llcorner V + [u]\nu^\perp \mathcal{H}^1 \llcorner S_u + D^c u^\perp,$$

where superscript $\perp$ denotes the counterclockwise rotation by $\frac{\pi}{2}$.

The optimal network associated with the mass flux $\mathcal{F}_u$ can be interpreted as the discontinuity set S_u and the height of the jump across S_u in some point $x \in V$ indicates the amount of mass flowing through x (cf. Figure 4.1). Note that, as we will see shortly, in case of branched transport, the image has to be piecewise constant, while in urban planning, the diffuse component of the cost functional admits regions with a non-zero Lebesgue-continuous gradient ∇u.

Remark 4.1.2. For $\mathcal{F}_u$ representing a mass flux on the domain V one requires that no mass is created or lost within the transport region. Indeed, $\mathcal{F}_u$ is divergence-free in the distributional sense: Let $\varphi \in C_0^\infty(V)$ be a smooth test function, then we have (cf. [19])

$$\int_V \varphi \; \mathrm{d}(\mathrm{div}\mathcal{F}_u) = -\int_V \nabla\varphi \cdot \mathrm{d}\mathcal{F}_u = \int_V \nabla\varphi^\perp \cdot \mathrm{d}Du = -\int_V \mathrm{div}(\nabla\varphi^\perp) u \; \mathrm{d}x = 0.$$

Similar to the definition of the set of admissible fluxes

$$\mathcal{A}_{\mathcal{F}}\left(\mu_+,\mu_-\right) = \{\mathcal{F} \in \mathcal{M}\left(V,\mathbb{R}^2\right) \ : \ \mathrm{spt}\mathcal{F} \subset \overline{\Omega},\ \mathrm{div}\mathcal{F} = \mu_+ - \mu_-\},$$

we want to define the set of admissible images. To this end, we need to reformulate the divergence constraint for the mass flux in the sense of images. As mentioned before, we restrict ourselves to μ_+, μ_- with support on $\partial\Omega$, which without loss of generality can be assumed to lie in the right halfplane of $\mathbb{R}^2$. Then we can define a parameterization of $\partial\Omega$ by

$$\gamma : [0, \mathcal{H}^1\left(\partial\Omega\right)) \to \partial\Omega$$

with $\gamma(0) = 0$ and $\partial\Omega_t := \gamma\left([0,t)\right)$. Furthermore, we define the orthogonal projection of a point $x \in \mathbb{R}^2$ onto $\partial\Omega$ via

$$\pi_{\partial\Omega} : \mathbb{R}^2 \to \partial\Omega,\ x \mapsto \mathrm{argmin}\{|x-y| : y \in \partial\Omega\}.$$

Note that the closest point might not be unique since Ω is not strictly convex in many applications; in this case, we pick the lexicographically first point (in the sense of the parameterization γ). Even for convex regions Ω, it happens that for $x \in \Omega$, $\pi_{\partial\Omega}(x)$ is not unique. Note that for the optimization problem, only the definition of $u(\mu_+,\mu_-)$ outside of Ω is important, since Ω takes the role of the inpainting region.

Definition 4.1.3 (Admissible image)**.** Given finite Borel measures μ_+, μ_-, we define the function

$$u\left(\mu_+,\mu_-\right) : \mathbb{R}^2 \to \mathbb{R},\ x \mapsto \left(\mu_+ - \mu_-\right)\left(\partial\Omega_{\gamma^{-1}(\pi_{\partial\Omega}(x))}\right).$$

Then the *set of admissible images* is given as

$$\mathcal{A}_u\left(\mu_+,\mu_-\right) := \{u \in BV\left(V\right) \ : \ u = u\left(\mu_+,\mu_-\right) \text{ on } V \setminus \overline{\Omega}\}.$$

Before we can translate the branched transport and urban planning cost functional to the image setting, we have to make sure that the relation between images and fluxes is one-to-one, which implicates that every admissible mass flux can indeed be identified with a unique grey value image and vice versa. Hence we state the following result [19].

Theorem 4.1.4 (Bijection between admissible fluxes and images)**.** *The mapping* $u \mapsto \mathcal{F}_u \llcorner \overline{\Omega}$ *from* $\mathcal{A}_u\left(\mu_+,\mu_-\right)$ *to* $\mathcal{A}_{\mathcal{F}}\left(\mu_+,\mu_-\right)$ *is one-to-one.*

Proof. The proof can be found in [19], Lemma 3.1.4. □

Remark 4.1.5. The reason why this method only works in two spatial dimensions is that a mass flux on a higher-dimensional domain does not admit an interpretation as an image gradient [19]. In other words, the gradient of a higher-dimensional grey value image is not a one-dimensional set, as required for the interpretation as a mass flux.

Now we have the tools for reformulating the energy functionals $\mathcal{M}^\alpha$ and $\mathcal{E}^{a,\varepsilon}$ with respect to u within the following definition.

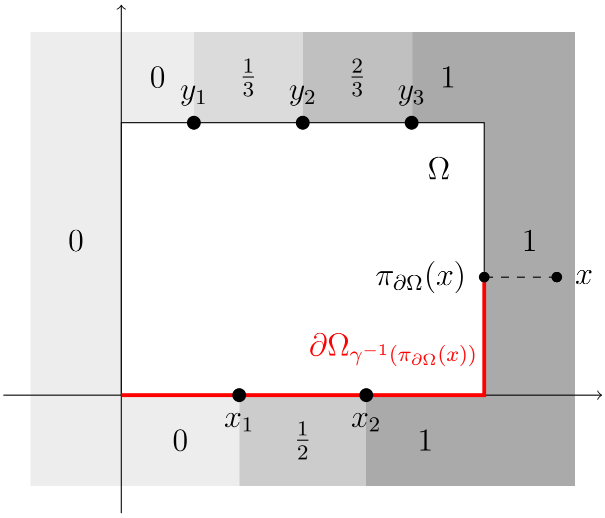


Figure 4.2.: Sketch of $u(\mu_+, \mu_-)$ for $\mu_+ = \frac{1}{2}(\delta_{x_1} + \delta_{x_2})$ and $\mu_- = \frac{1}{3}(\delta_{y_1} + \delta_{y_2} + \delta_{y_3})$. $u(\mu_+, \mu_-)$ is a grey value image taking values in $\{0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\}$. Note that $u(\mu_+, \mu_-)$ is also defined in the interior of Ω, but the set of admissible images $\mathcal{A}_u(\mu_+, \mu_-)$ only requires a definition of $u(\mu_+, \mu_-)$ on $V \setminus \overline{\Omega}$.

Definition 4.1.6 (Image cost functionals)**.** We define the functionals $\tilde{\mathcal{M}}^\alpha, \tilde{\mathcal{E}}^{a,\varepsilon} : BV(\overline{\Omega}) \to [0, \infty]$ as

$$\tilde{\mathcal{M}}^\alpha(u) = \int_{S_u \cap \overline{\Omega}} [u]^\alpha \mathrm{d}\mathcal{H}^1(x) + \iota_0\left(\left(\nabla u \mathcal{L}^2 + D^c u\right) \llcorner \overline{\Omega}\right),$$

$$\tilde{\mathcal{E}}^{a,\varepsilon}(u) = a \int_{\overline{\Omega} \setminus S_u} |Du| \mathrm{d}x + \int_{S_u \cap \overline{\Omega}} \min\{a[u], [u] + \varepsilon\} \mathrm{d}\mathcal{H}^1(x),$$

where

$$\iota_0(\mu) = \begin{cases} 0 & \text{if } \mu = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Furthermore, for two measures $\mu_+, \mu_- \in \mathcal{M}_+(\overline{\Omega})$ of equal mass with support on $\partial\Omega$ we define $\tilde{\mathcal{M}}^{\alpha,\mu_+,\mu_-}, \tilde{\mathcal{E}}^{a,\varepsilon,\mu_+,\mu_-} : BV(V) \to [0, \infty]$ as

$$\tilde{\mathcal{M}}^{\alpha,\mu_+,\mu_-}(u) = \begin{cases} \tilde{\mathcal{M}}^\alpha(u) & \text{if } u \in \mathcal{A}_u(\mu_+, \mu_-), \\ \infty & \text{otherwise,} \end{cases}$$

$$\tilde{\mathcal{E}}^{a,\varepsilon,\mu_+,\mu_-}(u) = \begin{cases} \tilde{\mathcal{E}}^{a,\varepsilon}(u) & \text{if } u \in \mathcal{A}_u(\mu_+, \mu_-), \\ \infty & \text{otherwise.} \end{cases}$$

The following results expresses the relation between the flux-related and the image-related cost functionals.

Theorem 4.1.7 (Lower bound on transport energy)**.** *For any flux $\mathcal{F} \in \mathcal{A}_{\mathcal{F}}(\mu_+, \mu_-)$ and the corresponding image $u_{\mathcal{F}} \in \mathcal{A}_u(\mu_+, \mu_-)$, we have*

$$\mathcal{M}^{\alpha,\mu_+,\mu_-}(\mathcal{F}) \geq \tilde{\mathcal{M}}^{\alpha,\mu_+,\mu_-}(u_{\mathcal{F}}), \ \ \mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\mathcal{F}) \geq \tilde{\mathcal{E}}^{a,\varepsilon,\mu_+,\mu_-}(u_{\mathcal{F}}).$$

Proof. The proof can be found in [19]. It follows from a few preliminary results (also shown in [19], namely the bijection between fluxes and images, the sequentially weak-* lower semi-continuity of $\tilde{\mathcal{M}}^{\alpha}$ and $\tilde{\mathcal{E}}^{a,\varepsilon}$, the equivalence for discrete measures (see Theorem 4.1.9) and

$$\mathcal{M}^{\alpha}(\mathcal{F}) = \inf\left\{\liminf_{k\to\infty}\mathcal{M}^{\alpha}(G_k) \ : \ (\mu_+^k, \mu_-^k, \mathcal{F}_{G_k}) \rightharpoonup^* (\mu_+, \mu_-, \mathcal{F}), \ \mathrm{spt}\ \mu_+, \mathrm{spt}\ \mu_- \subset \partial\Omega\right\}$$

for an approximating graph sequence G_k as in Definition 3.2.5 (the same result holds for $\mathcal{E}^{a,\varepsilon}(\mathcal{F})$). □

Remark 4.1.8. The authors believe, but so far could not prove, that the opposite inequality holds as well. This would imply that both energy functionals are equal and the branched transport and urban planning energy can indeed be reformulated as Mumford–Shah-type image inpainting problems. So far, it has only been shown that the original flux-related energy is bounded from below by the Mumford–Shah-type energy, which can still be used for numerical issues. The proof would require that the functionals $\tilde{\mathcal{M}}^{\alpha,\mu_+,\mu_-}$ and $\tilde{\mathcal{E}}^{a,\varepsilon,\mu_+,\mu_-}$ are the sequentially weakly-* lower semi-continuous envelopes of their discrete versions with respect to graphs.

In the case of discrete measures, one can show that equality in Theorem 4.1.7 holds true (cf. [19], Lemma A.0.2).

Theorem 4.1.9. *Let $\mu_+, \mu_- \in \mathcal{M}_+(\overline{\Omega})$ two measures of equal mass with* spt $\mu_+ \subset \partial\Omega$*,* spt $\mu_- \subset \partial\Omega$ *and let G be a transport path between μ_+ and μ_-. Let $\mathcal{F}_G = \sum_{e\in E(G)} w(e)(\mathcal{H}^1 \llcorner e)\hat{e}$. Then we have*

$$\mathcal{M}^{\alpha}(G) = \tilde{\mathcal{M}}^{\alpha}(u_{\mathcal{F}_G}), \ \ \mathcal{E}^{a,\varepsilon}(G) = \tilde{\mathcal{E}}^{a,\varepsilon}(u_{\mathcal{F}_G}).$$

Proof. For the sake of readability, we set $u := u_{\mathcal{F}_G}$. Since the mapping $u \mapsto \mathcal{F}_u \llcorner \overline{\Omega}$ from $\mathcal{A}_u(\mu_+, \mu_-)$ to $\mathcal{A}_{\mathcal{F}}(\mu_+, \mu_-)$ is a bijection (Theorem 4.1.4), $u \in \mathcal{A}_u(\mu_+, \mu_-)$ and

$$\mathcal{F}_G = \sum_{e\in E(G)} w(e)(\mathcal{H}^1 \llcorner e)\hat{e} = Du^{\perp} \llcorner \overline{\Omega} \ \Rightarrow \ Du \llcorner \overline{\Omega} = \sum_{e\in E(G)} w(e)(\mathcal{H}^1 \llcorner e)\hat{e}^{\perp}.$$

Thus, u is a piecewise constant constant function with discontinuity set

$$S_u \cap \overline{\Omega} = \bigcup_{e\in E(G)} e$$

with jump height $[u] = w(e)$. By inserting this into the definition of the image-related energies for branched transport and urban planning, we obtain

$$\tilde{\mathcal{M}}^{\alpha}(u) = \int_{S_u \cap \overline{\Omega}} [u]^{\alpha} \mathrm{d}\mathcal{H}^1(x) = \sum_{e \in E(G)} w(e)^{\alpha} l(e) = \mathcal{M}^{\alpha}(G),$$

$$\tilde{\mathcal{E}}^{a,\varepsilon}(u) = \int_{S_u \cap \overline{\Omega}} \min\{a[u], [u] + \varepsilon\} \mathrm{d}\mathcal{H}^1(x) = \sum_{e \in E(G)} l(e) \min\{aw(e), w(e) + \varepsilon\} = \mathcal{E}^{a,\varepsilon}(G).$$

□

The image-related energy functionals defined in Definition 4.1.6 for branched transport and urban planning admit an interpretation as image inpainting problems: The image u is prescribed on $V \setminus \overline{\Omega}$ and unknown inside $\overline{\Omega}$ and has to be reconstructed such that the energy becomes minimal. This minimization problem is comparable to the task of image inpainting via total variation regularization (see for instance [27]), where the cost functional is linear in the height of the jump along the discontinuity set. In our cases, the integrand of the jump set S_u is subadditive, while the part away from S_u is convex, which is similar to the behaviour of the Mumford–Shah image segmentation problem. This is the key observation for the numerical treatment presented in the following: It was shown (cf. Section 2.4.2) that the Mumford–Shah energy functional, although non-convex, admits a convex higher-dimensional reformulation, which can be used as a starting point for numerical simulations.

4.1.2. Functional lifting of the branched transport and urban planning energy

We want to apply the functional lifting approach introduced in Section 2.4.2 to the image-related branched transport and urban planning energy. For given initial and final measures $\mu_+, \mu_- \in \mathcal{M}_+(\mathbb{R}^2)$ we define $1_{u(\mu_+,\mu_-)}$ as the characteristic function of the subgraph of the function $u(\mu_+, \mu_-)$ defined in Definition 4.1.3. Note that both $\tilde{\mathcal{M}}^{\alpha,\mu_+,\mu_-}$ and $\tilde{\mathcal{E}}^{a,\varepsilon,\mu_+,\mu_-}$ can be written in the general framework of (2.6) (after restricting the image-related energy functionals to the slightly smaller space $SBV(V)$, which does not severely affect their behaviour) with the specific choice of g and h as

$$g(x, u, p) = \iota_0(p), \;\; h\left(x, u^+, u^-, \nu\right) = |u^+ - u^-|^{\alpha}$$

in case of branched transport and

$$g(x, u, p) = ap, \;\; h\left(x, u^+, u^-, \nu\right) = \min\{a|u^+ - u^-|, |u^+ - u^-| + \varepsilon\}$$

for the urban planning energy. Consequently, the sets $\mathcal{C}$ and $\mathcal{K}$ for the lifted saddle point problem become in our case

$$\begin{aligned}\mathcal{C} = \{&v \in SBV\left(V \times \mathbb{R}, [0,1]\right) \ : \\ &\lim_{t\to-\infty} v\left(x,t\right) = 1,\ \lim_{t\to\infty} v\left(x,t\right) = 0,\ v = 1_{u(\mu_+,\mu_-)} \text{ on } (V \setminus \overline{\Omega}) \times \mathbb{R}\}\end{aligned} \tag{4.1}$$

$$\begin{aligned}\mathcal{K}_1 = \{&\phi = (\phi^x, \phi^s) \in C_0^\infty\left(V \times \mathbb{R}, \mathbb{R}^2 \times \mathbb{R}\right) \ : \ \phi^s \geq 0, \\ &\left|\int_{s_1}^{s_2} \phi^x\left(x,s\right) \mathrm{d}s\right| \leq |s_2 - s_1|^\alpha \ \forall x \in V, s_1, s_2 \in \mathbb{R}\}\end{aligned} \tag{4.2}$$

$$\begin{aligned}\mathcal{K}_2 = \{&\phi = (\phi^x, \phi^s) \in C_0^\infty\left(V \times \mathbb{R}, \mathbb{R}^2 \times \mathbb{R}\right) \ : \ \phi^s \geq 0, \\ &|\phi^x| \leq a,\ \left|\int_{s_1}^{s_2} \phi^x\left(x,s\right) \mathrm{d}s\right| \leq \min\{|s_2 - s_1| + \varepsilon, a|s_2 - s_1|\} \ \forall x \in V, s_1, s_2 \in \mathbb{R}\}\end{aligned} \tag{4.3}$$

and the optimization problem for branched transport and urban planning reads

$$\inf_{v\in\mathcal{C}} \sup_{\phi\in\mathcal{K}} \int_{\overline{\Omega}\times\mathbb{R}} \phi \cdot \mathrm{d}Dv \tag{4.4}$$

with $\mathcal{K} = \mathcal{K}_1$ for branched transport and $\mathcal{K} = \mathcal{K}_2$ for urban planning respectively. Both cases yield convex optimization problems in v. Although v is allowed to take values in between 0 and 1, in some cases the optimal v still satisfies $v = 1_u$ for some $u \in SBV\left(V\right)$; however, some numerical examples show the discrepancy between the original image-related formulation and its convex relaxation (cf. Section 4.2). We will further investigate the behaviour of the relaxation and its impact on the optimal network in Section 4.2.1.

4.2. Analysis

In this section, we aim at pointing out some analytical aspects regarding the numerical treatment of the previously described model. In particular, we investigate the tightness of the convexification by reference to a particular example in order to obtain a better understanding of the involved energies. Furthermore, we show that the convex set $\mathcal{K}$ only contains a finite number of constraints in case of discrete functions.

4.2.1. Original formulation versus convexification

We have shown that the original flux-based formulation of the branched transport and urban planning problem admits a convex reformulation as an image inpainting problem via functional lifting. Putting together all the exhibited inequalities, we obtain

$$\begin{aligned}
&\min_{\mathcal{F}\in\mathcal{A}_{\mathcal{F}}(\mu_+,\mu_-)} \mathcal{M}^{\alpha,\mu_+,\mu_-}(\mathcal{F}) \geq \min_{u_{\mathcal{F}}\in\mathcal{A}_u(\mu_+,\mu_-)} \tilde{\mathcal{M}}^{\alpha,\mu_+,\mu_-}(u_{\mathcal{F}}) \\
&\geq \min_{u_{\mathcal{F}}\in\mathcal{A}_u(\mu_+,\mu_-)} \sup_{\phi\in\mathcal{K}_1} \int_{\Omega\times\mathbb{R}} \phi\cdot \mathrm{d}D1_{u_{\mathcal{F}}} \geq \inf_{v\in\mathcal{C}} \sup_{\phi\in\mathcal{K}_1} \int_{\Omega\times\mathbb{R}} \phi\cdot\mathrm{d}Dv, \\
&\min_{\mathcal{F}\in\mathcal{A}_{\mathcal{F}}(\mu_+,\mu_-)} \mathcal{E}^{a,\varepsilon,\mu_+,\mu_-}(\mathcal{F}) \geq \min_{u_{\mathcal{F}}\in\mathcal{A}_u(\mu_+,\mu_-)} \tilde{\mathcal{E}}^{a,\varepsilon,\mu_+,\mu_-}(u_{\mathcal{F}}) \\
&\geq \min_{u_{\mathcal{F}}\in\mathcal{A}_u(\mu_+,\mu_-)} \sup_{\phi\in\mathcal{K}_2} \int_{\Omega\times\mathbb{R}} \phi\cdot \mathrm{d}D1_{u_{\mathcal{F}}} \geq \inf_{v\in\mathcal{C}} \sup_{\phi\in\mathcal{K}_2} \int_{\Omega\times\mathbb{R}} \phi\cdot\mathrm{d}Dv.
\end{aligned}$$

This correlation naturally raises the question whether some of the inequalities can be replaced by an equality under certain conditions.

The first inequality was shown in [19] as stated in Theorem 4.1.7. As mentioned, the authors were not able to prove the opposite inequality, but believe that it holds as well. However, for the special case of discrete mass fluxes, equality is obtained by some simple calculations (cf. Theorem 4.1.9).

The second inequality arises from the functional lifting approach. Equality can be achieved if one is able to construct a vector field $\phi\in\mathcal{K}$ such that the value of the functional on the right-hand side of the second inequality equals the one of the left-hand side. While for some special cases, such as the Mumford–Shah functional, the existence of such a vector field is well established (see for instance [54]), the explicit construction remains a quite technical issue and depends on the particular choice of the integrands. In [2], this subject is revised in more details, additionally an exemplary construction of ϕ for the minimal partition problem is provided.

The most interesting case for numerical purposes is the last inequality, arising from the convex relaxation of the binary characteristic function 1_u to functions which are allowed to take values in between 0 and 1. On the one hand, we have already observed in Section 2.4, Theorem 2.4.8, that the existence of a *divergence-free* vector field ϕ guarantees equality of the minima of the original and the relaxed minimization problem (note that in our case, the convex set $\mathcal{C}$ naturally contains the necessary conditions of v being prescribed on the boundaries on $\partial\Omega\times\mathbb{R}$). On the other hand, it still remains an open question whether a minimizer of the relaxed functional also admits a minimizer of the branched transport or urban planning energy.

In case of the one-dimensional Mumford–Shah functional, in [23] it was shown that there exists a an equivalent convex representation consisting of a slightly more precise version of the functional lifting approach. However, this idea could not be extended to higher dimensions so far [23], so that in particular it cannot be applied to the branched transport and urban planning energies. Hence, we will investigate the tightness of the convex relaxation for these problems within this section.

The question of equality of the minima is related to the task of finding an optimal divergence-free $\phi\in\mathcal{K}$ which realizes the supremum on the right-hand side of the inequality. In order to gain a better understanding of the problem structure, in the following we want

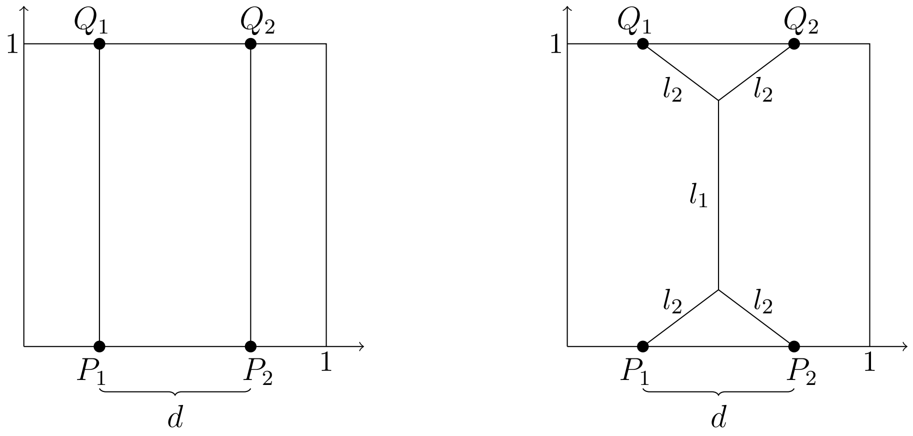


Figure 4.3.: Transport from two to two mass points in $\mathbb{R}^2$ as described in Example 4.2.1. Left: Topology of the graph G_1. Right: Topology of the graph G_2.

to restrict ourselves to the branched transport case and investigate the particular example of transport from two source points to two sinks of equal mass.

Example 4.2.1 (Branched transport cost for transport from two to two mass points)**.** Let $\Omega = [0,1]^2$, $V = B_1(\Omega)$ and $P_1, P_2, Q_1, Q_2 \in \partial\Omega$ be the four vertices of a rectangle with side lengths 1 and $d \leq 1$ (see Figure 4.3). Let $\mu_+ = m(\delta_{P_1} + \delta_{P_2})$, $\mu_- = m(\delta_{Q_1} + \delta_{Q_2})$ for some $m > 0$. Depending on $\alpha \in (0,1)$, there exist two possible topologies for the optimal graph G minimizing the branched transport cost $\mathcal{M}^\alpha(G)$ (Figure 4.3). Denoting G_1 as the graph consisting of two straight lines and G_2 as the single tree and $\mathcal{F}_1$ ($\mathcal{F}_2$ respectively) the discrete mass flux associated with the graph G_1 (G_2 respectively), then we have

$$\mathcal{M}^\alpha(\mathcal{F}_1) = 2m^\alpha,\ \mathcal{M}^\alpha(\mathcal{F}_2) = 2^\alpha m^\alpha l_1 + 4m^\alpha l_2,$$

where l_1, l_2 are the lengths as shown in Figure 4.3, depending on the positions of the branching points. For α small, the single tree has the lower costs, whereas for α close to 1, the two straight lines will be preferred. Thus, there exists a bifurcation point $\hat{\alpha}$ where both topologies have equal costs (depending on the distance d between the mass points). In the following, for fixed m and d, let $\alpha = \hat{\alpha}$ be chosen such that

$$\min_{\mathcal{F}\in\mathcal{A}_\mathcal{F}} \mathcal{M}^\alpha(\mathcal{F}) = \mathcal{M}^\alpha(\mathcal{F}_1) = \mathcal{M}^\alpha(\mathcal{F}_2)$$

and $\mathcal{F}_1, \mathcal{F}_2$ (where the exact structure of $\mathcal{F}_2$ depends on α) be the two optimal fluxes related to the two topologies (indeed, one can easily verify that such an α exists). Due to Theorem 4.1.9, we have

$$\min_{u\in\mathcal{A}_u} \tilde{\mathcal{M}}^\alpha(u) = \tilde{\mathcal{M}}^\alpha(u_1) = \tilde{\mathcal{M}}^\alpha(u_2),$$

where u_1, u_2 denote the images related to the mass fluxes $\mathcal{F}_1, \mathcal{F}_2$.
Now we define a functional $J : SBV(V \times \mathbb{R}) \to [0, \infty]$ as

$$J(v) = \sup_{\phi \in \mathcal{K}_1} \int_{\Omega \times \mathbb{R}} \phi \cdot \mathrm{d}Dv + \iota_{\mathcal{C}}(v)$$

and set $v = \lambda 1_{u_1} + (1 - \lambda) 1_{u_2}$ for some $\lambda \in [0, 1]$. Under the assumption that we can achieve

$$\tilde{\mathcal{M}}^\alpha(u_1) = \sup_{\phi \in \mathcal{K}_1} \int_{\Omega \times \mathbb{R}} \phi \cdot \mathrm{d}D1_{u_1}, \tag{4.5}$$

if u_1 and u_2 are minimizers of $\tilde{\mathcal{M}}^\alpha$, 1_{u_1} and 1_{u_2} are minimizers of the right-hand side. Since J is convex, it follows that

$$J(v) \leq J(1_{u_1}) = J(1_{u_2}).$$

If now we were able to construct a $\hat{\phi} \in \mathcal{K}_1$ which satisfies equation (4.5) and is additionally divergence-free, then Theorem 2.4.8 would yield

$$\min_{v \in \mathcal{C}} J(v) = \min_{u \in \mathcal{A}_u} \tilde{\mathcal{M}}^\alpha(u),$$

and as a consequence, $v = \lambda 1_{u_1} + (1 - \lambda) 1_{u_2}$ for every $\lambda \in [0, 1]$ would be a minimizer of J. Unfortunately, the task of finding such an optimal vector field $\hat{\phi}$ that incorporates the desired properties is a crucial issue which results in a rather technical construction. One approach, leading to an *almost optimal* vector field is deferred to Appendix A. Although we were not able to determine a truly optimal $\hat{\phi}$, we believe that it is nevertheless possible and as a consequence, the relaxed minimization problem in some cases yields the convex envelope (at least in some subset of the domain of J) of the original image-related problem. On the other hand J is not necessarily strictly convex, which is also reflected by some numerical examples (cf. Figure 4.4).

If the construction of an optimal $\hat{\phi}$ could be achieved, Example 4.2.1 would suggest that in a discrete setting, there exist at least some cases where the relaxed energy corresponds to the convex envelope of the original branched transport functional. Although still lacking a rigorous proof, by transferring the construction strategy locally to all (at least discrete) transportation networks, one might be able to extend this result to the more general setting. However, a further investigation of this issue goes beyond the scope of this thesis and might be a subject of future work.

Remark 4.2.2. In [23], the author investigates the addressed problem in a very general framework and is able to prove that in one space dimension and under certain assumptions, the Mumford–Shah-type functional admits an equivalent convex representation in the sense that the relaxed functional indeed equals the convex lower semi-continuous envelope of the original functional. As a consequence, if u is a minimizer of the latter, then 1_u is a minimizer of the relaxed problem. Unfortunately, the proof could so far not be extended

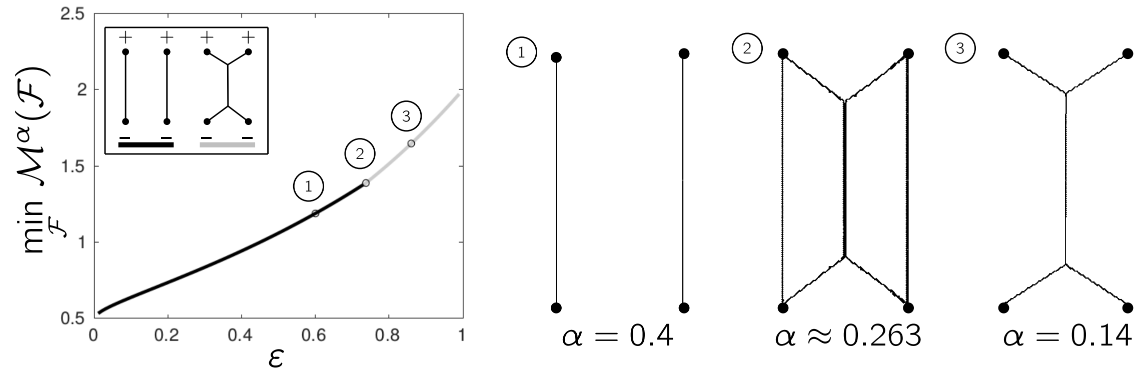


Figure 4.4.: Numerical examples for transport from two to two mass points as defined in Example 4.2.1. Left: Plot of the manually computed minimal energy for different $\varepsilon = 1 - \alpha$. The line type indicates the optimal network topology. Right: Numerically computed optimal fluxes for three different values of α. Example ② corresponds to the critical value of α, where $\mathcal{M}^{\alpha}$ has two minimizers. The numerical result corresponds to a linear combination of both minimizers.

to the case of non-scalar functions u such that it does not provide a solution for our case.

4.2.2. Reduction of the set $\mathcal{K}$ for piecewise constant functions

With regard to the numerical realization, the projection onto the convex set $\mathcal{K}$ is challenging. On the one hand, both sets contain for every ground point $x \in V$ an infinite number of inequality constraints. On the other hand, at first there exists an infinite number of ground points for which the sets of inequality constraints do not necessarily have to be independent. The following theorem shows that in case of piecewise constant functions along the lifted dimension, the number of inequality constraints for a fixed $x \in V$ is finite.

Theorem 4.2.3. *For any function $\phi : V \times [0,1] \to \mathbb{R}^2 \times \mathbb{R}$ which is piecewise constant in the third variable, i.e. $\phi(x_1, x_2, s) = C_i(x_1, x_2)$ for all $(x_1, x_2) \in V$, $s \in [ih_s, (i+1)h_s)$, $p \in \mathbb{N}$, $h_s = \frac{1}{p}$, $i = 0, \dots, p-1$, we define*

$$\begin{aligned}
\tilde{\mathcal{K}}_1 = \Big\{ &\phi = (\phi^x, \phi^s) \in L^\infty(V \times [0,1], \mathbb{R}^2 \times \mathbb{R}) \ : \ \phi^s \geq 0, \\
&\left| \int_{s_1}^{s_2} \phi^x\,(x, s)\, \mathrm{d}s \right| \leq |s_2 - s_1|^\alpha \ \forall x \in V, s_1, s_2 \in \{0, h_s, \dots, ph_s\} \Big\}, \\
\tilde{\mathcal{K}}_2 = \Big\{ &\phi = (\phi^x, \phi^s) \in L^\infty(V \times [0,1], \mathbb{R}^2 \times \mathbb{R}) \ : \ \phi^s \geq 0, \ |\phi^x| \leq a, \\
&\left| \int_{s_1}^{s_2} \phi^x\,(x, s)\, \mathrm{d}s \right| \leq \min\{|s_2 - s_1| + \varepsilon, a|s_2 - s_1|\} \ \forall x \in V, s_1, s_2 \in \{0, h_s, \dots, ph_s\} \Big\}.
\end{aligned}$$

Then we have

$$\begin{aligned} \phi \in \tilde{\mathcal{K}}_1 \;&\Rightarrow\; \phi \in \hat{\mathcal{K}}_1, \\ \phi \in \tilde{\mathcal{K}}_2 \;&\Rightarrow\; \phi \in \hat{\mathcal{K}}_2, \end{aligned}$$

where we define

$$\begin{aligned} \hat{\mathcal{K}}_1 = \Big\{&\phi = (\phi^x, \phi^s) \in L^\infty(V \times [0,1], \mathbb{R}^2 \times \mathbb{R}) \;:\; \phi^s \geq 0, \\ &\left|\int_{s_1}^{s_2} \phi^x(x,s)\,\mathrm{d}s\right| \leq |s_2 - s_1|^\alpha \;\forall x \in V, s_1, s_2 \in [0,1]\Big\}, \\ \hat{\mathcal{K}}_2 = \Big\{&\phi = (\phi^x, \phi^s) \in L^\infty(V \times [0,1], \mathbb{R}^2 \times \mathbb{R}) \;:\; \phi^s \geq 0,\; |\phi^x| \leq a, \\ &\left|\int_{s_1}^{s_2} \phi^x(x,s)\,\mathrm{d}s\right| \leq \min\{|s_2 - s_1| + \varepsilon, a|s_2 - s_1|\} \;\forall x \in V, s_1, s_2 \in [0,1]\Big\}. \end{aligned}$$

as the spaces $\mathcal{K}_1, \mathcal{K}_2$ without the smoothness requirement.

Proof. We denote by $t_i := ih_s$ for $i = 0, \dots, p$ the limit points of every constant region of ϕ in s-direction. Since $\phi \in \tilde{\mathcal{K}}_1$, respectively $\phi \in \tilde{\mathcal{K}}_2$, and $\phi(x_1, x_2, s) = C_i(x_1, x_2)$, we have

$$\left|\int_{t_i}^{t_j} \phi^x(x,s)\,ds\right| = \left|h_s \sum_{k=i}^{j-1} C_k(x_1, x_2)\right| \leq \begin{cases} |h_s(j-i)|^\alpha & \text{for (BT)}, \\ \min\{h_s(j-i) + \varepsilon, ah_s(j-i)\} & \text{for (UP)} \end{cases}$$

for all $t_i < t_j$ for all $x = (x_1, x_2)$. Due to the independence of the sets with respect to x, we need to show for a fixed $x = (x_1, x_2)$

$$\left|\int_{s_1}^{s_2} \phi^x(x,s)ds\right| \leq \begin{cases} |s_2 - s_1|^\alpha & \text{for (BT)}, \\ \min\{|s_2 - s_1| + \varepsilon, a\,|s_2 - s_1|\} & \text{for (UP)} \end{cases}$$

for arbitrary $s_1, s_2 \in [0,1]$. We will apply the following notation (cf. Figure 4.5):

- $C_i = C_i(x_1, x_2) \in \mathbb{R}^2$ for $s \in [t_i, t_{i+1})$, $i = 0, \dots, p-1$,
- $t_i \leq s_1 \leq t_{i+1}$, $t_{i+q-1} \leq s_2 \leq t_{i+q}$ for $q \leq p$,
- $h_1 = t_{i+1} - s_1$, $h_2 = s_2 - t_{i+q-1}$.

Let us now consider the branched transport and urban planning case separately.

Proof for branched transport:
We want to show

$$\left|h_1 C_i + h_s \sum_{k=i+1}^{i+q-2} C_k + h_2 C_{i+q-1}\right| \leq (h_1 + h_s(q-2) + h_2)^\alpha \tag{4.6}$$

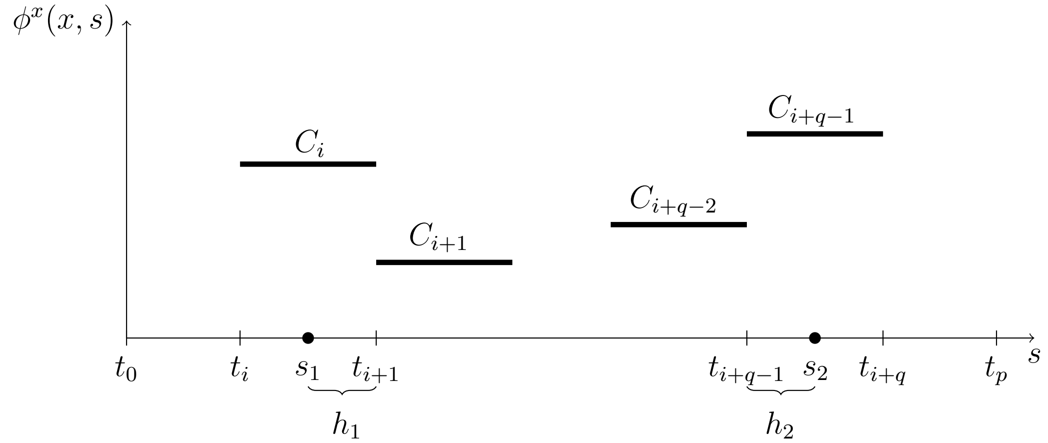


Figure 4.5.: Sketch of the profile of $\phi^x(x,s)$ for a fixed x as a one-dimensional function (note that $C_k \in \mathbb{R}^2$ for all k).

for $0 \leq h_1, h_2 \leq h_s$. We define

$$f(h_1,h_2) := \left|h_1 C_i + h_s \sum_{k=i+1}^{i+q-2} C_k + h_2 C_{i+q-1}\right|, \quad g(h_1,h_2) := -(h_1 + h_s(q-2) + h_2)^\alpha,$$

and show equivalently that $f(h_1,h_2) + g(h_1,h_2) \leq 0$ for all $0 \leq h_1, h_2 \leq h_s$. Set $h = (h_1,h_2)^T$, then f has the form $f(h) = F(Ah+b)$ with $F(x) = |x|$, a matrix $A = (C_i, C_{i+q-1})$ and a vector $b = h_s \sum_{k=i+1}^{i+q-2} C_k$. The function F is twice differentiable almost everywhere, thus we compute the Hessian of f as

$$D^2 f(h) = AD^2F(Ah+b)A^T.$$

Since F is convex, $D^2F(Ah+b)$ is positive semi-definite, thus $D^2f(h)$ is positive semi-definite and as a consequence, f is convex in h. Additionally, one can easily verify that g is twice differentiable and its Hessian has the form $D^2g(h) = \left(\begin{smallmatrix} c & c \\ c & c \end{smallmatrix}\right)$ for a $c \geq 0$, thus g is convex as well. Consequently, the function $f(h_1,h_2) + g(h_1,h_2)$ is convex on the domain defined by $[0,h_s] \times [0,h_s]$. Additionally, from the given constraints we have that $f(h_1,h_2) + g(h_1,h_2) \leq 0$ for the domain vertices $(h_1,h_2) \in \{(0,0),(h_s,0),(0,h_s),(h_s,h_s)\}$, which proves the desired statement.

Proof for urban planning:
Similar as before, we want to show

$$\left|h_1 C_i + h_s \sum_{k=i+1}^{i+q-2} C_k + h_2 C_{i+q-1}\right| \leq \min\{h_1 + h_s(q-2) + h_2 + \varepsilon, a(h_1 + h_s(q-2) + h_2)\}$$

for $0 \leq h_1, h_2 \leq h_s$. Defining f as before and $g(h_1,h_2) = -\min\{h_1 + h_s(q-2) + h_2 +$

$\varepsilon, a(h_1 + h_s(q-2) + h_2)\}$, it remains to show that g is convex on the domain $[0, h_s] \times [0, h_s]$, which can be achieved easily by applying the definition of convexity and some technical computation and case analysis. The proof then follows with the same argument as in the branched transport case. □

As a consequence, a piecewise constant approximation of the variables in the lifted direction makes sense and is easy to handle. Additionally, Theorem 4.2.3 does not hold for piecewise linear functions in s-direction, which can be shown by constructing a counterexample.

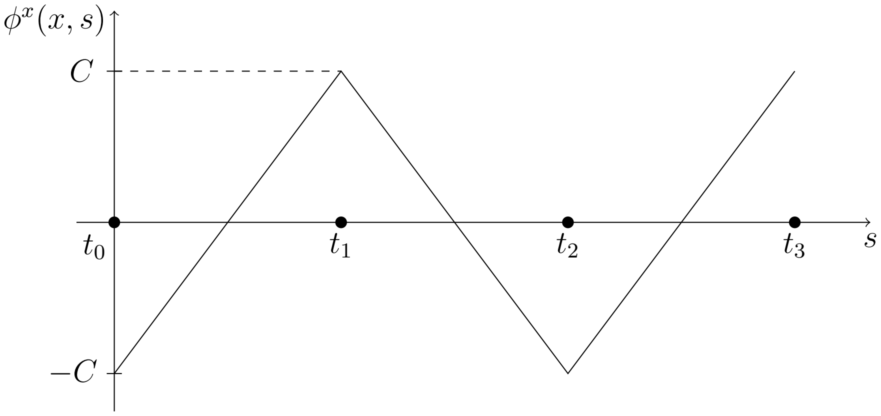


Figure 4.6.: Sketch of the function $\phi^x(x_1, x_2, s)$ for a fixed $(x_1, x_2) \in V$ as defined in Remark 4.2.4, $C = C(x_1, x_2)$.

Remark 4.2.4. Let $p \in \mathbb{N}$, $h_s = \frac{1}{p}$ and $t_i := ih_s$ for all $i = 0, \dots, p$. We define a function $\phi \in C(V \times [0,1], \mathbb{R}^2 \times \mathbb{R})$ as

$$\phi^x(x_1, x_2, s) = \begin{cases} \frac{2C(x_1,x_2)}{h_s}(s - t_i) - C(x_1, x_2) & \text{if } i \text{ even,} \\ \frac{2C(x_1,x_2)}{h_s}(t_i - s) + C(x_1, x_2) & \text{if } i \text{ odd} \end{cases}$$

for all $s \in [t_i, t_{i+1}]$, $(x_1, x_2) \in V$ for $C(x_1, x_2)$ independent of s (see Figure 4.6). Then it follows that

$$\left| \int_{t_i}^{t_{i+1}} \phi^x(x_1, x_2, s) \, \mathrm{d}s \right| = \left| \tfrac{1}{2} h_s (C(x_1, x_2) - C(x_1, x_2)) \right| = 0$$

and therefore

$$\left| \int_{t_j}^{t_l} \phi^x(x_1, x_2, s) \, \mathrm{d}s \right| = \left| \int_{t_j}^{t_{j+1}} \phi^x(x_1, x_2, s) \, \mathrm{d}s + \ldots + \int_{t_{l-1}}^{t_l} \phi^x(x_1, x_2, s) \, \mathrm{d}s \right| = 0$$

for all $j \leq l$, thus, all constraints between points t_j and t_l are satisfied. On the other hand, we have for $\tilde{s} = \frac{t_i + t_{i+1}}{2} \in [t_i, t_{i+1}]$

$$\left| \int_{t_i}^{\tilde{s}} \phi^x(x_1, x_2, s) \, \mathrm{d}s \right| = \tfrac{h_s}{4} \left| C(x_1, x_2) \right| ,$$

and the right-hand side can become arbitrarily large.

4.3. Numerical optimization with finite differences

In this section, we describe and discuss a finite difference discretization approach to the three-dimensional problem (4.4). This method comes along with several advantages: On the one hand, the interpretation of the variables as three-dimensional matrices makes the arising operators easy to handle, which keeps the implementation clear and simple. On the other hand, the problem naturally involves only a finite number of inequality constraints, since all variables are piecewise constant on voxels by definition (cf. Theorem 4.2.3). This also enables a straightforward communication between the two-dimensional image and its lifted counterpart.

4.3.1. Discretization

Let us for the sake of simplicity and without loss of generality assume $V \subset \mathbb{R}^2$ to be a rectangular domain with bottom left corner at the origin. Then we can discretize the domain $V \times \mathbb{R}$ by a finite three-dimensional $(n+1) \times (m+1) \times (p+1)$ grid

$$\mathcal{G} = \left\{ (ih_1, jh_2, lh_s) : \; i = 0, \dots, n, j = 0, \dots, m, l = 0, \dots p \right\} ,$$

where $h_1, h_2, h_s > 0$ denote the grid size in each direction. Then we can define the discrete counterparts of the variables $v \in SBV \left(V \times \mathbb{R}, [0,1] \right)$ and $\phi \in C_0^\infty \left(V \times \mathbb{R}, \mathbb{R}^2 \times \mathbb{R} \right)$ as $v^h : \mathcal{G} \to [0,1]$ and $\phi^h : \mathcal{G} \to \mathbb{R}^2 \times \mathbb{R}$. For every $(ih_1, jh_2, lh_s) \in \mathcal{G}$, we write $v_{ijl}^h = v^h \left(ih_1, jh_2, lh_s \right)$ and $\phi_{ijl}^h = \phi^h \left(ih_1, jh_2, lh_s \right)$. We discretize the gradient operator by forward finite differences,

$$\left(D_1 v^h \right)_{ijl} = \frac{v_{i+1,j,l}^h - v_{i,j,l}^h}{h_1}, \; \left(D_2 v^h \right)_{ijl} = \frac{v_{i,j+1,l}^h - v_{i,j,l}^h}{h_2}, \; \left(D_s v^h \right)_{ijl} = \frac{v_{i,j,l+1}^h - v_{i,j,l}^h}{h_s}$$

and set $D = \left(D_1, D_2, D_s \right)^T$. Hence, the discrete form of the saddle point problem for branched transport and urban planning defined in (4.4) reads

$$\min_{v^h \in \mathcal{C}^h} \max_{\phi^h \in \mathcal{K}^h} \left[\sum_{i,j,l} \phi_{ijl}^h \left(D v^h \right)_{ijl} = \langle \phi^h, D v^h \rangle \right] \tag{4.7}$$

where the discrete versions of the convex sets $\mathcal{C}$, $\mathcal{K}_1$ and $\mathcal{K}_2$ are given by

$$\begin{aligned}
\mathcal{C}^h &= \Big\{ v^h : \mathcal{G} \to [0,1] : \ v^h_{ij0} = 1, \ v^h_{ijp} = 0 \ \forall i,j, \ v^h = 1^h_{u(\mu_+,\mu_-)} \text{ on } \mathcal{G} \setminus \Big(\overline{\Omega} \times \mathbb{R}\Big) \Big\} \\
\mathcal{K}^h_1 &= \Big\{ \phi^h = (\phi^h_x, \phi^h_s) : \mathcal{G} \to \mathbb{R}^2 \times \mathbb{R} : \ \phi^h_s \geq 0, \\
&\qquad |h_s \textstyle\sum_{l=s_1}^{s_2} (\phi^h_x)_{ijl}| \leq h_s^\alpha |s_2 - s_1 + 1|^\alpha \ \forall i,j,s_1 \leq s_2 \Big\}, \\
\mathcal{K}^h_2 &= \Big\{ \phi^h = (\phi^h_x, \phi^h_s) : \mathcal{G} \to \mathbb{R}^2 \times \mathbb{R} : \ \phi^h_s \geq 0, \ |\phi^h_x| \leq a, \\
&\qquad |h_s \textstyle\sum_{l=s_1}^{s_2} (\phi^h_x)_{ijl}| \leq \min\{h_s|s_2 - s_1 + 1| + \varepsilon, a h_s |s_2 - s_1 + 1|\} \ \forall i,j,s_1 \leq s_2 \Big\}.
\end{aligned}$$

Above, $1^h_{u(\mu_+,\mu_-)}$ denotes the discretization of the function $1_{u(\mu_+,\mu_-)}$ with respect to the grid $\mathcal{G}$. Note that in $\mathcal{K}^h_1$ and $\mathcal{K}^h_2$ the infinite number of constraints has now reduced to a finite number of inequalities, which was shown by exploiting the piecewise constancy in Theorem 4.2.3.

4.3.2. Algorithm

Since the optimization problem already has a classical saddle point form, a straightforward choice is the well-known first order primal-dual algorithm for convex problems initially introduced in [54] and further investigated in [26]. To this end, we write problem (4.7) as

$$\min_v \max_\phi \Big[\mathcal{L}(v,\phi) = \langle \phi, Dv \rangle + \iota_{\mathcal{C}^h}(v) - \iota_{\mathcal{K}^h}(\phi) \Big], \tag{4.8}$$

where we dropped the superscript h in the variables for the sake of readability. The proposed algorithm then alternatingly performs a gradient descent step in v and a gradient ascent step in ϕ, with an additional overrelaxation and step sizes τ and σ. Denoting by v^k and ϕ^k the k^{th} approximation of v and ϕ, the next iterates v^{k+1} and ϕ^{k+1} are computed as

$$\begin{cases}
\phi^{k+1} &= \mathcal{P}_{\mathcal{K}^h}\left(\phi^k + \sigma D \bar{v}^k\right), \\
v^{k+1} &= \mathcal{P}_{\mathcal{C}^h}\left(v^k - \tau D^* \phi^k\right), \\
\bar{v}^{k+1} &= v^{k+1} + \theta\left(v^{k+1} - v^k\right),
\end{cases}$$

starting with an initial approximation (v^0, ϕ^0), $\bar{v}^0 = v^0$. Here, $\mathcal{P}_{\mathcal{C}^h}$ (and $\mathcal{P}_{\mathcal{K}^h}$, respectively) denotes the orthogonal projection onto the convex set.

While the projection onto the set $\mathcal{C}^h$ is straightforward to implement, the projection onto $\mathcal{K}^h$ remains more challenging since it involves a large set of non-local constraints. We rewrite the convex sets $\mathcal{K}_1^h$ and $\mathcal{K}_2^h$ as

$$\mathcal{K}_1^h = \bigcap_{s_1 \le s_2} \mathcal{K}_1^{h,s_1,s_2},\ \mathcal{K}_2^h = \left(\bigcap_{s_1 \le s_2} \mathcal{K}_2^{h,s_1,s_2} \right) \cap \mathcal{K}_2^{h,a}$$

with

$$\begin{aligned}
\mathcal{K}_1^{h,s_1,s_2} &:= \Big\{ \phi = (\phi^x, \phi^s) : \mathcal{G} \to \mathbb{R}^2 \times \mathbb{R} :\ \phi^s \ge 0,\ |h_s \sum_{l=s_1}^{s_2} (\phi^x)_{ijl} | \le h_s^\alpha |s_2 - s_1 + 1|^\alpha\ \forall i,j \Big\},\\
\mathcal{K}_2^{h,a} &:= \Big\{ \phi = (\phi^x, \phi^s) : \mathcal{G} \to \mathbb{R}^2 \times \mathbb{R} :\ \phi^s \ge 0, |\phi^x| \le a \Big\}\\
\mathcal{K}_2^{h,s_1,s_2} &:= \Big\{ \phi = (\phi^x, \phi^s) : \mathcal{G} \to \mathbb{R}^2 \times \mathbb{R} :\ \phi^s \ge 0,\\
&\qquad |h_s \textstyle\sum_{l=s_1}^{s_2} (\phi^x)_{ijl} | \le \min\{ h_s |s_2 - s_1 + 1| + \varepsilon, a h_s |s_2 - s_1 + 1| \}\ \forall i,j \Big\}.
\end{aligned}$$

The orthogonal projection onto each $\mathcal{K}^{h,s_1,s_2}$ can be computed directly. Since the integral inequality constraints are independent for each i,j, we can compute the projection for a fixed (i,j). For $l \notin \{s_1, \dots, s_2\}$, we set

$$(\mathcal{P}_{\mathcal{K}^{h,s_1,s_2}}(\phi^x))_{ijl} = (\phi^x)_{ijl}\,.$$

Now let $q := s_2 - s_1 + 1 \ge 1$ and define vectors $\psi^1, \psi^2, \theta^1, \theta^2 \in \mathbb{R}^q$ with

$$\psi^1 = \left(\phi_{ijl}^1\right)_{l=s_1,\dots,s_2},\ \psi^2 = \left(\phi_{ijl}^2\right)_{l=s_1,\dots,s_2},\ (\theta^1, \theta^2) = (\mathcal{P}_{\mathcal{K}^{h,s_1,s_2}}(\phi^x)_{ijl})_{l=s_1,\dots,s_2}\,.$$

Then we can write

$$(\theta^1, \theta^2) = \mathcal{P}_{\tilde{\mathcal{K}}}(\psi^1, \psi^2) = \operatorname*{argmin}_{w \in \tilde{\mathcal{K}}} |w - \psi|^2$$

with $\psi = (\psi^1, \psi^2)$ and

$$\tilde{\mathcal{K}} := \{ w \in (\mathbb{R}^q)^2\ :\ |h_s \sum_{l=1}^q w_l|^2 \le C^2 \}$$

for $C = h_s^\alpha |q|^\alpha$ for $\mathcal{K}_1^{h,s_1,s_2}$ and $C = \min\{h_s|q| + \varepsilon, a h_s |q|\}$ for $\mathcal{K}_2^{h,s_1,s_2}$. The corresponding optimality conditions for the minimization problem read

$$0 = \theta_k^r - \psi_k^r + \mu h_s \sum_{l=1}^q \theta_l^r,\ 0 = \mu \left(|h_s \sum_{l=1}^q \theta_l|^2 - C^2 \right),\ \mu \ge 0,\ \mu \in \mathbb{R}$$

for $r = 1, 2$. Summing up the first condition over all k yields

$$\sum_{k=1}^{q} \theta_k^r = \frac{1}{1 + q\mu h_s} \sum_{k=1}^{q} \psi_k^r,$$

which leads to an explicit formula for the θ_k^r as

$$\theta_k^r = \psi_k^r - \frac{\mu h_s}{1 + q\mu h_s} \sum_{l=1}^{q} \psi_l^r.$$

For μ, we have $\mu = 0$ if $|h_s \sum_{l=1}^{l} \theta_l|^2 < C^2$ or

$$|h_s \sum_{l=1}^{l} \theta_l|^2 - C^2 = 0 \;\Rightarrow\; \mu = \frac{h_s |\sum_{l=1}^{q} \psi_l| - C}{C h_s q}.$$

Together, we obtain

$$\mu = \min\left\{0, \frac{h_s |\sum_{l=s_1}^{s_2} \psi_l| - C}{C h_s q}\right\}. \tag{4.9}$$

Consequently, the projection of ϕ^x onto $\mathcal{K}^{h,s_1,s_2}$ can be computed component-wise as

$$\begin{aligned}(\mathcal{P}_{\mathcal{K}^{h,s_1,s_2}}(\phi^x))_{ijl} &= \phi_{ijl}^x - \left(\frac{\mu h_s}{1 + \mu h_s (s_2 - s_1 + 1)} \sum_{k=s_1}^{s_2} \phi_{ijk}^x\right) \chi_{l \in \{s_1, \ldots, s_2\}}, \\ (\mathcal{P}_{\mathcal{K}^{h,s_1,s_2}}(\phi^s))_{ijl} &= \max\{0, \phi_{ijl}^s\}\end{aligned}$$

with μ_{ij} as in (4.9) (depending on i, j).

For the projection onto the whole set $\mathcal{K}_1^h$, $\mathcal{K}_2^h$ respectively, we make use of an iterative approach known as Dykstra's projection method [15], which employs the fact that the set can be decomposed as described above. Suppose that we have a convex set $C = C_1 \cap \ldots \cap C_r$, where each C_l is closed and convex. Consider the sequence (x_{kl}) for any $0 \le k \le r, 0 \le l \le r$ defined by the following iteration process:

```
Set $x_{0,r} = \tilde{x},\ d_{0,1} = \ldots = d_{0,r} = 0$
for $k = 1, 2, 3, \ldots$
    $x_{k,0} = x_{k-1,r}$
    for $l = 1, \ldots, r$
        $x_{kl} = \mathcal{P}_{C_l}(x_{k,l-1} - d_{k-1,l})$
        $d_{kl} = x_{kl} - x_{k,l-1} + d_{k-1,l}$
    end
end
```

Then, the authors have shown that for any $1 \le k \le r$, the sequence converges strongly to

$x^* = \mathcal{P}_C(\tilde{x})$, i.e.

$$|x_{kl} - x^*| \to 0 \text{ for } k \to \infty.$$

Roughly spoken, Dykstra's projection method alternatingly projects onto all single sets containing only one integral inequality independently, after removing the previous increment from the last projection round.
Summarizing, the full primal-dual algorithm is given in Algorithm 4.

Algorithm 4 Primal-dual algorithm for urban planning and branched transport

```
function OptimalTransportNetworkFD($u^0$,$\tau$,$\sigma$,$\theta$)
    Set $v^0 = 1_{u^0}$, $\bar{v}^0 = v^0$, $\phi^0 = (\phi_1^0, \phi_2^0, \phi_s^0) = 0$
    while Not converged do
        $\phi^{k+1} = \mathcal{P}_{\mathcal{K}^h}(\phi^k + \sigma D\bar{v}^k)$
        $v^{k+1} = \mathcal{P}_{\mathcal{C}^h}(v^k - \tau D^*\phi^{k+1})$
        $\bar{v}^{k+1} = v^{k+1} + \theta(v^{k+1} - v^k)$
        $k \leftarrow k + 1$
    end while
end function
return $v^{end}$, $\phi^{end}$
```

4.3.3. Convergence of the algorithm

In [26], the authors prove convergence of the primal-dual algorithm for $\tau\sigma|D|^2 < 1$ and $\theta = 1$, where $|D|$ denotes the operator norm. However, the proof assumes that the single updating steps are computed *exactly*, which as a consequence requires convergence of the Dykstra subroutine for the projection onto $\mathcal{K}^h$ in our case. Since this projection is computationally expensive, we would like to restrict ourselves to only a few subiteration steps and briefly investigate the convergence behaviour of the primal-dual method in case of an inexact projection of ϕ.
This problem was handled in [58] in a rather general setting for different types of errors in the proximal updates. Picking up the notation introduced in their article, we define

$$y \approx_\delta \mathcal{P}_{\mathcal{K}^h}(x) \;:\Leftrightarrow\; |y - y^*| \leq \delta$$

for a $\delta > 0$ and $y^* = \mathcal{P}_{\mathcal{K}^h}(x)$ being the exact projection, where we allow that the computed projection lies within a δ-ball around the exact value. Note that this includes the case of an infeasible solution y, which does not necessarily lie in $\mathcal{K}^h$. We consider the following algorithm

$$\begin{aligned} \phi^{k+1} &\approx_{\delta_k} \mathcal{P}_{\mathcal{K}^h}(\phi^k + \sigma D(2v^k - v^{k-1})) \\ v^{k+1} &= \mathcal{P}_{\mathcal{C}^h}(v^k - \tau D^*\phi^{k+1}), \end{aligned} \tag{4.10}$$

which corresponds to a choice of $\theta = 1$ in the general method and set $v^{-1} = v^0$. In the following we briefly apply the method introduced in [58] to our saddle point problem and state their convergence result for this special case.

Since allowing an error within the first projection step implies that the iterate is not necessarily feasible, an explicit convergence rate in the sense of an upper bound on the primal-dual gap in each iteration cannot be achieved easily [58]. Instead, the authors suggest to shift the error in the first step to a slightly simpler error in the second update, resulting in a an estimate on the exact projection. Precisely, we let

$$\phi^{*,k+1} = \mathcal{P}(\phi^k + \sigma D(2v^k - v^{k-1}))$$

be the true projection, which gives

$$|\phi^{k+1} - \phi^{*,k+1}| \leq \delta_k$$

by definition. Setting $\psi^{k+1} := \phi^{k+1} - \phi^{*,k+1}$, for the second update we obtain

$$\begin{aligned} v^{k+1} &= \mathcal{P}_{\mathcal{C}^h}(v^k - \tau D^* \phi^{k+1}) \\ &= \mathcal{P}_{\mathcal{C}^h}(v^k - \tau D^*(\phi^{*,k+1} + \psi^{k+1})) \\ &= \mathcal{P}_{\mathcal{C}^h}(v^k - \tau D^* \phi^{*,k+1} - \tau D^* \psi^{k+1}). \end{aligned}$$

Set $d^{k+1} = \tau D^* \psi^{k+1}$, then d^{k+1} satisfies

$$|d^{k+1}| = |\tau D^* \psi^{k+1}| \leq \tau |D^*| \; |\psi^{k+1}| \leq \tau L \delta_k$$

with $L = |D|$. By defining

$$y \mathrel{\hat{\approx}_\delta} \mathcal{P}_{\mathcal{C}^h}(x) \;\; :\Leftrightarrow \;\; y = \mathcal{P}_{\mathcal{C}^h}(x + d) \text{ for a } |d| \leq \delta,$$

we can replace the algorithm in (4.10) by

$$\begin{aligned} \phi^{*,k+1} &= \mathcal{P}_{\mathcal{K}^h}(\phi^{*,k} + \sigma D(2v^k - v^{k-1})) \\ v^{k+1} &\mathrel{\hat{\approx}_{\tau L \delta_k}} \mathcal{P}_{\mathcal{C}^h}(v^k - \tau D^* \phi^{*,k+1}). \end{aligned}$$

Finally, we obtain the following result for $\mathcal{L}(v, \phi) = \langle \phi, Dv \rangle + \iota_{\mathcal{C}^h}(v) - \iota_{\mathcal{K}^h}(\phi)$.

Theorem 4.3.1. *Let $L = |D|$, $\tau, \sigma > 0$ such that $\sigma\tau L^2 + \tau\beta L < 1$ for $\beta \ll 1$ and set $V^N := \frac{1}{N}\sum_{k=1}^N v^k$, $\Phi^{*,N} := \frac{1}{N}\sum_{k=1}^N \phi^{*,k}$. Let (v^*, ϕ^*) be a saddle point of (4.8), then we have*

$$\mathcal{L}(V^N, \phi^*) - \mathcal{L}(v^*, \Phi^{*,N}) \leq \frac{1}{2\tau N}\left(|v^* - v^0| + \sqrt{\frac{\tau}{\sigma}}|\phi^* - \phi^{*,0}| + 2\tau L \sum_{k=1}^N \delta_k \right)^2 .$$

Proof. The proof is a simple application of [58], Theorem 4.9 and Corollary 4.27. □

4.3.4. Results

We implemented the algorithm described above in MATLAB© and simulated the transportation networks for different sets of sources and sinks. In a first test, we defined a simple geometric setting with four evenly-spaced sources of equal mass at the top of a rectangular domain Ω and four evenly-spaced sinks at the bottom. In order to test the reliability of the proposed method, we computed the resulting optimal network for different parameters α (a, ε respectively) by hand. Then we compared the numerical results with the true global minimizers. Figures 4.7 and 4.8 show that in almost every case, the algorithm converged to the correct solution, except for some boundary cases, where the energy gap between two different topologies is very small. This discrepancy might be caused by the vertical alignment in the grid, which slightly prefers vertical network structures over others.

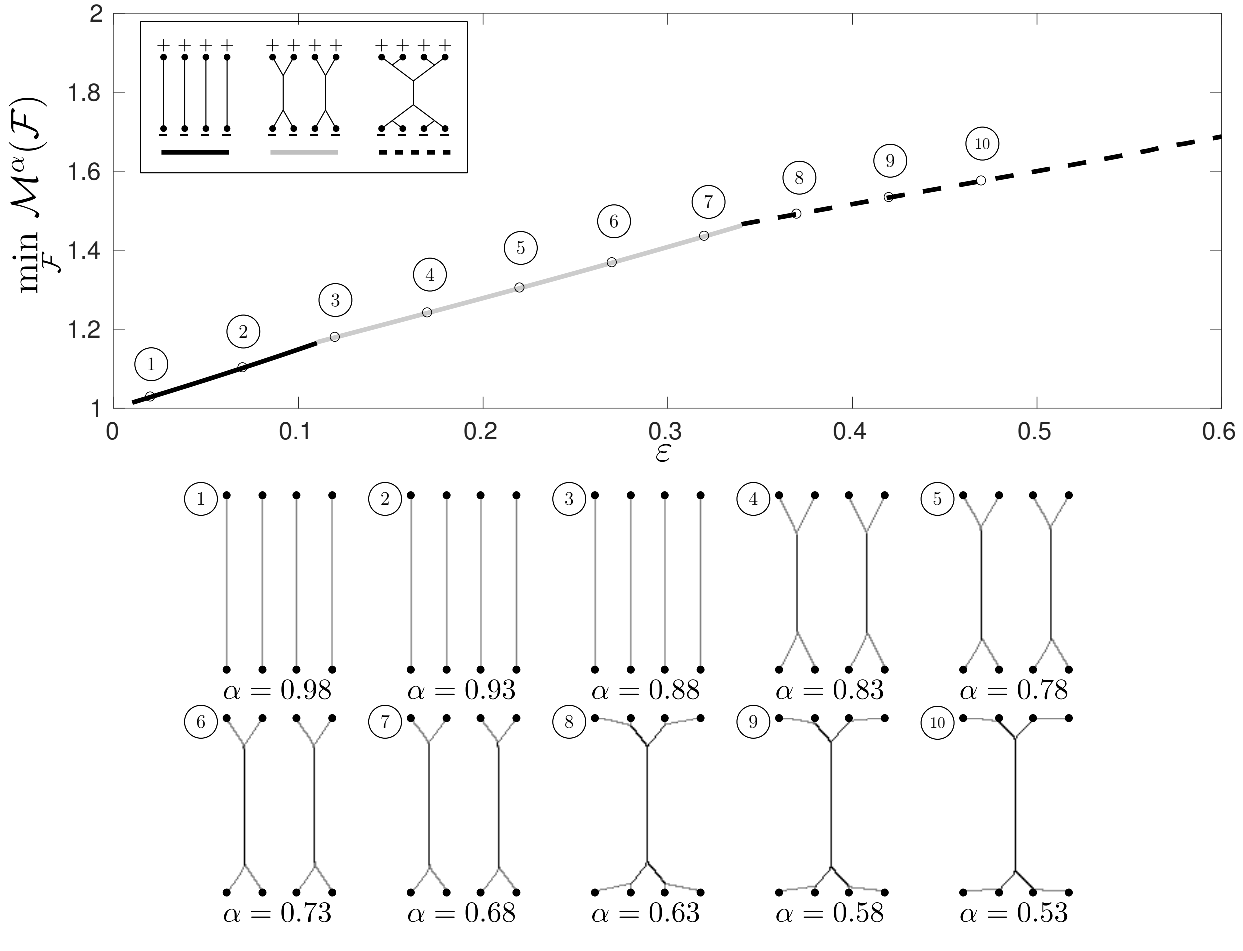


Figure 4.7.: Parameter study for branched transport. Top: Plot of the manually computed minimal energy for different values of $\varepsilon = 1 - \alpha$. The line type indicates the optimal network topology. Bottom: Numerically computed optimal fluxes for evenly spaced values of α in the same range. The numerically obtained network topologies match the predicted ones except for example ③.

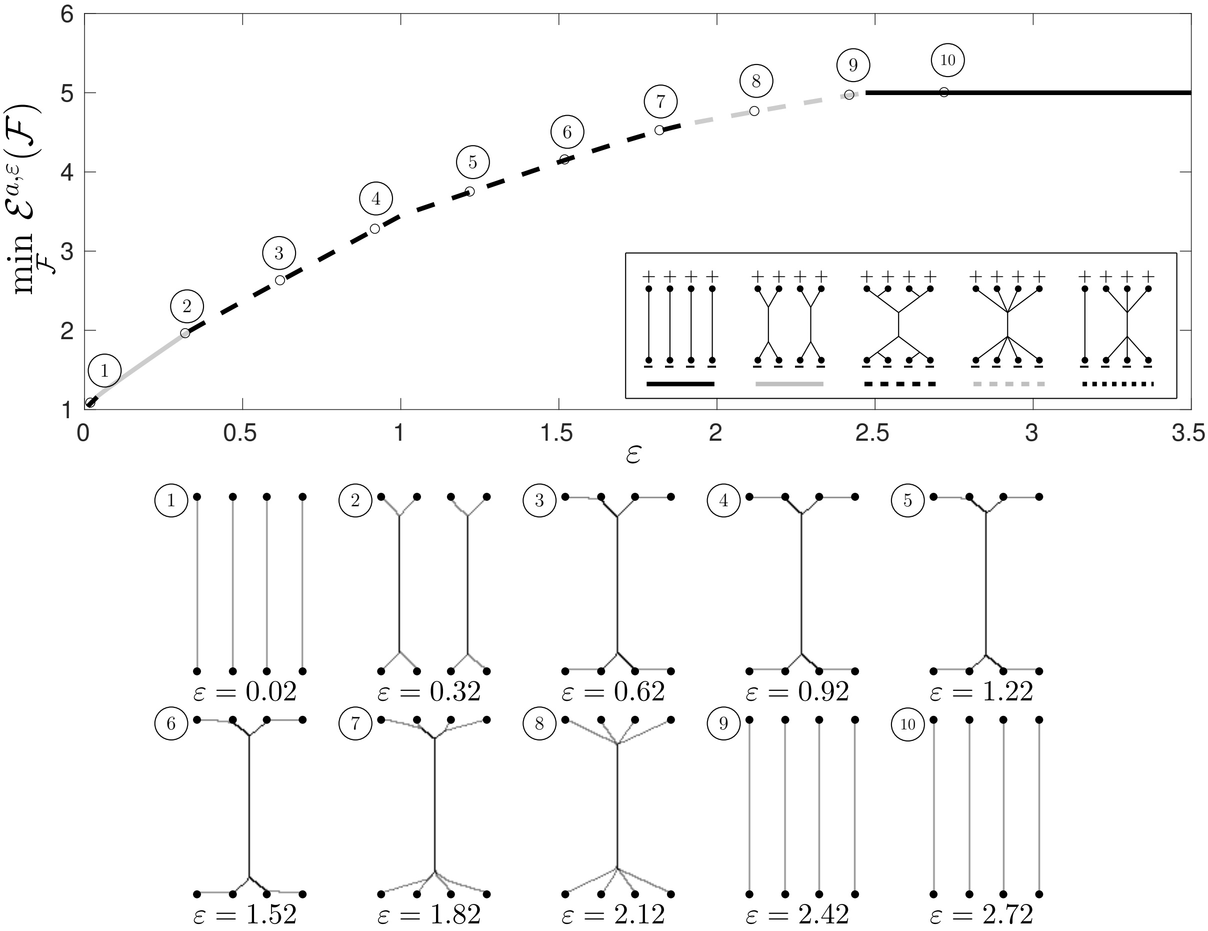


Figure 4.8.: Parameter study for urban planning. Top: Plot of the manually computed minimal energy for different values of ε and fixed $a = 5$. The line type indicates the optimal network topology. Bottom: Numerically computed optimal fluxes for evenly spaced values of ε in the same range. The numerically obtained network topologies match the predicted ones except for example ⑨. Note that the fifth topology is never optimal for this choice of $a = 5$, but can be for different values of a.

If the parameters are chosen such that the global optimal network lies very close to a bifurcation point, where the topology suddenly changes, the resulting variable v might not be binary. Figure 4.9 shows an example where v takes values in $\{0, 0.5, 1\}$. This behaviour probably results from the convex relaxation as shown in Section 4.2.1. If the optimal network is not unique for some choice of parameters, the optimal network obtained via the functional lifting approach might consist of a combination of two network topologies. Indeed, one can easily check that the non-binary solution in the example in Figure 4.9 consists of a convex combination of two topologies with equal costs (also shown in Figure 4.9).

In a second test, we simulated more complex branching structures in a test with 16 evenly-spaced sources and sinks respectively, where all points have equal mass. Since in

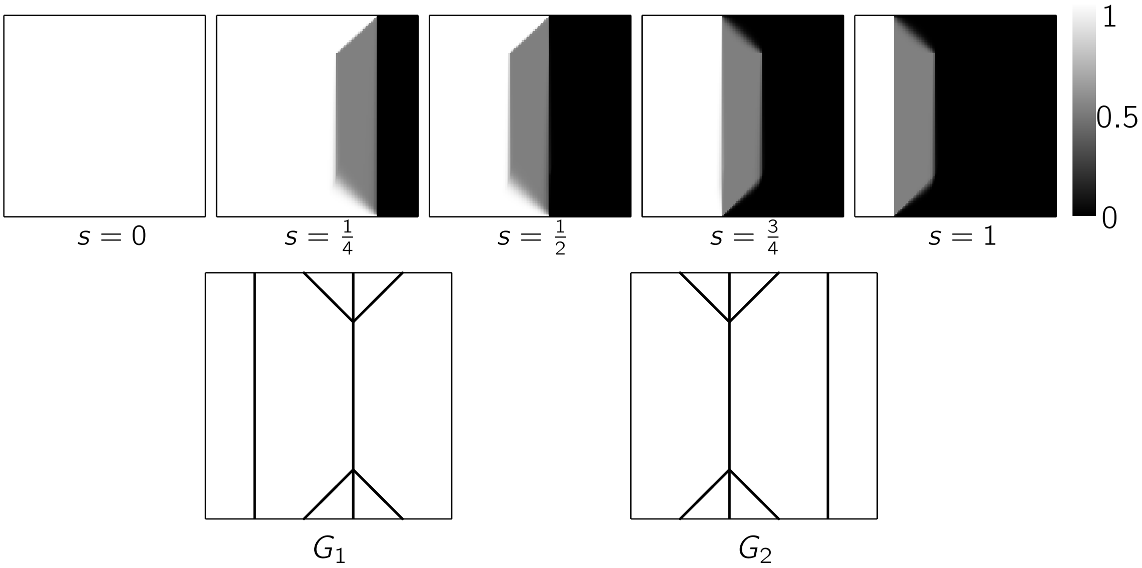


Figure 4.9.: Example of a numerical optimization for urban planning, resulting in a non-binary solution v (the images show different cross-sections). This indicates the effect of the convex relaxation for the chosen parameters ($a = 2.13$ and $\varepsilon = 0.5$). Bottom: Optimal topologies with exactly the same costs for the chosen parameters. The non-binary result is a convex combination of the two minimizers.

case of branched transport, the degree of branching is governed by the parameter α, one would expect more bifurcations in case of a small α and less for α being close to one. In urban planning, a small value of ε is expected to result in a higher number of single trees, whereas for increasing ε, the degree of branching should increase likewise. This effect is reflected by the numerical simulations in Figure 4.10.

For the functional lifting approach, we assumed for simplicity a rectangular image domain Ω. However, one can easily extend the approach to more general cases such as a circular shape. Figure 4.11 shows an example with some sources and sinks of different mass on the boundaries of a circle. Furthermore, the transport from a single source in the middle to 32 almost evenly-spaced points on the boundaries is displayed in Figure 4.12. Here, one has to face the additional difficulty that the initial and final measure have to satisfy spt μ_+, spt $\mu_- \subset \partial\Omega$. This can be achieved by setting the image domain $\Omega = B_1(0) \setminus \{0\}$ and assuming the image u to take values in S^1. Here, we used a periodic colour coding to visualize the image range of u.

4.3.5. Discussion

In this section, we presented a finite difference discretization approach to tackle the transportation problem arising from functional lifting of the branched transport and urban

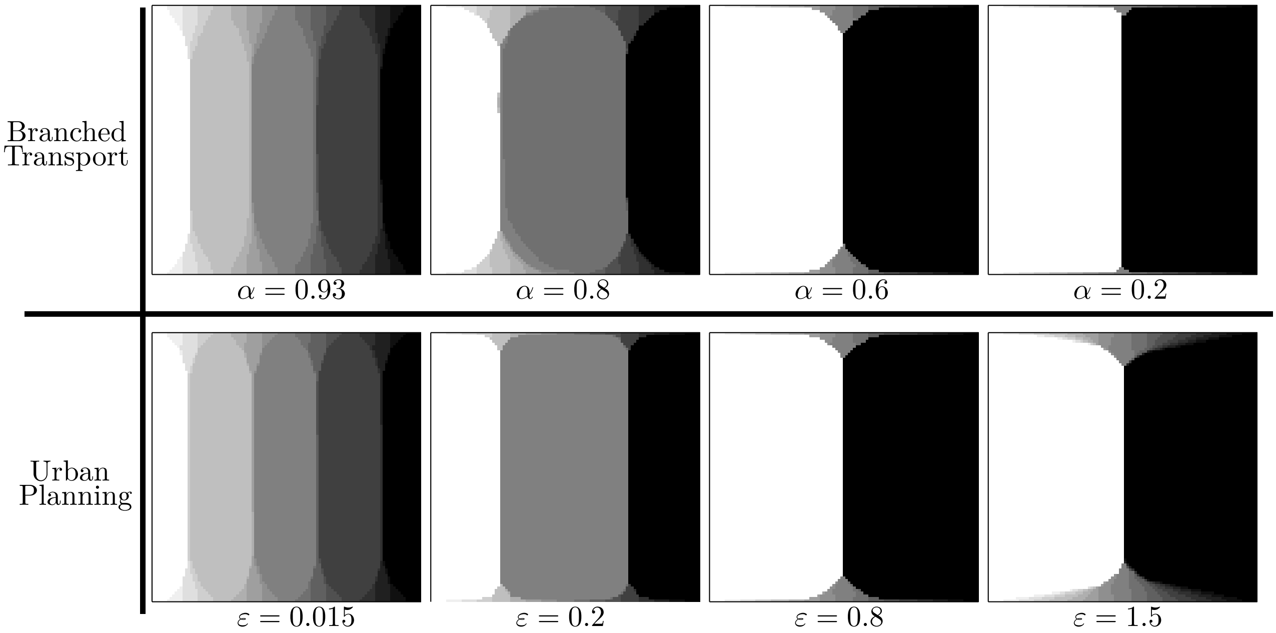


Figure 4.10.: Numerical optimization results for transport from 16 more or less evenly spaced point sources of same mass at the top to 16 evenly spaced point sinks (of same mass as well) at the bottom ($a = 5$ in case of urban planning). Instead of the optimal flux we show the corresponding optimal image u.

planning energies. We presented a simple algorithmic framework based on the well-known primal-dual method by [26] and, according to their proof, state convergence of the method even in case of an inexact projection. Finally, we presented some results obtained on different image domains.

A finite difference approximation is in most cases easy to handle and seems to be a natural choice in mathematical imaging since an image usually appears in the form of a matrix. This has the advantage, among others, that neighbouring relations of image pixels are directly obtained by the matrix form, which is a useful feature for the inequality constraint handling.

On the other hand, the matrix representation does not tackle the problem of the high-dimensionality arising from the functional lifting. Regarding the structure of the primal solution to the saddle point problem naturally suggests to adopt a locally varying image resolution, which is higher close to jump parts in order to define clear network pipes, and lower in constant regions. Compared to a uniform high resolution, this approach would decrease the number of degrees of freedom and, as a consequence, the runtime of the algorithm. Hence, we want to dissociate from the matrix form of the image and instead implement the ideas of adaptive finite elements.

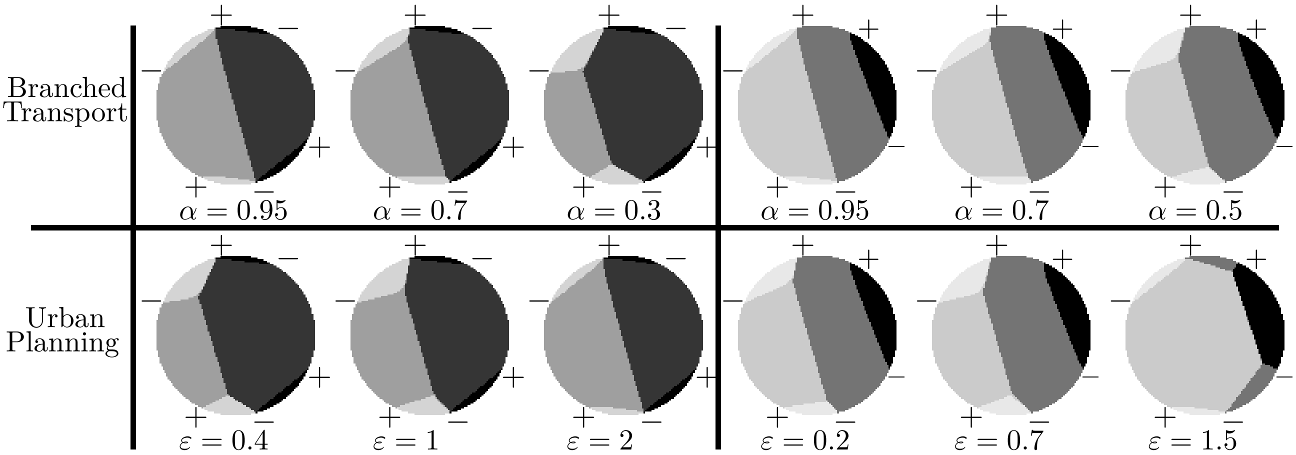


Figure 4.11.: Numerical optimization results for branched transport and urban planning with different parameters ($a = 5$ in case of urban planning). In the left column, the prescribed masses are $+\frac{1}{2}, -\frac{1}{8}, +\frac{1}{8}, -\frac{1}{2}, +\frac{1}{8}, -\frac{1}{8}$ (counterclockwise from top), in the right column $+\frac{1}{2}, -\frac{1}{8}, +\frac{1}{8}, -\frac{1}{2}, -\frac{1}{2}, +\frac{1}{2}$.

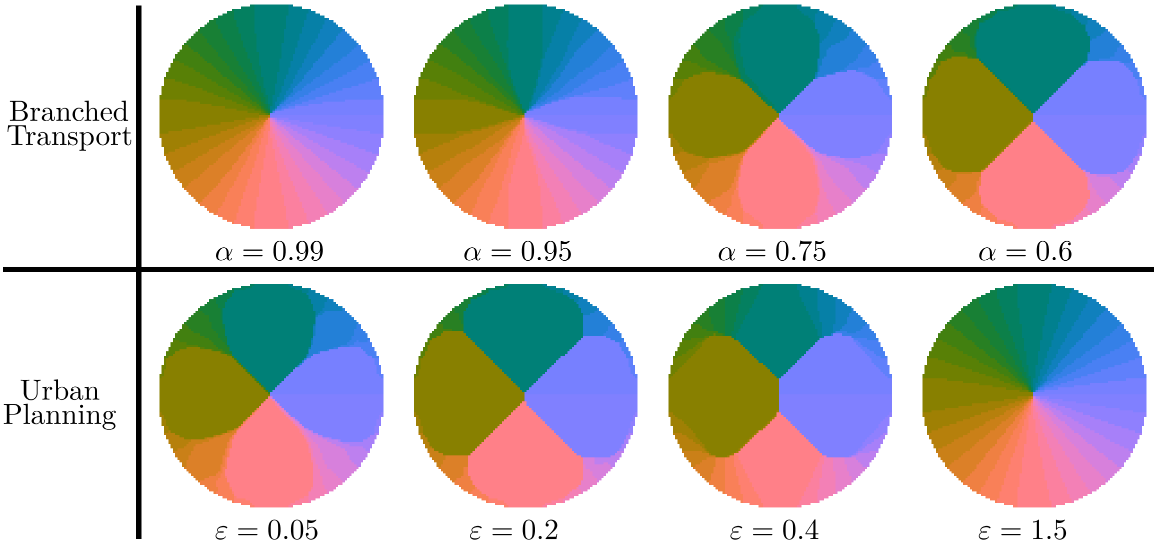


Figure 4.12.: Numerical optimization results for branched transport and urban planning for a single point source at the centre of the circular domain to 32 point sinks on the boundary ($a = 5$ in case of urban planning). The discontinuity set of the image corresponds to the optimal network. For this geometry, the image u takes values in S^1, which is here indicated by the periodic colour scale.

4.4. Numerical optimization with finite elements on adaptive triangular prism grids

In this section, we will describe and analyse in detail another approach where the discretization relies on a finite element scheme defined on an adaptive triangular prism grid as designed in Section 2.5 for the purpose of optimizing functional lifting problems. The basic idea is to overcome the disadvantages of functional lifting, like the high-dimensionality and the large number of inequality constraints, by a local grid refinement in image and lifted dimension independently. Hence, on the one hand, one can keep a low resolution in regions where the image is constant and a higher resolution at edges or affine parts. On the other hand, one hopes to reduce the large number of constraints to a minimum without violating those involving "non-grid points" (cf. Theorem 4.2.3).

Whereas the idea of adaptive finite element methods in general bears the possibility to vastly decrease the programme's runtime, it also entails some difficulties in our case. As mentioned before, a triangulation of the whole three-dimensional domain into tetrahedrons is not suitable due to the structure of the integral inequality constraints. In case of an adaptive discretization, a separation into rectangular elements automatically leads to the appearance of so-called hanging nodes, which (especially with regard to the constraint set) have to be treated carefully. In the second place, one has to think of a suitable refinement criterion which guarantees the convergence of the method to the same solution as in the non-adaptive case.

In the course of this section, we present a discretization of the lifted branched transport and urban planning formulations based on triangular prism elements as introduced in Section 2.5. We describe the algorithmic framework including the projection onto the convex set $\mathcal{K}^h$ and discuss different refinement criteria. Furthermore, we present some promising simulation results and discuss the advantages and drawbacks coming along with this type of discretization.

4.4.1. Discretization

In the following, we want to restrict ourselves to a three-dimensional image domain $[0,1]^3$. This implies that $V = [0,1]^2$ and the original image $u \in SBV(\Omega)$ only takes values in $[0,1]$ (note that this can be achieved for any real-valued image with bounded range by simple rescaling). In order to reuse the result of Theorem 4.2.3, we decide for piecewise constant Ansatz functions in the lifted dimension. Thus, we chose a semi-regular triangular prism grid $\mathcal{T}$ and a function space $S^{1,0}(\mathcal{T})$ (cf. Section 2.5).

For a basis $(\psi_1, \ldots, \psi_q)$ as defined in Section 2.5.2, the discrete solutions v^h, ϕ^h can be written in terms of basis functions as

$$v^h(x,s) = \sum_{k=1}^{q} \tilde{v}_k^h \psi_k(x,s),$$

$\phi_1^h, \phi_2^h, \phi_s^h$ respectively, for a coefficient vector $\tilde{v}^h \in \mathbb{R}^q$ ($\tilde{\phi}_1^h, \tilde{\phi}_2^h, \tilde{\phi}_s^h \in \mathbb{R}^q$), and where q denotes the total number of degrees of freedom. Hence we have $\tilde{v}(P_k) = \tilde{v}_k^h$ for all $P_k \in \mathcal{D}(\mathcal{T})$.

Based on this finite element discretization, we can now reformulate the convex saddle point problem (4.4) in terms of the coefficient vectors $\tilde{v}^h, \tilde{\phi}^h$:

$$\min_{\tilde{v}^h \in \mathcal{C}^h} \max_{\tilde{\phi}^h \in \mathcal{K}^h} \langle \tilde{\phi}^h, M\tilde{v}^h \rangle, \tag{4.11}$$

where $M = (M^1, M^2, M^s)^T$ denotes the mixed mass-stiffness matrix, i.e.

$$M_{kr}^1 = \int_{[0,1]^3} \psi_k \cdot \nabla_{x_1}\psi_r \, \mathrm{d}x\mathrm{d}s, \; M_{kr}^2 = \int_{[0,1]^3} \psi_k \cdot \nabla_{x_2}\psi_r \, \mathrm{d}x\mathrm{d}s, \; M_{kr}^s = \int_{[0,1]^3} \psi_k \cdot D_s\psi_r \, \mathrm{d}x\mathrm{d}s.$$

The gradient in s-direction D_s is interpreted in a finite difference sense, i.e.

$$D_s\psi_r(x,s) = \frac{\psi_r(x, s + h_{xs}) - \psi_r(x,s)}{h_{xs}} \tag{4.12}$$

with $h_{xs} := \{h(T) : (x,s) \in T\}$ being the height of the corresponding triangular prism element that contains (x,s).

In order to define the discrete constraint sets $\mathcal{C}^h$ and $\mathcal{K}^h$, we recall the definition of an s-line L_N for a ground node $N = (N^1, N^2, N^s) \in \mathcal{N}(\mathcal{T})$ (cf. Definition 2.5.16)

$$L_N = \{P = (P^1, P^2, P^s) \in \mathcal{N}(\mathcal{T}) \; : \; P^1 = N^1, \; P^2 = N^2\}$$

and set additionally

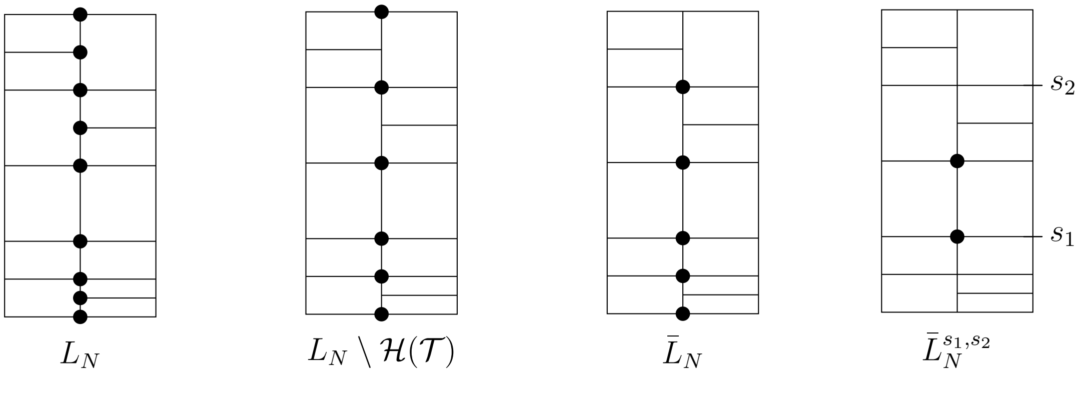


Figure 4.13.: Comparison between an s-line L_N (all nodes with the same xy-coordinates as the ground node N), $L_N \setminus \mathcal{H}(\mathcal{T})$ (s-line without hanging nodes), $\bar{L}_N$ (all degrees of freedom along L_N) and $\bar{L}_N^{s_1,s_2}$ (all degrees of freedom along L_N which lie in the interval $[s_1, s_2)$).

$$\bar{L}_N := L_N \cap \mathcal{D}(\mathcal{T}),\ \bar{L}_N^{s_1,s_2} := \bar{L}_N \cap [s_1, s_2)$$

and for every $P = (P^1, P^2, P^s) \in \bar{L}_N$, we define $h_P := \min_{\tilde{P} \in L_N \setminus \mathcal{H}(\mathcal{T}),\ \tilde{P}^s > P^s} |\tilde{P} - P|$ (see Figure 4.13). Then the constraint set $\mathcal{C}^h$, $\mathcal{K}_1^h$ and $\mathcal{K}_2^h$ can be defined as

$$\begin{aligned}
\mathcal{C}^h = \Big\{&\tilde{v}^h \in \mathbb{R}^q :\ \tilde{v}_k^h \in [0,1],\ \tilde{v}_k^h = 1 \text{ if } P_k^s = 0,\ \tilde{v}_k^h = 0 \text{ if } P_k^s = 1, \\
&\tilde{v}_k^h = 1_{u(\mu_+,\mu_-)}^h \text{ if } P_k \in \mathcal{D}(\mathcal{T}) \cap \partial\left([0,1]^2\right) \times [0,1]\Big\}, \\
\mathcal{K}_1^h = \Big\{&\tilde{\phi}^h = (\tilde{\phi}_x^h, \tilde{\phi}_s^h) \in (\mathbb{R}^q)^3 :\ \tilde{\phi}_s^h \geq 0, \\
&\Big|\sum_{P_k \in \bar{L}_N^{s_1,s_2}} h_{P_k} (\tilde{\phi}_x^h)_k\Big| \leq |s_2 - s_1|^\alpha\ \forall \text{ ground nodes } N \in \mathcal{N}(\mathcal{T})\ \forall\ s_1 < s_2\Big\}, \\
\mathcal{K}_2^h = \Big\{&\tilde{\phi}^h = (\tilde{\phi}_x^h, \tilde{\phi}_s^h) \in (\mathbb{R}^q)^3 :\ \tilde{\phi}_s^h \geq 0,\ |\tilde{\phi}_x^h| \leq a, \\
&\Big|\sum_{P_k \in \bar{L}_N^{s_1,s_2}} h_{P_k} (\tilde{\phi}_x^h)_k\Big| \leq \min\{|s_2 - s_1| + \varepsilon, a|s_2 - s_1|\} \\
&\quad \forall \text{ ground nodes } N \in \mathcal{N}(\mathcal{T})\ \forall\ s_1 < s_2\Big\},
\end{aligned}$$

where, as before, $(\tilde{\phi}_x^h)_k = \tilde{\phi}_x^h(P_k)$ for a $P_k \in \mathcal{D}(\mathcal{T})$.

Remark 4.4.1. In Section 2.5.2, we stated the independence of s-lines in the absence of xy-hanging nodes. Now that we defined the constraint sets $\mathcal{K}_1^h$ and $\mathcal{K}_2^h$ in terms of s-lines, we can benefit of this property: By setting

$$\begin{aligned}
\mathcal{K}_1^{h,s_1,s_2} := \Big\{&\tilde{\phi}^h = (\tilde{\phi}_x^h, \tilde{\phi}_s^h) \in (\mathbb{R}^q)^3 :\ \tilde{\phi}_s^h \geq 0, \\
&\Big|\sum_{P_k \in \bar{L}_N^{s_1,s_2}} h_{P_k} (\tilde{\phi}_x^h)_k\Big| \leq |s_2 - s_1|^\alpha\ \forall \text{ ground nodes } N \in \mathcal{N}(\mathcal{T})\Big\}, \\
\mathcal{K}_2^{h,s_1,s_2} := \Big\{&\tilde{\phi}^h = (\tilde{\phi}_x^h, \tilde{\phi}_s^h) \in (\mathbb{R}^q)^3 :\ \tilde{\phi}_s^h \geq 0, \\
&\Big|\sum_{P_k \in \bar{L}_N^{s_1,s_2}} h_{P_k} (\tilde{\phi}_x^h)_k\Big| \leq \min\{|s_2 - s_1| + \varepsilon, a|s_2 - s_1|\}\ \forall \text{ ground nodes } N \in \mathcal{N}(\mathcal{T})\Big\}, \\
\mathcal{K}_2^{h,a} := \Big\{&\tilde{\phi}^h = (\tilde{\phi}_x^h, \tilde{\phi}_s^h) \in (\mathbb{R}^q)^3 : |\tilde{\phi}_x^h| \leq a\Big\},
\end{aligned}$$

we obtain

$$\mathcal{K}_1^h = \bigcap_{s_1 < s_2} \mathcal{K}_1^{h,s_1,s_2},\ \mathcal{K}_2^h = \left(\bigcap_{s_1 < s_2} \mathcal{K}_2^{h,s_1,s_2}\right) \cap \mathcal{K}_2^{h,a},$$

thus, the projection onto the sets can be performed independently from each other, resulting in an increased efficiency of the overall method.

Remark 4.4.2. If the triangular prism grid is designed as a partition of the image domain

$[0,1]^3$, the set of degrees of freedom as defined in Section 2.5.2 does not contain any point with s-coordinate equal to 1, due to the fact that triangular prism elements must be formally defined as half-open in order to guarantee well-posedness of the function space $S^{1,0}(\mathcal{T})$. However, to include nodes with s-coordinate 1 in $\mathcal{D}(\mathcal{T})$, one can simply add another horizontal "slice" of prism elements such that the full grid contains the region $[0,1]^3$ as a proper subset.

4.4.2. Algorithm

Similar to the case of a finite difference discretization, we apply a primal-dual algorithm [26] to the discrete saddle point problem and perform the projection onto the convex set $\mathcal{K}^h$ via an iterative Dykstra routine [15].

Starting on a uniform low-resolution grid $\mathcal{T}_0$, we iterate until a certain convergence criterion is satisfied. Afterwards, some selected grid elements are refined with respect to a specified refinement criterion (which will be further discussed in Section 4.4.4) to obtain $\mathcal{T}_1$ and the solution (v^h, ϕ^h) is interpolated to the new set of degrees of freedom in $\mathcal{T}_1$. Starting with the result from the first round, we repeat the iteration on the new adaptive grid. For the sake of simplicity, in the following we will denote the coefficient vector with v^h, if ambiguity is beyond question.

The whole procedure is presented in Algorithm 5. The crucial difficulties of this method appear in the projection onto the convex set $\mathcal{K}^h$.

4.4.3. Projection onto $\mathcal{K}^h$

Let us have a closer look at the projection in the update of ϕ within Algorithm 5. In case of a full uniform grid $\mathcal{T}_0$, the projection routine stays the same as in case of the finite difference discretization (cf. Section 4.3). Since we maintained the piecewise constancy of the variables along s-direction, it is straightforward to show that, as before, it is sufficient to consider constraints between degrees of freedom (Theorem 4.2.3, with the simple extension that within one element the finite element function is linear in x, and one can easily show that for any $x \in [0,1]^2$ which does not coincide with a degree of freedom, the corresponding constraints are satisfied).

In case of an adaptive refinement $\mathcal{T}_t$, the situation is slightly changed by the occurrence of s-hanging nodes (we refer to Section 2.5 for a clear definition of a hanging node in case of a semi-regular triangular prism grid). As mentioned before, these nodes are not treated as degrees of freedom, but are interpolated from the finite element function within the corresponding element. Thus, these nodes are simply not considered within the discrete version of the constraint sets $\mathcal{K}_1^h, \mathcal{K}_2^h$. As a consequence, the projection routine roughly stays the same as in case of finite differences, where only the space between consecutive degrees of freedom within one s-line might vary (cf. Figure 4.14).

Remark 4.4.3. Note that the projection method would change significantly in the presence of xy-hanging nodes. On the one hand side, the independence of different s-lines is lost,

Algorithm 5 Adaptive primal-dual algorithm for urban planning and branched transport

```
function OptimalTransportNetworkFE($u^{start}$,$\tau$,$\sigma$,$\theta$,$numRefinements$)
    Set $v^{start} = 1^h_{u^{start}}$, $\phi^{start} = (\phi^{start}_1, \phi^{start}_2, \phi^{start}_s) = 0$
    for $run = 0, \dots, numRefinements$ do
        Set matrix $M = (M^1, M^2, M^s)^T$
        if run=0 then
            $v^0 = v^{start}$, $\phi^0 = \phi^{start}$
        else
            Interpolate results: $v^0 = Int(v^{lastRun})$, $\phi^0 = Int(\phi^{lastRun})$, $\bar{v}^0 = v^0$
        end if
        while Not converged do
            $\phi^{k+1} = \mathcal{P}_{\mathcal{K}^h}(\phi^k + \sigma M \bar{v}^k)$
            $v^{k+1} = \mathcal{P}_{\mathcal{C}^h}(v^k - \tau M^* \phi^{k+1})$
            $\bar{v}^{k+1} = v^{k+1} + \theta(v^{k+1} - v^k)$
            $k \leftarrow k + 1$
        end while
        $v^{lastRun} = v^{end}$, $\phi^{lastRun} = \phi^{end}$
        $v = v^{end}$, $\phi = \phi^{end}$
        if $run < numRefinements$ then
            Refine grid
        end if
    end for
end function
return $v, \phi$
```

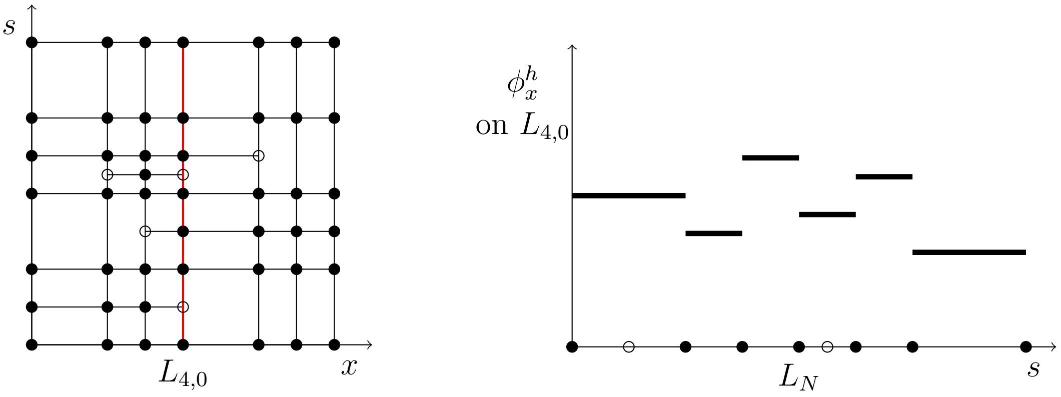


Figure 4.14.: Left: Two-dimensional adaptive grid with s-hanging nodes. All s-lines are independent of each other. Right: Possible values of ϕ^h_x on s-line L_N in the grid on the left. Hanging nodes are interpolated from the node below.

which causes the refinement routine to become much more inefficient since s-lines cannot be treated separately. On the other hand side, the function ϕ_x^h might have jumps in s-direction at points which are neither a degree of freedom nor a hanging node. Thus, it is no longer sufficient to test the integral inequality constraints between degrees of freedom only (not even if hanging nodes are included).

4.4.4. Refinement criteria

In the last step of Algorithm 5, certain elements of the current grid are refined. The goal of any adaptive refinement technique naturally is to achieve the same solution as computed on a fully uniform grid on the finest desired resolution. Hence, the choice of a suitable refinement criterion is crucial.

For different classes of partial differential equations, there have been several suggestions of local error estimates which admit lower and upper bounds for the true error under the assumption that the true solution is known (see for instance [64]). In the context of variational problems, a posteriori error estimates for uniformly convex energy functionals were introduced ([59], [10], among others). Unfortunately, the requirements for these error estimates do not hold in the case of the lifted branched transport and urban planning functionals.

Another natural and intuitive idea is to refine elements where the local gradient of the three-dimensional solution is high, in other words, where the solution is not close to being constant. Restricting ourselves to the variable v^h, we define

$$\eta_T(v^h) := \frac{1}{|T|} \int_T |D^h v^h| \, \mathrm{d}x \, \mathrm{d}s$$

with the operator $D^h = (\nabla_{x_1}, \nabla_{x_2}, D_s)^T$ as in (4.12) and refine elements T where

$$\eta_T(v^h) \geq \lambda \max_{S \in \mathcal{T}} \eta_S(v^h)$$

for a $\lambda \in [0,1]$. Although this strategy is computationally cheap and easy to handle, gradient refinement only takes the current grid structure into account and neglects any information about proximate steps (possibly leading to redundantly refined elements).

In order to obtain the same result as on a fully uniform high resolution grid $\bar{\mathcal{T}}$, one would need to compare the result on a locally refined grid $\mathcal{T}_t$ with the solution on $\bar{\mathcal{T}}$ and identify those regions where the local refinement needs to be improved. This strategy indeed would annihilate the advantages taken from grid adaptivity. In order to relax this approach, one can take one step backward and approximate a solution on a grid $\tilde{\mathcal{T}}_{t+1}$, which arises from $\mathcal{T}_t$ by refinement of every element in every direction. Instead of performing a complete primal-dual algorithm on $\tilde{\mathcal{T}}_{t+1}$, we want to exploit a local version of the primal-dual gap in order to identify elements which have to be refined.

To this end, we recall the definition of the primal-dual gap and afterwards define a localized version. Let X, Y be two Hilbert spaces and M be a continuous linear operator $M : X \to Y$.

Then we consider a general saddle point problem

$$\inf_{x\in X}\sup_{y\in Y}\langle Mx,y\rangle_Y + G(x) - F^*(y), \tag{4.13}$$

where $G : X \to \mathbb{R}$ and $F : Y \to \mathbb{R}$ are two convex, lower semi-continuous functions and F^* denotes the convex conjugate of F,

$$F^*(\tilde{y}) = \sup_{y\in Y}\Big(\langle \tilde{y},y\rangle_Y - F(\tilde{y})\Big).$$

The primal-dual gap for the saddle point problem (4.13) is given by

$$\Delta(x,y) := G(x) + F(Mx) + F^*(y) + G^*(-M^*y).$$

We can rewrite $\Delta(x,y)$ by inserting the definition of the convex conjugate and $F = F^{**}$ (since F is convex and lower semi-continuous) and obtain

$$\begin{aligned}\Delta(x,y) &= G(x) + F^{**}(Mx) + F^*(y) + G^*(-M^*y)\\ &= G(x) + \sup_{\tilde{y}\in Y}\Big(\langle \tilde{y},Mx\rangle_Y - F^*(\tilde{y})\Big) + \sup_{\tilde{x}\in X}\Big(\langle \tilde{x},-M^*y\rangle_X - G(\tilde{x})\Big)\\ &= \sup_{\tilde{y}\in Y}\Big(\langle \tilde{y},Mx\rangle_Y - F^*(\tilde{y}) + G(x)\Big) - \inf_{\tilde{x}\in X}\Big(\langle y,M\tilde{x}\rangle_Y + G(\tilde{x}) - F^*(y)\Big).\end{aligned}$$

For a primal-dual optimal pair (x^*,y^*), the primal-dual gap equals zero. Thus, the discrete primal-dual gap can be used as a stopping criterion for the convergence of a primal-dual algorithm. If the algorithm converged, the discrete solution $(x^h,y^h)\in X^h\times Y^h$ satisfies

$$\Delta^h(x^h,y^h) = \sup_{\tilde{y}\in Y^h}\Big(\langle \tilde{y},Mx^h\rangle_{Y^h} - F^*(\tilde{y}) + G(x^h)\Big) - \inf_{\tilde{x}\in X^h}\Big(\langle y^h,M\tilde{x}\rangle_{Y^h} + G(\tilde{x}) - F^*(y^h)\Big) = 0.$$

Here, $X^h\subset X$ and $Y^h\subset Y$ are finite-dimensional spaces, for instance the spaces of piecewise polynomial continuous functions as in the standard finite element case. Note that the occurring supremum (and the infimum, respectively) is taken in y^h (x^h) by definition of (x^h,y^h).

The idea is now to replace the function spaces X^h and Y^h in the occurring infimum and supremum term by larger function spaces $\tilde{X}^h$ and $\tilde{Y}^h$ with $X^h\subset\tilde{X}^h$, $Y^h\subset\tilde{Y}^h$. By this we obtain

$$\sup_{\tilde{y}\in\tilde{Y}^h}\Big(\langle \tilde{y},Mx^h\rangle_{\tilde{Y}^h} - F^*(\tilde{y}) + G(x^h)\Big) - \inf_{\tilde{x}\in\tilde{X}^h}\Big(\langle y^h,M\tilde{x}\rangle_{\tilde{Y}^h} + G(\tilde{x}) - F^*(y^h)\Big) \geq 0. \tag{4.14}$$

Since the subproblems are solved on a larger space, the variables $\tilde{x}$ and $\tilde{y}$ have more degrees of freedom and thus might yield a better result than x^h and y^h. In terms of adaptive finite element spaces, if all terms involved can be evaluated locally (i.e. on each element

independently) and if $\tilde{X}^h$ and $\tilde{Y}^h$ are chosen such that the underlying grid is a uniform refinement of the one from X^h and Y^h, (4.14) might be used to decide whether a finer element leads to a better solution in some sense.
To be more precise, let us go back to the saddle point problem (4.11). Let $(v, \phi) \in S^{1,0}(\mathcal{T}_t) \times (S^{1,0}(\mathcal{T}_t))^3$ be the discrete primal-dual optimal pair on an adaptive grid $\mathcal{T}_t$. If $\mathcal{T}_{t+1}$ is a refinement of $\mathcal{T}_t$, then obviously $S^{1,0}(\mathcal{T}_t) \subset S^{1,0}(\mathcal{T}_{t+1})$. For the sake of readability, we set $S_t := S^{1,0}(\mathcal{T}_t)$. Then, by (4.14) we have

$$\sup_{\tilde{\phi} \in S^3_{t+1}} \Big(\langle \tilde{\phi}, D^h v \rangle_{S^3_{t+1}} + \underbrace{\iota_{\mathcal{C}^h}(v)}_{=0} - \iota_{\mathcal{K}^h}(\tilde{\phi}) \Big) - \inf_{\tilde{v} \in S_{t+1}} \Big(\langle \phi, D^h \tilde{v} \rangle_{S^3_{t+1}} + \iota_{\mathcal{C}^h}(\tilde{v}) - \underbrace{\iota_{\mathcal{K}^h}(\phi)}_{=0} \Big) \geq 0,$$

where the second and last term vanish because of (v, ϕ) being an optimal pair. Assume that the supremum (the infimum, respectively) is taken in ϕ^{opt} (v^{opt}). Thus we obtain

$$\begin{aligned}
&\sup_{\tilde{\phi} \in S^3_{t+1}} \Big(\langle \tilde{\phi}, D^h v \rangle_{S^3_{t+1}} - \iota_{\mathcal{K}^h}(\tilde{\phi}) \Big) - \inf_{\tilde{v} \in S_{t+1}} \Big(\langle \phi, D^h \tilde{v} \rangle_{S^3_{t+1}} + \iota_{\mathcal{C}^h}(\tilde{v}) \Big) \\
&= \langle \phi^{opt}, D^h v \rangle_{S^3_{t+1}} - \langle D^{h*} \phi, v^{opt} \rangle_{S_{t+1}} + \langle \phi, D^h v \rangle_{S^3_t} - \langle \phi, D^h v \rangle_{S^3_t} \\
&= \langle \phi^{opt} - \phi, D^h v \rangle_{S^3_{t+1}} - \langle D^{h*} \phi, v^{opt} - v \rangle_{S_{t+1}} \\
&= \int_{\Omega \times \mathbb{R}} (\phi^{opt} - \phi) \cdot D^h v - D^{h*} \phi \cdot (v^{opt} - v) \, \mathrm{d}x \mathrm{d}s \\
&= \sum_{T \in \mathcal{T}_t} \int_T (\phi^{opt} - \phi) \cdot D^h v - D^{h*} \phi \cdot (v^{opt} - v) \, \mathrm{d}x \mathrm{d}s
\end{aligned}$$

where we added zero in the second line. Let us fix the following definition.

Definition 4.4.4 (Local primal-dual gap)**.** Let (v, ϕ), $\mathcal{T}_t$ and $\mathcal{T}_{t+1}$ as above, then for every element $T \in \mathcal{T}_t$ we define

$$\Delta_T(v, \phi) := \int_T (\phi^{opt} - \phi) \cdot D^h v - D^{h*} \phi \cdot (v^{opt} - v) \, \mathrm{d}x \mathrm{d}s,$$

with

$$\begin{aligned}
v^{opt} &= \operatorname*{argmin}_{\tilde{v} \in S_{t+1}} \Big(\langle \phi, D^h \tilde{v} \rangle_{S^3_{t+1}} + \iota_{\mathcal{C}^h}(\tilde{v}) \Big) \\
\phi^{opt} &= \operatorname*{argmax}_{\tilde{\phi} \in S^3_{t+1}} \Big(\langle \tilde{\phi}, D^h v \rangle_{S^3_{t+1}} - \iota_{\mathcal{K}^h}(\tilde{\phi}) \Big).
\end{aligned}$$

$\Delta_T(v, \phi)$ is called *local primal-dual gap.*

Note that v^{opt} and ϕ^{opt} live on a finer grid than v and ϕ.

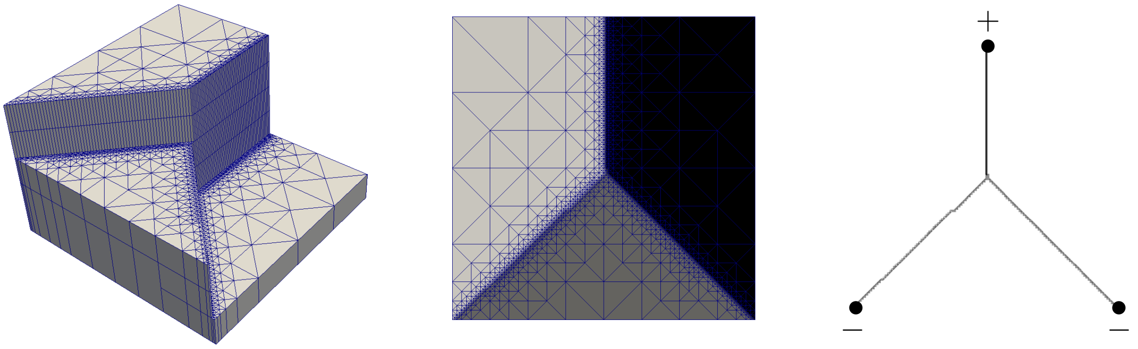


Figure 4.15.: Optimal network for branched transport from one mass point at the top to two mass points at the bottom of the domain. Left: Profile of the three-dimensional solution v. Middle: Two-dimensional solution obtained by delifting of v. Right: Optimal network structure.

4.4.5. Results

We implemented the algorithm described above in C++, where the grid and corresponding finite element classes are based on the QuocMesh library [57], where we added some additional features such as the adaptive prism grid and the corresponding operators. As in the case of finite differences, we simulated different network structures for both branched transport and urban planning problems. For all the results which will be presented in the following, we chose $\theta = 1$ and $\sigma = \tau = \frac{1}{|M|^2}$ with $|M|$ denoting the Frobenius norm of the finite element matrix M. Furthermore, the applied refinement criterion is a combination of large gradient and local primal-dual gap refinement as explained in Section 4.4.4.

In order to illustrate the adaptive grid structure in both three-dimensional and two-dimensional image, we created a simple example with one mass point at the top and two mass points at the bottom of the domain Ω and computed the optimal branched transportation network. Here, one can clearly see that the algorithm keeps a coarse resolution in regions where the lifted image remains constant and refines adaptively close to the edge set (see Figure 4.15).

To compare the results obtained on an adaptive triangular prism grid with the finite difference approach introduced in Section 4.3, we repeated the study of transport between four evenly-spaced points at the top and bottom of the domain (cf. Figures 4.7 and 4.8) for different parameter sets. The optimal networks are shown in Figures 4.16 and 4.17. As expected, most of the results equal the finite difference solutions except for example ③ in the branched transport experiments, where the parameter α lies close to a bifurcation point, where the topology of the optimal network changes. While the finite difference algorithm preferred four straight lines over the true optimal network, the method based on triangular prism grids converges to a three-dimensional solution which contains values of

0.5 and so does not correspond to a characteristic function of the subgraph of a piecewise constant two-dimensional image. This behaviour suggests that the minimizer of the relaxed energy functional can indeed be non-binary (cf. Section 4.2) and the discretization might suffer from some kind of grid bias.

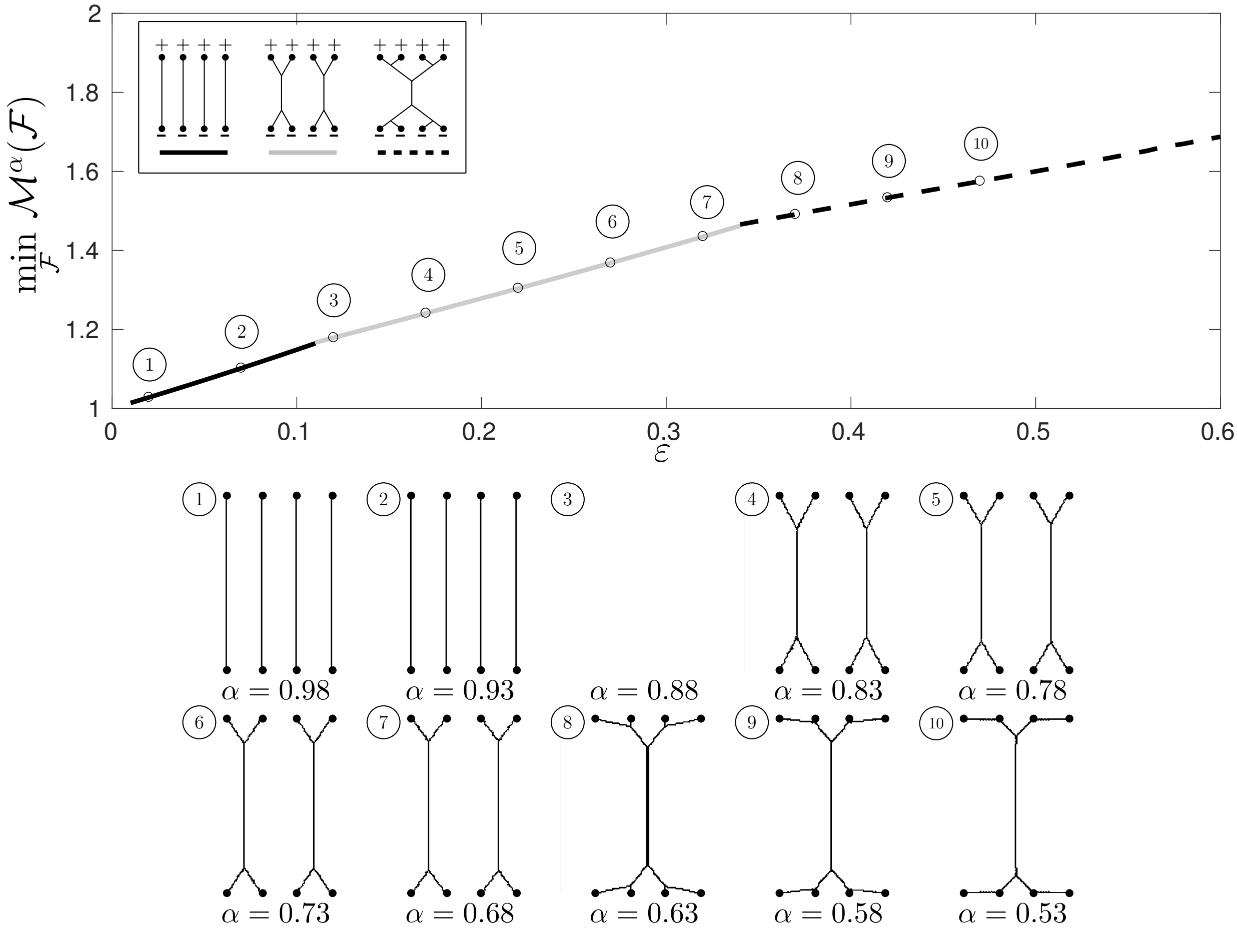


Figure 4.16.: Parameter study for branched transport. Top: Plot of the manually computed minimal energy for different values of $\varepsilon = 1 - \alpha$. The line type indicates the optimal network topology. Bottom: Numerically computed optimal fluxes for evenly spaced values of α in the same range. The numerically obtained network topologies match the predicted ones except for example ③, where the three-dimensional solution is not binary, but a convex combination of the binary solutions to the topologies consisting of four straight lines and three trees, where the latter possibly represents a local minimizer.

In order to simulate more complex branching structures, we repeated the test with 16 evenly-spaced sources and sinks respectively as well as the transport from a single source in the middle to 32 sinks on the boundaries of a circular domain (Figures 4.18 and 4.19). The higher complexity of the network structures requires a relatively high spatial resolution, which is a very crucial issue in case of the previously introduced uniform finite difference

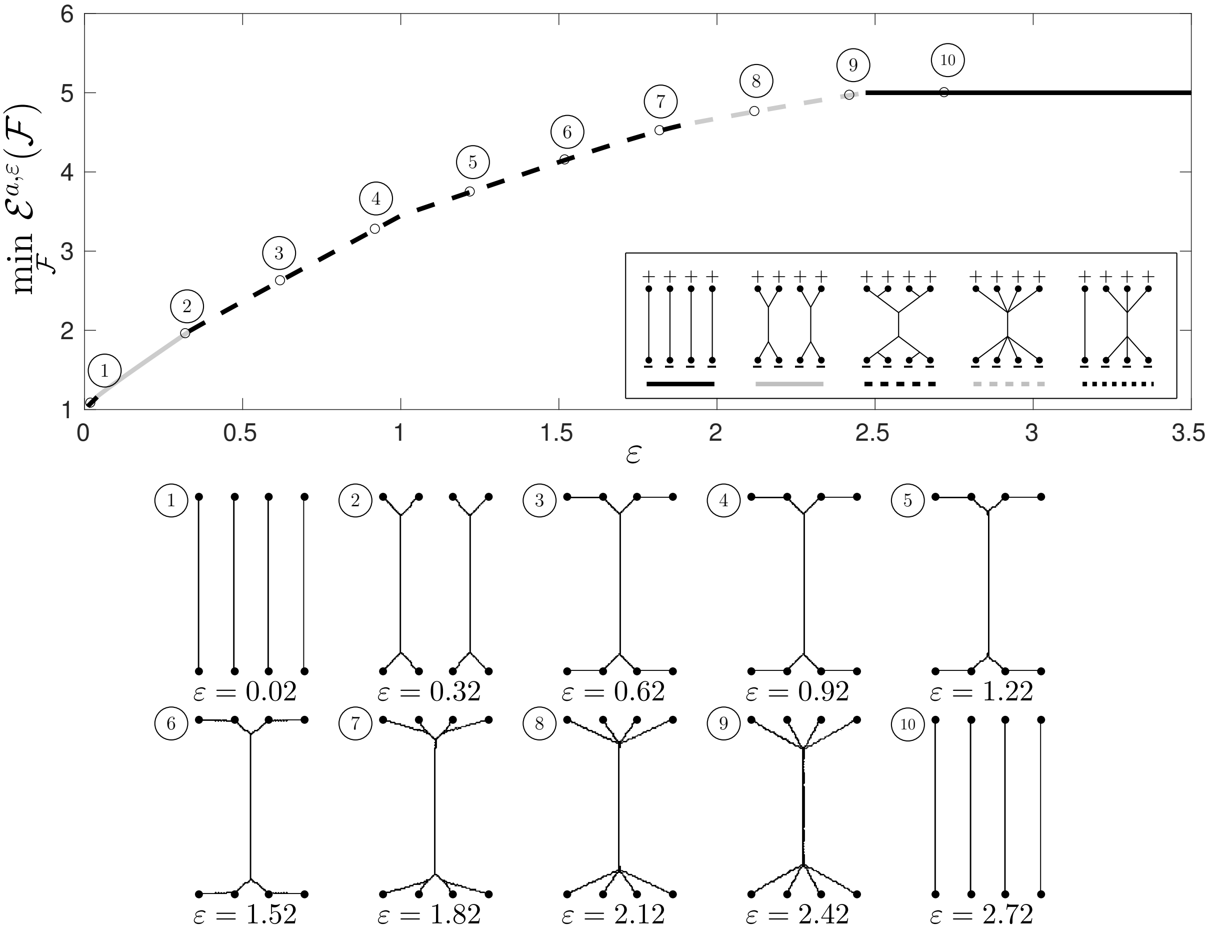


Figure 4.17.: Parameter study for urban planning. Top: Plot of the manually computed minimal energy for different values of ε and fixed $a = 5$. The line type indicates the optimal network topology. Bottom: Numerically computed optimal fluxes for evenly spaced values of ε in the same range. The numerically obtained network topologies match the predicted ones. Note that the fifth topology is never optimal for this choice of $a = 5$, but can be for different values of a.

discretization (cf. Figures 4.10 and 4.12). Due to the adaptive grid refinement, one can now easily achieve a local resolution of $(2^{10})^2$ grid nodes in two dimensions without drastically extending the overall computation time. We will discuss the impact of adaptivity in more details in Section 4.4.6.

A disadvantage of the functional lifting approach lies in the fact that the given measures μ_+, μ_- need to have support on the boundaries of the image domain. In case of transport from a single source in the middle to sinks on the boundaries of a circle, this can be overcome by additionally defining some parts of $\Omega \setminus \partial\Omega$ as "boundaries" and, as a consequence, reducing the inpainting region by fixing the image values in some elements. A similar trick can be applied in order to handle examples where both the given initial and final measure

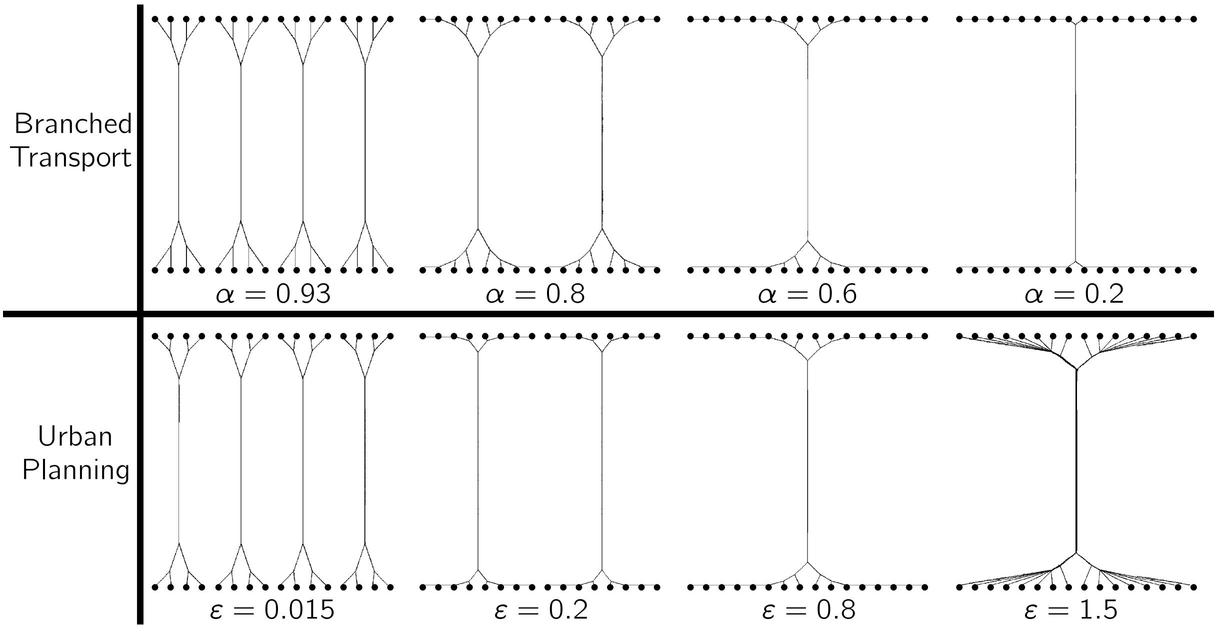


Figure 4.18.: Numerical optimization results for transport from 16 almost evenly spaced point sources to 16 point sinks of the same mass. For the urban planning results, we chose $a = 5$.

live in the interior of the image domain. In [13], the authors propose a method where an initial backwards transport path from μ_- to μ_+ is predefined and fixed during the iteration process. In terms of the functional lifting approach, this is equivalent to fixing some parts of the interior of Ω which correspond to the backwards path $\tilde{\Sigma}$ as "boundaries". This procedure allows an interpretation of the desired transport network as a circular flow, where only the forward part is optimized. We made use of the approach in order to simulate the transport form one source to two, three, four or five evenly distributed sinks, respectively. We approximated the Steiner tree problem by choosing a value of α close to zero in case of branched transport and compared the results with the urban planning network structure for a cost functional which is affine in the transported mass (see Figure 4.20). In the later course of this work, we repeat this experiment making use of a phase field approximation approach described in Chapter 5.

4.4.6. Uniform versus adaptive grid

In the course of this chapter, we introduced an adaptive grid approach to tackle the functional lifting problem arising from the branched transport and urban planning problem.

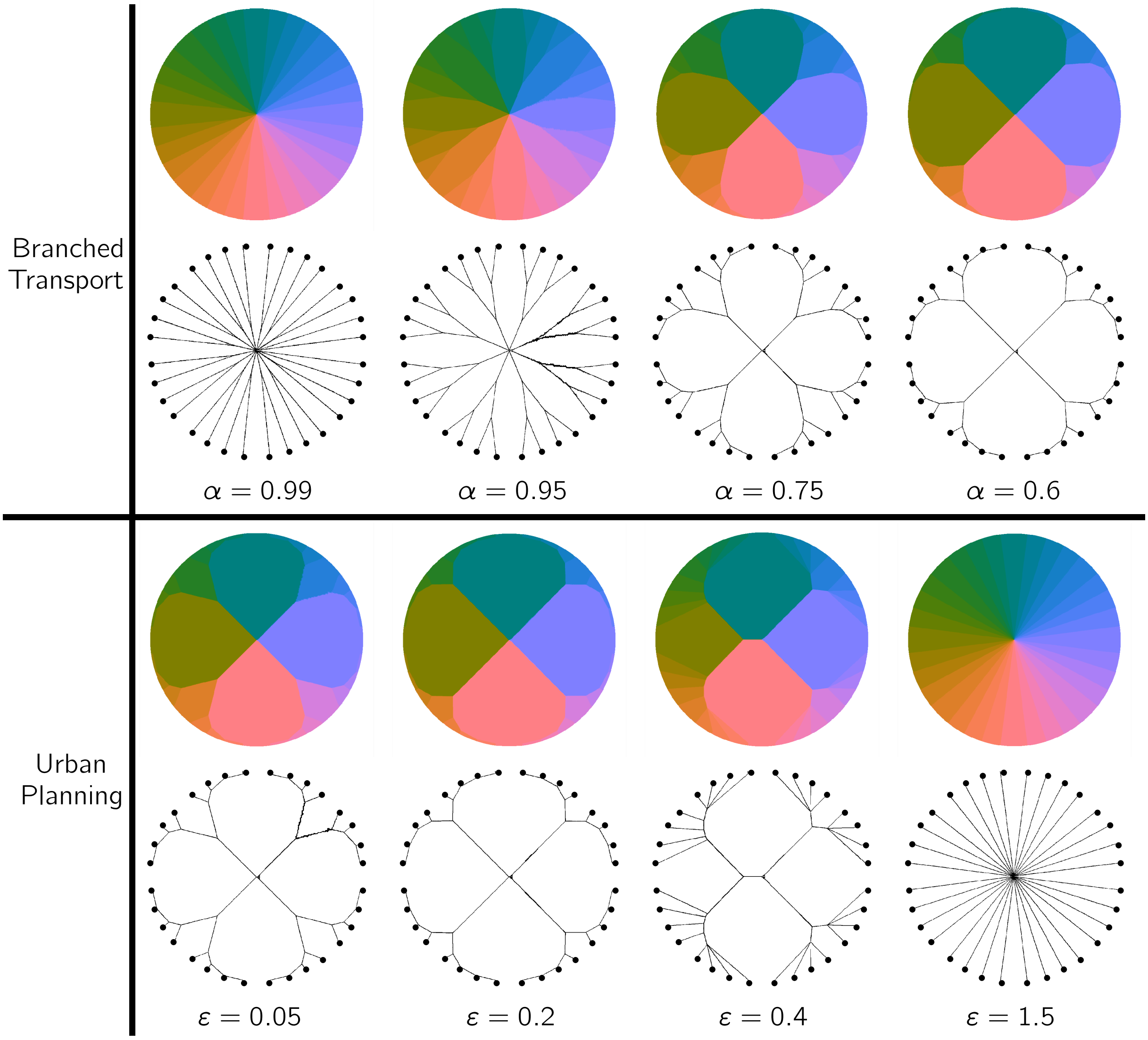


Figure 4.19.: Numerical optimization results for transport from one source in the middle of a circle to 32 almost evenly spaced point sinks of the same mass on the boundaries. For the urban planning results, we chose $a = 5$. The two-dimensional results are displayed using a periodic colour coding, the discontinuity set corresponds to the transportation network.

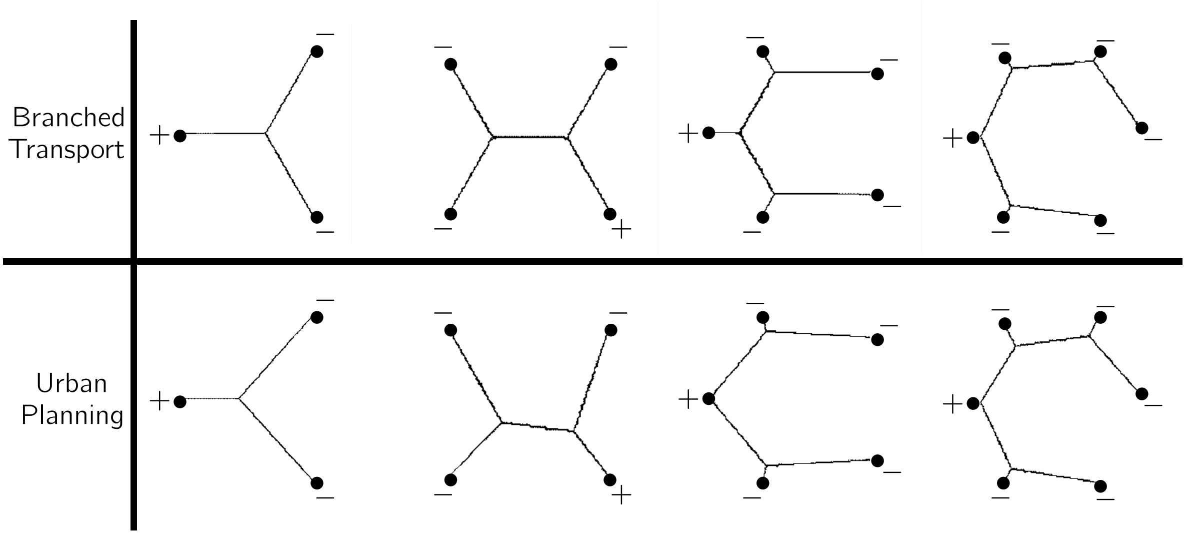


Figure 4.20.: Numerical optimization results from one point source to two, three, four or five point sinks respectively. The result for branched transport correspond to the parameter $\alpha = 0.001$ and for the urban planning results we chose $a = 10^{10}$, $\varepsilon = 1$.

It thus remains to accumulate evidence that the presented ideas are suitable in some sense, that is

(a) the obtained adaptive grid solution should qualitatively equal the solutions obtained on a fully uniform grid with the same resolution,

(b) the programme runtime is severely decreased by adaptivity depending on the problem size.

Concerning equality of the solutions, a comparison of the result in Section 4.3.4 using the finite difference method on a uniform grid with those presented in Section 4.4.5 already suggests that there is only a minor grid structure effect on the qualitative behaviour of the optimal network.

In order to reduce the comparison to local grid size effects (and thus exclude any effect arising from the discretization method), we performed some additional numerical examples using finite elements on triangular prism grids for both adaptive and uniform experiments. Figure 4.21 displays the results for two different branching parameters. As expected, although the grid structure is significantly different in some regions, it roughly coincides along the transportation network edges and as a consequence, the resulting solutions look very much alike.

In order to compare the efficiency of the adaptive grid approach, we performed numerical experiments for a particular example on multiple different problem sizes. As a test case, we used the example of branched transport from four to four mass points with a branching

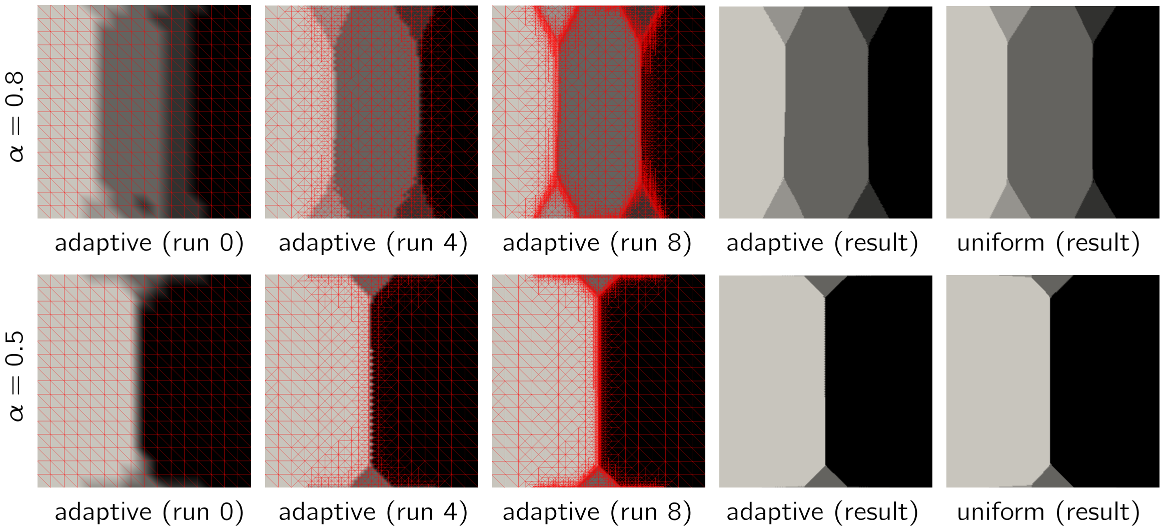


Figure 4.21.: Comparison between between an optimal network obtained on an adaptive and a uniform grid for branched transport from four to four mass points with different branching parameters. Columns one to three show the result on an adaptive grid after a certain number of refinement steps, columns four and five compare the result (without showing the grid) on the resulting adaptive grid and the corresponding uniform grid with the same final resolution.

parameter $\alpha = 0.5$, such that a transportation network consisting of a single tree is clearly favoured (cf. Figure 4.16). In order to decrease the runtime even further, we additionally applied a preconditioning step in each refinement round by adjusting the finite element operators with locally varying step sizes as described in [25]. In the adaptive case, after every iteration round we computed the absolute value of the local gradient of the primal variable and refined all elements where the value is higher than 15% of the overall maximum. The number of iterations in each round was fixed to 10000 in the primary rounds and 100000 in the last adaptive round and in the uniform case respectively. The maximal number of projection iterations was set to 100, where the projection stops if the dual solution satisfies every constraint. In order to quantify the final results, we computed the primal-dual gap after the iteration process was completed and compared the resulting transport network with the underlying ground truth.
We tested the result obtained on a fully uniform grid on different resolutions with the corresponding adaptive result, where we distinguish between refinement of the image domain (xy) only, the image range (s) only and both at the same time. The grid resolution is described by its *level*, where a uniform grid of xy-level $l_1 \in \mathbb{N}$ and s-level $l_2 \in \mathbb{N}$ consists of $2^{2l_1+1} \cdot 2^{l_2}$ prism elements in total. In case of adaptivity, we started with a fixed low resolution and performed as many refinement rounds as necessary to obtain the same minimal element size as in the uniform case (note that two refinement rounds are necessary to proceed from a xy-level l_1 grid to an adaptive xy-level $l_1 + 1$ grid, since xy-refinement

	Uniform				Adaptive						
xy	numEls	numDofs	time	error	runs	numEls	numDofs	%Els	%Dofs	time	error
4	4096	2601	26 sec.	0.0225	0	4096	2601	100	100	26 sec.	0.0225
5	16384	9801	96 sec.	0.0192	2	8400	5112	51.3	52.2	53 sec.	0.0096
6	65536	38025	375 sec.	0.0165	4	15856	9486	24.2	24.9	101 sec.	0.0048
7	262144	149769	1333 sec.	0.0017	6	32904	19431	12.6	12.9	211 sec.	0.0022
8	1048576	594441	7769 sec.	0.0009	8	70096	40914	6.7	6.9	441 sec.	0.0012
9	4194304	2368521	28042 sec.	0.0007	10	142976	82881	3.4	3.5	987 sec.	0.0007
10	-	-	-	-	12	347800	200025	2.1	2.1	2448 sec.	0.0003

Table 4.1.: Comparison between uniform and adaptive grid iteration for a fixed s-resolution of level 3, where only the image domain (xy) is refined. For each adaptive experiment, the initial xy-level is 4. The first column refers to the xy-level of the uniform grid and highest local xy-resolution of the final adaptive grid respectively. The computed error equals the primal-dual gap in the last iteration. The experiment on a uniform grid of xy-level 10 resolution is omitted due to its high runtime and memory consumption.

is accomplished via bisection of an element). For all results, we stated the number of elements, the number of degrees of freedom (dofs), the overall runtime and the final error corresponding to a numerically computed primal-dual gap using the MOSEK optimization software [5]. All experiments were executed on an Intel Core i7 6 × 3.60 GHz CPU with hyper-threading, where the projection onto the convex constraint set $\mathcal{K}_1$ is parallelized. The results are displayed in Tables 4.1 to 4.3. Figures 4.22 to 4.24 visualize the overall runtime for different problem sizes on a logarithmic scale.

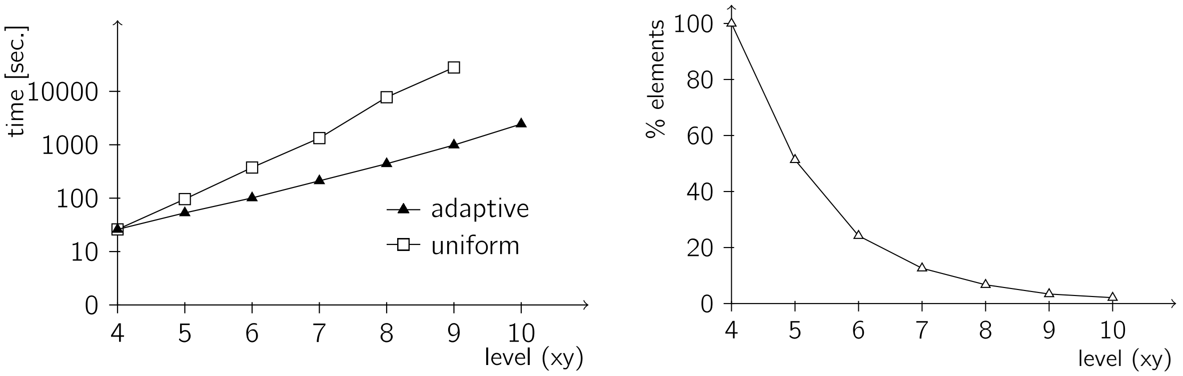


Figure 4.22.: Left: Runtime comparison between a uniform grid and an adaptive refinement for a fixed s-resolution on a logarithmic scale. Right: Percentage of the number of elements in each adaptive experiment compared to the total number of elements in a fully uniform grid of the same level. The values correspond to the numbers in Table 4.1.

The results clearly prove the superiority of the adaptive refinement approach compared to the iteration on a uniform grid in this particular example. While the resulting optimal network looks similar, the adaptive iteration is significantly faster and, referring to very high xy- or s-resolutions, allows for a suitable solution in the first place. Especially in the

	Uniform				Adaptive						
s	numEls	numDofs	time	error	runs	numEls	numDofs	%Els	%Dofs	time	error
2	8192	5445	50 sec.	0.0021	0	8192	5445	100	100	50 sec.	0.0021
3	16384	9801	96 sec.	0.0192	1	10453	6743	63.8	68.8	63 sec.	0.0067
4	32768	18513	230 sec.	0.0192	2	12983	8296	39.6	44.8	81 sec.	0.0239
5	65536	35937	756 sec.	0.0175	3	15912	10066	24.3	17.0	104 sec.	0.0091
6	131072	70785	3893 sec.	0.1529	4	19202	12049	14.7	17.0	148 sec.	0.0590
7	262144	140481	54715 sec.	0.0238	5	22969	14301	8.8	10.2	166 sec.	0.0084
8	-	-	-	-	6	27290	16892	5.2	6.0	218 sec.	5.1695

Table 4.2.: Comparison between uniform and adaptive grid iteration for a fixed xy-resolution of level 5, where only the image range (s) is refined. For each adaptive experiment, the initial s-level is 2. The first column refers to the s-level of the uniform grid and highest local s-resolution of the final adaptive grid respectively. The computed error equals the primal-dual gap in the last iteration. The experiment on a uniform grid of s-level 8 resolution is omitted due to its high runtime and memory consumption. Note that the very last adaptive experiment did not seem to converge as stated by a relatively high primal-dual gap, which we believe is caused by a slow convergence of the dual problem, however the result looks as expected.

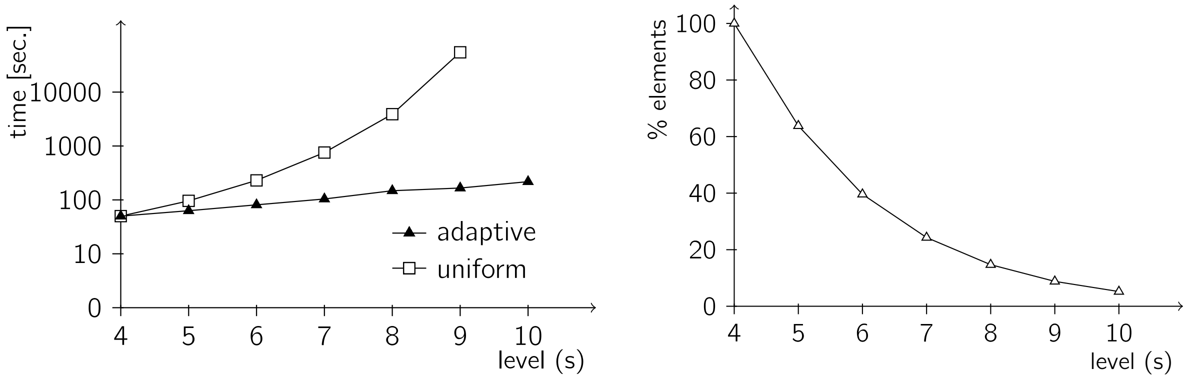


Figure 4.23.: Left: Runtime comparison between a uniform grid and an adaptive refinement for a fixed xy-resolution on a logarithmic scale. Right: Percentage of the number of elements in each adaptive experiment compared to the total number of elements in a fully uniform grid of the same level. The values correspond to the numbers in Table 4.2.

case of an adaptive s-refinement, the runtime increases linearly instead of exponentially, which is probably due to the scaling of the number of constraints of order $\mathcal{O}(n^2)$ for n nodes in one s-line. The acceleration is explained by the decreasing percentage of the total number of elements and degrees of freedom compared to the uniform grid of the same level, which falls below one percent in some examples. Note also that in most experiments the primal-dual gap is roughly of the same order in the uniform and adaptive case after a fixed number of iterations.

Naturally the displayed results depend on the structure of the given initial and final

	Uniform				Adaptive						
xy/s	numEls	numDofs	time	error	runs	numEls	numDofs	%Els	%Dofs	time	error
4/2	2048	1445	14 sec.	0.0069	0	2048	1445	100	100	14 sec.	0.0069
5/3	16384	9801	96 sec.	0.0192	2	7111	4576	43.4	46.7	44 sec.	0.0101
6/4	131072	71825	855 sec.	0.0165	4	30961	18800	23.6	26.2	184 sec.	0.0431
7/5	1048576	549153	20014 sec.	0.0013	6	91391	53596	8.7	9.8	632 sec.	0.0027
8/6	8388608	4293185	224221 sec.	0.0047	8	146825	84749	1.7	2.0	1405 sec.	0.0019
9/7	-	-	-	-	10	295227	167030	0.4	0.5	3438 sec.	0.0008
10/8	-	-	-	-	12	667289	370570	0.1	0.1	9767 sec.	0.0003

Table 4.3.: Comparison between uniform and adaptive grid iteration. For each adaptive experiment, the initial xy/s-level is 4/2. The first column refers to the xy/s-level of the uniform grid and highest local xy/s-resolution of the final adaptive grid respectively. The computed error equals the primal-dual gap in the last iteration. The experiments on a uniform grid of xy/s-level 9/7 or higher are omitted due to their high runtime and memory consumption.

measures. A higher network complexity such as in case of transport from 16 to 16 mass points requires a higher number of refined elements in each round and as a consequence, the runtime of the adaptive algorithm deviates less from the uniform iteration. However, even for more complex examples there always exists a level threshold from which the adaptive approach starts to pay off eventually.

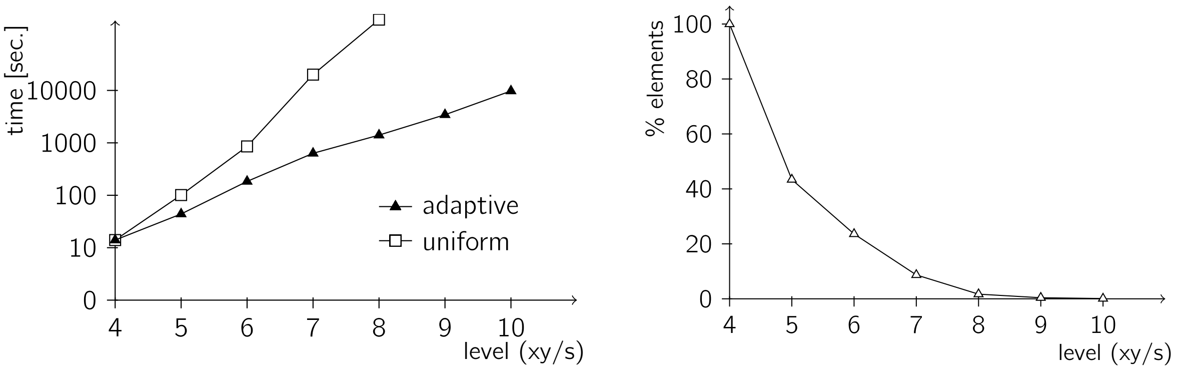


Figure 4.24.: Left: Runtime comparison between a uniform grid and an adaptive refinement on a logarithmic scale. Right: Percentage of the number of elements in each adaptive experiment compared to the total number of elements in a fully uniform grid of the same level. The values correspond to the numbers in Table 4.3.

4.4.7. Comparison of different refinement strategies

In order to obtain the results presented in Section 4.4.5, we employed a combined gradient and local primal-dual gap refinement strategy. However, since both methods come along with certain advantages and drawbacks, we want to investigate the impact of both approaches separately.

The first aspect which comes to mind considering the differences of the two strategies is the fact that the gradient refinement is *static*, meaning that it can only incorporate information from the *current* state of the primal variable, neglecting information concerning the dual variable as well as any future perceptions. In general, if the primal variable does not admit a large gradient within one element, this element will not be refined although it might contain some parts of the exact transportation network. Thus, there is no guarantee that a successive gradient refinement yields the optimal solution on a fully uniform high resolution grid. Additionally, gradient refinement certainly leads to redundantly refined elements especially concerning the lifted dimension, since any element containing a jump from 1 to 0 in s-direction will be refined regardless of the necessity of an additional grid layer.

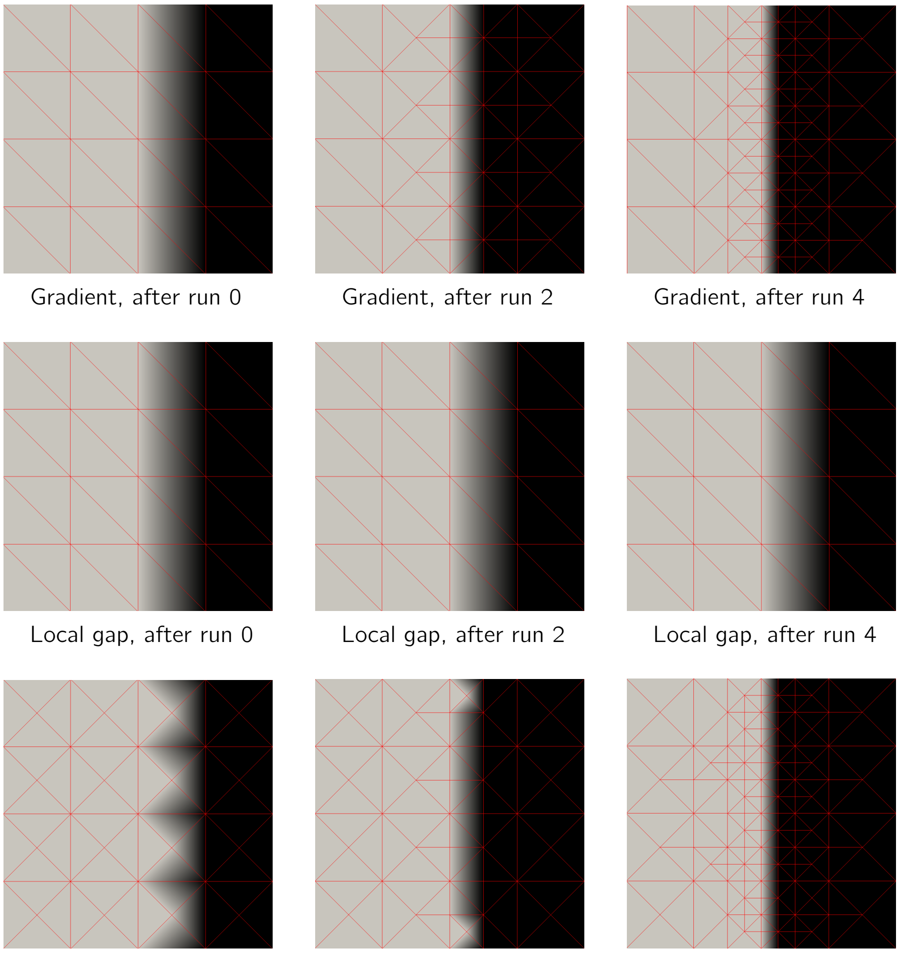
On the other hand, the local primal-dual gap by definition is unable to look more than one refinement step ahead. As a consequence, one can explicitly construct examples depending on the underlying grid structure where the gradient refinement is of certain advantage for obtaining sharp edges in the primal variable. In Figure 4.25, we display an example where the primal-dual gap fails to refine any element. Note that since the refinement is performed via element bisection, refinement of the initial grid does not provide any new node on the region boundaries. Therefore, the primal variable v (being prescribed on the boundaries) does not change on a higher resolution grid, thus the local primal-dual gap equals zero everywhere and no element is refined. In contrast to this, starting with a grid which is already refined once and thus symmetric, by the next refinement new boundary nodes are introduced, thus the gap is non-zero at least in some elements.

Another drawback of the gap refinement lies in the lack of a criterion for distinguishing between refinement in xy- or s-direction. While this criterion is naturally implemented in the gradient refinement, it would require an independent local gap computation on both a xy-refined and a s-refined grid, coming along with an additional runtime increase. Just as in case of gradient refinement, this might lead to some redundantly refined elements.

Although both methods admit different desirable features, we observe that the qualitative variations in comparison to the fully uniform grid solution are scarcely perceptible. In Figure 4.26, we juxtapose the numerical solutions obtained on an adaptive grid via gradient refinement, local primal-dual gap refinement (where we additionally refine the upper and lower boundary elements to prevent the boundary problems explained before) and on a fully uniform grid. Both methods refine the relevant regions where the network establishes, while the gradient method seems to be slightly advantageous since the refined elements remain closer to the network edges.

4.4.8. Discussion

We have discussed a novel adaptive finite element approach for numerical simulations of the lifted branched transport and urban planning problems. The presented approach tackles the main difficulties arising from functional lifting such as higher dimensionality and a



Gradient, after run 0

Gradient, after run 2

Gradient, after run 4

Local gap, after run 0

Local gap, after run 2

Local gap, after run 4

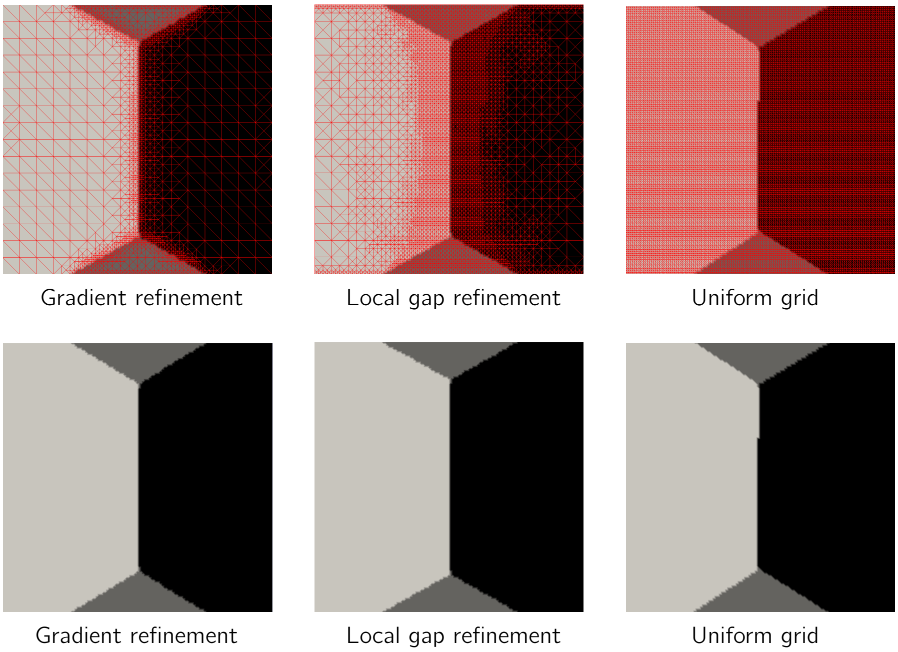
Local gap, after run 0

Local gap, after run 2

Local gap, after run 4

Figure 4.25.: Comparison between gradient (first row) and local primal-dual gap refinement (second and third row). In the second row, refining all elements of the leftmost image does not provide any new node on the boundaries, thus according to the local gap, no element needs to be refined. In contrast, by starting from a refined grid in the third row, the boundary elements are refined by the local gap criterion.

possibly infinite number of non-local constraints and is therefore most likely also applicable to more general problems of this form. We developed an algorithmic framework for the saddle point problem based on [26] as well as the projection onto the non-local constraint set, discussed some refinement criteria and presented numerical simulation results. The latter yield good approximations of the optimal transportation networks and prove to be clearly beneficial for higher network complexities and higher image resolutions.
The adaptive grid approach promises to be a useful tool for functional lifting problems, however it comes along with several difficulties. In order to exploit the full potential of


Gradient refinement Local gap refinement Uniform grid

Gradient refinement Local gap refinement Uniform grid

Figure 4.26.: Comparison between different refinement strategies on the example of branched transport from two to two mass points for $\alpha = 0.01$.

adaptivity, one would need to carefully study the effects of optimal refinement criteria and rigorously prove that a refinement strategy yields the same solution as on a uniform grid. Besides, an additional element coarsening could eliminate the overhead caused by redundantly refined elements. Finally, the runtime discrepancy between uniform and adaptive iterations might be further increased by a more efficient implementation of the refinement routine, which we do not claim to be optimized in every detail.

5 A phase field approximation approach

Inspired by elliptic approximations of free-discontinuity problems, where a part of the desired output of a model consists of some kind of lower-dimensional set, phase field approximations have proven to be a practical tool for numerical purposes. Similar to the approach of Ambrosio–Tortorelli (see Section 2.4.1) in approximating the Mumford–Shah problem, there have been several successful attempts to make use of this idea in the context of finding optimal transportation networks. For the branched transport cost, in [51] the authors introduce the functional

$$\mathcal{M}_\varepsilon^\alpha(v) = \int_\Omega \varepsilon^{\alpha-1}|v(x)|^\beta + \varepsilon^{\alpha+1}|\nabla v(x)|^2 \mathrm{d}x$$

on the space $W^{1,2}(\Omega)$ and prove Γ-convergence to $c\mathcal{M}^\alpha$ in dimension two for some constant c (cf. Theorem 3.5.1 and 3.5.2). The underlying idea, similar to the Ambrosio–Tortorelli approach, is to approximate the measure $\mathcal{F}$ concentrated on a one-dimensional set by a “smoothed” version v, where the smoothness is governed by a parameter ε. A comparable result has been provided, for instance, by [24] for the Steiner tree problem, a special case of both the branched transport and the urban planning model. The authors introduce the energy

$$\mathcal{S}_\varepsilon^\alpha(\sigma,\varphi) = \int_\Omega \frac{1}{2\varepsilon}\varphi^2|\sigma|^2 \mathrm{d}x + \int_\Omega \frac{\varepsilon}{2}|\nabla\varphi|^2 + \frac{1}{2\varepsilon}(1-\varphi)^2 \mathrm{d}x$$

for div $\sigma = (\mu_+ - \mu_-) * \rho_\varepsilon$ and $\eta \le \varphi \le 1$ for some smoothing kernel ρ_ε and $\frac{\eta}{\varepsilon} \to \alpha$ for $\varepsilon \to 0$. Here, the one-dimensional measure $\mathcal{F}$ is represented by a smooth vector-valued measure σ, which is forced to be non-zero on the desired one-dimensional set Σ. The phase field φ is close to 1 away from this set. Thus, a similar Γ-convergence result can be achieved, stating that $\mathcal{S}_\varepsilon^\alpha$ Γ-converges to the functional

$$\mathcal{S}^\alpha(\sigma,\varphi) = \int_\Sigma (1+\alpha\tilde{F})\mathrm{d}\mathcal{H}^1$$

if $\varphi \equiv 1$ and $\sigma = \tilde{F}\hat{e}(\mathcal{H}^1 \llcorner \Sigma)$ (see Theorem 3.5.6).
While both the branched transport and the Steiner tree problem have already been investigated using a phase field approximation, the urban planning problem (and thus, the generalized urban planning problem as a superordinate case) have only been tackled in a theoretical manner by a general framework to approximate transportation network problems via phase field energy potentials in [67]. In this chapter we propose a more specific phase field model which was developed in [33] based on the one introduced by [24]. It covers the class of generalized urban planning problems as defined in Section 3.3.4. We start with a description of the model in detail, cite some analytical results and provide a description of the numerical optimization procedure used to produce the computational results. We conclude with a short discussion about the advantages and disadvantages of the proposed method.

5.1. Model

Following the course of [33], we aim at investigating a piecewise affine transportation cost function

$$c(m) = \min_{i=1,\dots,N}\{a_0 m, a_1 m + b_1, \dots, a_N m + b_N\}$$

with $a_0 > a_1 > \ldots > a_N$, $b_1 < \ldots < b_N < \infty$. In order to cover the case where no diffuse part is allowed, we make the additional assumption that for $a_0 = \infty$,

$$c(m) = \begin{cases} 0 & \text{if } m = 0, \\ \min\limits_{i=1,\dots,N}\{a_0 m, a_1 m + b_1, \dots, a_N m + b_N\} & \text{otherwise.} \end{cases}$$

The generalized urban planning cost functional (as defined in Section 3.3.4) is then given by

$$\mathcal{E}_g^{a,b}(\mathcal{F}) = \int_\Sigma c(\tilde{F}(x))\mathrm{d}\mathcal{H}^1(x) + c'(0)|\mathcal{F}^\perp|(\overline{\Omega})$$

for $a_0 < \infty$ and

$$\mathcal{E}_g^{a,b}(\mathcal{F}) = \begin{cases} \int_\Sigma c(\tilde{F}(x))\mathrm{d}\mathcal{H}^1(x) & \text{if } \mathcal{F}^\perp = 0, \\ \infty & \text{otherwise} \end{cases}$$

for $a_0 = \infty$, where $\mathcal{F} = \tilde{F}\hat{e}\mathcal{H}^1 \llcorner \Sigma + \mathcal{F}^\perp$ with a rectifiable set $\Sigma \subset \Omega$, a multiplicity $\tilde{F} : \Sigma \to [0,\infty)$, an orientation $\hat{e} : \Sigma \to S^1$ and a $\mathcal{H}^1$-diffuse part $\mathcal{F}^\perp$, which consists of a Lebesgue-continuous and a Cantor part (cf. Section 2.2).

By seizing the intuitive approach of [24], we define a functional $\tilde{\mathcal{E}}_\varepsilon^{a,b,\mu_+,\mu_-} : L^2(\Omega,\mathbb{R}^2) \times W^{1,2}(\Omega)^N \to [0,\infty]$ by

$$\begin{aligned}
&\tilde{\mathcal{E}}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma,\varphi_1,\dots,\varphi_N)\\
&= \begin{cases} \int_\Omega \min\left\{a_0|\sigma(x)|, \min\limits_{i=1,\dots,N}\{\varphi_i(x)^2 + \frac{a_i^2\varepsilon^2}{b_i}\}\frac{|\sigma(x)|^2}{2\varepsilon}\right\} + \sum_{i=1}^N \frac{b_i}{2}\left[\varepsilon|\nabla\varphi_i(x)|^2 + \frac{(\varphi_i(x)-1)^2}{\varepsilon}\right] \mathrm{d}x \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \operatorname{div}\sigma = \mu_+^\varepsilon - \mu_-^\varepsilon, \\ \infty \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\ \ \text{otherwise}, \end{cases}
\end{aligned}$$

where $\mu_+^\varepsilon, \mu_-^\varepsilon$ are smoothed versions of μ_+, μ_-. Here, the vector measure σ approximates the mass flux $\mathcal{F}$, while the phase fields $\varphi_1,\dots,\varphi_N$ take value 1 away from Σ and are equal to 1 in the limit $\varepsilon \to 0$. Moreover, φ_j approaching 0 indicates the parts of the network where the j-th term in the cost functional is cheapest, in other words where $a_jm + b_j = \min\{a_0m, a_1m+b_1,\dots,a_Nm+b_N\}$. In those parts where the linear term has minimal costs, no phase field is active (meaning equal to 0), which corresponds to transport outside of the network Σ.
The functional $\tilde{\mathcal{E}}_\varepsilon^{a,b,\mu_+,\mu_-}$, though quite intuitive, involves a minimum term which is non-convex with respect to σ. To ensure existence of a minimizer and simplify the numerical handling, we define the following relaxed version of the cost functional [33].

Definition 5.1.1 (Phase field cost functional)**.** Let $\varepsilon > 0$ and $\rho : \mathbb{R}^2 \to [0,\infty)$ be a smoothing kernel with support on the unit ball and $\int_{\mathbb{R}^2} \rho \,\mathrm{d}x = 1$. Given the initial and final measures $\mu_+, \mu_- \in \mathcal{P}(\overline{\Omega})$, we set $\mu_\pm^\varepsilon = \rho_\varepsilon * \mu_\pm$ with $\rho_\varepsilon = \frac{1}{\varepsilon^2}\rho(\frac{\cdot}{\varepsilon})$. Moreover, we define the set of admissible functions as

$$\begin{aligned}
X_\varepsilon^{\mu_+,\mu_-} = \{(\sigma,\varphi_1,\dots,\varphi_N) &\in L^2(\Omega,\mathbb{R}^2)\times W^{1,2}(\Omega)^N \ : \\
&\operatorname{div}\sigma = \mu_+^\varepsilon - \mu_-^\varepsilon,\ \varphi_1 = \ldots = \varphi_N = 1 \text{ on } \partial\Omega\}.
\end{aligned}$$

Then, the *phase field cost functional* is given by $\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-} : X_\varepsilon^{\mu_+,\mu_-} \to [0,\infty)$,

$$\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma,\varphi_1,\dots,\varphi_N) = \int_\Omega \omega_\varepsilon\left(a_0, \frac{\gamma_\varepsilon(x)}{\varepsilon}, |\sigma(x)|\right)\mathrm{d}x + \sum_{i=1}^N b_i\mathcal{L}_\varepsilon(\varphi_i), \tag{5.1}$$

where

$$\begin{aligned}
\mathcal{L}_\varepsilon(\varphi) &= \frac{1}{2}\int_\Omega \varepsilon|\nabla\varphi(x)|^2 + \frac{(\varphi(x)-1)^2}{\varepsilon}\,\mathrm{d}x\,,\\
\gamma_\varepsilon(x) &= \min_{i=1,\dots,N}\left\{\varphi_i(x)^2 + a_i^2\varepsilon^2/b_i\right\}\,,\\
\omega_\varepsilon\left(\alpha_0, \frac{\gamma_\varepsilon(x)}{\varepsilon}, |\sigma(x)|\right) &= \left.\begin{cases} \frac{\gamma_\varepsilon(x)}{\varepsilon}\frac{|\sigma(x)|^2}{2} & \text{if } |\sigma(x)| \le \frac{a_0\varepsilon}{\gamma_\varepsilon(x)} \\ a_0(|\sigma(x)| - \frac{a_0\varepsilon}{2\gamma_\varepsilon(x)}) & \text{if } |\sigma(x)| > \frac{\alpha_0\varepsilon}{\gamma_\varepsilon(x)} \end{cases}\right\} + \varepsilon^p|\sigma(x)|^2 \quad \text{for } a_0 < \infty\,,\\
\omega_\varepsilon\left(a_0, \frac{\gamma_\varepsilon(x)}{\varepsilon}, |\sigma(x)|\right) &= \frac{\gamma_\varepsilon(x)}{\varepsilon}\frac{|\sigma|^2}{2} \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } a_0 = \infty
\end{aligned}$$

for some $p > 1$.

For fixed $\varphi_1, \dots, \varphi_N$ and ignoring the term $\varepsilon^p|\sigma(x)|^2$, the term $\omega_\varepsilon\left(a_0, \frac{\gamma_\varepsilon(x)}{\varepsilon}, |\sigma(x)|\right)$ is the lower semi-continuous envelope with respect to σ of the original minimum term in the functional $\tilde{\mathcal{E}}_\varepsilon^{a,b,\mu_+,\mu_-}$ (see Figure 5.1). Taking the convex envelope does not affect the Γ-convergence result, but ensures existence of a minimizer (see Section 5.2).

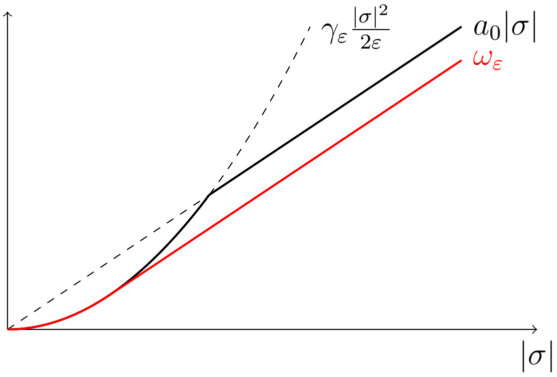


Figure 5.1.: Difference between the functionals $\tilde{\mathcal{E}}_\varepsilon^{a,b,\mu_+,\mu_-}$ and $\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}$. Black curve: Sketch of the function $f(|\sigma|) := \min\{a_0|\sigma|, \gamma_\varepsilon \frac{|\sigma|^2}{2\varepsilon}\}$ for fixed γ_ε. f is non-convex with respect to $|\sigma|$. Red curve: Sketch of the function $g(|\sigma|) := \omega_\varepsilon(a_0, \frac{\gamma_\varepsilon}{\varepsilon}, |\sigma|)$ for fixed γ_ε, which is the lower semi-continuous envelope of f.

Remark 5.1.2 (Regularization term $\varepsilon^p|\sigma(x)|^2$). For $a_0 < \infty$, the term $\varepsilon^p|\sigma(x)|^2$ ensures $L^2(\Omega, \mathbb{R}^2)$-coercivity in σ, which is necessary for sequentially weak compactness of subsets of $X_\varepsilon^{\mu_+,\mu_-}$ with finite cost and thus ensures existence of a minimizer (again, see Section 5.2). Besides, the term has no other purpose and is especially not essential for the numerical treatment.

5.2. Analysis

Before we describe the numerical treatment of the phase field cost functional, we want to prove existence of a minimizer in Section 5.2.1 and state the Γ-convergence and equi-coercivity result justifying its usability as an approximation of the generalized urban planning functional in Section 5.2.2.

5.2.1. Existence of a minimizer

As a preliminary result, we state existence of a minimizer of the phase field cost functional [33].

Theorem 5.2.1 (Existence of a minimizer)**.** *The phase field cost functional* $\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}$ *has a minimizer* $(\sigma, \varphi_1, \dots, \varphi_N) \in X_\varepsilon^{\mu_+,\mu_-}$.

Proof. The functional is bounded below by 0. Moreover, choose $\hat{\varphi}_1 \equiv \ldots \equiv \hat{\varphi}_N \equiv 1$ and $\hat{\sigma} = \nabla\psi$ for a function ψ which solves the equation $\Delta\psi = \mu_+^\varepsilon - \mu_-^\varepsilon$ on Ω with Neumann boundary conditions $\nabla\psi \cdot \nu_{\partial\Omega} = 0$ (where $\nu_{\partial\Omega}$ is the outward unit normal on $\partial\Omega$). One can easily show that the solution to the Poisson problem exists since $\int_\Omega \mu_+^\varepsilon - \mu_-^\varepsilon \mathrm{d}x = 0$ and satisfies $\psi \in W^{2,2}(\Omega)$. With this we obtain that $(\hat{\sigma}, \hat{\varphi}_1, \ldots, \hat{\varphi}_N) \in X_\varepsilon^{\mu_+,\mu_-}$ and $\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}(\hat{\sigma}, \hat{\varphi}_1, \ldots, \hat{\varphi}_N) < \infty$. Thus, the functional has a non-empty domain.
Let $(\sigma_k, \varphi_1^k, \ldots, \varphi_N^k) \in X_\varepsilon^{\mu_+,\mu_-}$ be a minimizing sequence with

$$\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma_k, \varphi_1^k, \ldots, \varphi_N^k) \to \inf \mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}$$

monotonically for $k \to \infty$. Since $\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}$ is coercive with respect to $L^2(\Omega, \mathbb{R}^2) \times W^{1,2}(\Omega)^N$, the sequence $(\sigma_k, \varphi_1^k, \ldots, \varphi_N^k)$ is uniformly bounded. Thus, there exists a weakly converging subsequence (which is still indexed with k for the sake of simplicity) $(\sigma_k, \varphi_1^k, \ldots, \varphi_N^k) \rightharpoonup (\sigma, \varphi_1, \ldots, \varphi_N) \in X_\varepsilon^{\mu_+,\mu_-}$ due to the closedness of $X_\varepsilon^{\mu_+,\mu_-}$ with respect to weak convergence in $L^2(\Omega, \mathbb{R}^2) \times W^{1,2}(\Omega)^N$. Now consider a subsequence along which each term $\mathcal{L}_\varepsilon(\varphi_i^k)$ converges and the φ_i^k converge pointwise almost everywhere such that $\gamma_\varepsilon^k(x) = \min_{i=1,\ldots,N}\{\varphi_i^k(x)^2 + a_i^2\varepsilon^2/b_i\}$ converges for almost every $x \in \Omega$. In addition, from Mazur's lemma, a sequence of convex combinations $\sum_{j=k}^{m_k} \lambda_j^k \sigma^j$ of the σ^k converges strongly. Thus, up to another subsequence, pointwise and we can apply Fatou's lemma. Hence we have

$$\begin{aligned}
\inf \mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-} &= \lim_{k\to\infty} \mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma^k, \varphi_1^k, \ldots, \varphi_N^k) \\
&= \lim_{k\to\infty} \int_\Omega \omega_\varepsilon\left(a_0, \frac{\gamma_\varepsilon^k(x)}{\varepsilon}, |\sigma^k(x)|\right) \mathrm{d}x + \sum_{i=1}^N b_i \lim_{k\to\infty} \mathcal{L}_\varepsilon(\varphi_i^k) \\
&\geq \lim_{k\to\infty} \sum_{j=k}^{m_k} \lambda_j^k \int_\Omega \omega_\varepsilon\left(a_0, \frac{\gamma_\varepsilon^j(x)}{\varepsilon}, |\sigma^j(x)|\right) \mathrm{d}x + \sum_{i=1}^N b_i \mathcal{L}_\varepsilon(\varphi_i) \\
&\geq \int_\Omega \liminf_{k\to\infty} \sum_{j=k}^{m_k} \lambda_j^k \omega_\varepsilon\left(a_0, \frac{\gamma_\varepsilon^j(x)}{\varepsilon}, |\sigma^j(x)|\right) \mathrm{d}x + \sum_{i=1}^N b_i \mathcal{L}_\varepsilon(\varphi_i) \\
&\geq \int_\Omega \liminf_{k\to\infty} \sum_{j=k}^{m_k} \lambda_j^k \omega_\varepsilon\left(a_0, \inf_{i=k,\ldots,m_k} \frac{\gamma_\varepsilon^i(x)}{\varepsilon}, |\sigma^j(x)|\right) \mathrm{d}x + \sum_{i=1}^N b_i \mathcal{L}_\varepsilon(\varphi_i) \\
&\geq \int_\Omega \liminf_{k\to\infty} \omega_\varepsilon\left(a_0, \inf_{i=k,\ldots,m_k} \frac{\gamma_\varepsilon^i(x)}{\varepsilon}, \sum_{j=k}^{m_k} \lambda_j^k |\sigma^j(x)|\right) \mathrm{d}x + \sum_{i=1}^N b_i \mathcal{L}_\varepsilon(\varphi_i) \\
&= \mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma, \varphi_1, \ldots, \varphi_N).
\end{aligned}$$

□

Remark 5.2.2. The phase field cost functional is convex in σ for fixed phase fields $\varphi_1, \ldots, \varphi_N$, but strongly non-convex in $\varphi_1, \ldots, \varphi_N$ due to the minimum term in γ_ε. Thus, the energy might admit some local minima and uniqueness of a minimizer cannot be guaranteed.

5.2.2. Γ-convergence and equi-coercivity

As in the case of the phase field approximation of the Steiner tree problem [24], the functional $\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}$ Γ-converges to the generalized urban planning energy. The result is stated in the following, for a detailed proof, we refer the reader to [33].

Theorem 5.2.3 (Γ-convergence of the phase field cost functional)**.** *Let $X^{\mu_+,\mu_-} = \{\mathcal{F} \in \mathcal{M}(\overline{\Omega}, \mathbb{R}^2) \ : \ \operatorname{div} \mathcal{F} = \mu_+ - \mu_-\}$. We define the functionals*

$$E^{a,b,\mu_+,\mu_-}(\sigma, \varphi_1, \dots, \varphi_N) = \begin{cases} \mathcal{E}^{a,b,\mu_+,\mu_-}(\sigma) & \textit{if } \sigma \in X^{\mu_+,\mu_-}, \ \varphi_1 = \ldots = \varphi_N = 1 \textit{ a.e.}, \\ \infty & \textit{otherwise}, \end{cases}$$

$$E_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma, \varphi_1, \dots, \varphi_N) = \begin{cases} \mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma, \varphi_1, \dots, \varphi_N) & \textit{if } (\sigma, \varphi_1, \dots, \varphi_N) \in X_\varepsilon^{\mu_+,\mu_-}, \\ \infty & \textit{otherwise}. \end{cases}$$

Then, for admissible $\mu_+, \mu_- \in \mathcal{M}_+(\overline{\Omega})$, we have

$$E_\varepsilon^{a,b,\mu_+,\mu_-} \xrightarrow{\Gamma} E^{a,b,\mu_+,\mu_-},$$

where the Γ-limit is with respect to the weak- convergence in $\mathcal{M}(\overline{\Omega}, \mathbb{R}^2)$ and strong convergence in $L^1(\Omega)^N$.*

Proof. See [33]. □

By Theorem 5.2.3, the generalized urban planning functional can indeed be approximated by the relaxed phase field cost functional. The following result from [33] states that the *minimizers* of the functional $\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}$ indeed approximate minimizers of the original functional $\mathcal{E}^{a,b,\mu_+,\mu_-}$ as well.

Theorem 5.2.4 (Equi-coercivity of the phase field cost functional)**.** *For $\varepsilon \to 0$ let $(\sigma^\varepsilon, \varphi_1^\varepsilon, \dots, \varphi_N^\varepsilon)$ be a sequence with uniformly bounded phase field cost functional $E_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma^\varepsilon, \varphi_1^\varepsilon, \dots, \varphi_N^\varepsilon) < C < \infty$. Then, along a subsequence, $\sigma^\varepsilon \rightharpoonup^* \sigma$ in $\mathcal{M}(\overline{\Omega}, \mathbb{R}^2)$ for some $\sigma \in \mathcal{M}(\overline{\Omega}, \mathbb{R}^2)$ and $\varphi_i^\varepsilon \to 1$ in $L^1(\Omega)$, $i = 1, \dots, N$.*
As a consequence, if $\mu_+, \mu_- \in \mathcal{M}_+(\overline{\Omega})$ are admissible and such that there exists $\mathcal{F} \in X^{\mu_+,\mu_-}$ with $\mathcal{E}^{a,b,\mu_+,\mu_-}(\mathcal{F}) < \infty$, then any sequence of minimizers of $E_\varepsilon^{a,b,\mu_+,\mu_-}$ contains a subsequence converging to a minimizer of E^{a,b,μ_+,μ_-} as $\varepsilon \to 0$.

Proof. See [33]. □

For numerical purposes, Theorem 5.2.4 yields an essential result. Keeping ε fixed, one can minimize $\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}$ instead of $\mathcal{E}^{a,b,\mu_+,\mu_-}$ and the minimizers of the phase field cost functional indeed approximate the true optimal network with respect to the generalized urban planning energy.

Remark 5.2.5 (Discrete Γ-convergence). Note that it is not straightforward to show that the Γ-convergence result from Theorem 5.2.3 also holds for a discretized version of the involved functionals. This requires some additional assumptions on the relation of the Γ-convergence parameter ε and a discrete mesh size. Such a result in case of a discrete Ambrosio–Tortorelli approximation of the Mumford–Shah functional with finite differences has been stated by [6].

5.3. Numerical optimization

In this section, we describe the numerical discretization using finite elements on a simple triangular grid, present a suitable optimization scheme and show some computational results.

5.3.1. Discretization

As before, we consider the energy functional

$$\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma,\varphi_1,\ldots,\varphi_N)=\int_\Omega \omega_\varepsilon\left(a_0,\frac{\gamma_\varepsilon(x)}{\varepsilon},|\sigma(x)|\right)\mathrm{d}x+\sum_{i=1}^N b_i\mathcal{L}_\varepsilon(\varphi_i).$$

The proposed phase field approximation allows a simple numerical treatment with piecewise constant and piecewise linear finite elements for the variables σ and $\varphi_1,\ldots,\varphi_N$, respectively. To this end, we introduce a triangulation $\mathcal{T}_h$ of the space $\Omega=(0,1)^2$ with minimal mesh size h, such that $\overline{\Omega}=\bigcup_{T\in\mathcal{T}_h}\bar{T}$. For the variables, we use the finite element function spaces

$$\begin{aligned}X_h^0&=\{v_h\in L^\infty(\Omega)\ :\ v_h|_T\in\mathbb{P}^0\ \forall\, T\in\mathcal{T}_h\},\\ X_h^1&=\{v_h\in C(\overline{\Omega})\ :\ v_h|_T\in\mathbb{P}^1\ \forall\, T\in\mathcal{T}_h\},\end{aligned}$$

where $\mathbb{P}^m$ denotes the space of polynomials of degree m. Hence, for a given basis of the spaces X_h^0, X_h^1, we can write the discrete counterparts of $\sigma,\varphi_1,\ldots,\varphi_N$ as a linear combination of basis functions. Then the discrete version of the phase field energy functional reads

$$\min_{\substack{(\sigma,\varphi_1,\ldots,\varphi_N)\in X_h^0\times(X_h^1)^N\\ \varphi_1|_{\partial\Omega}=\ldots=\varphi_N|_{\partial\Omega}=1}}\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma,\varphi_1,\ldots,\varphi_N)$$

under the constraint

$$\int_\Omega -\sigma\cdot\nabla\lambda\,\mathrm{d}x=\int_\Omega f_\varepsilon v_h\,\mathrm{d}x\quad\forall\lambda\in X_h^1,$$

where we abbreviated $f_\varepsilon=\rho_\varepsilon*(\mu_+-\mu_-)$ and the divergence constraint is enforced in its weak formulation. The integrals reduce to integration over the finite element basis functions and can be evaluated using midpoint quadrature.

5.3.2. Optimization

Here we describe the numerical optimization strategy used to find a minimizer of the phase field energy. Depending on the chosen parameter values, the problem requires a very careful treatment. Due to the minimum term in γ_ε, the problem is strongly non-convex even for larger values of ε, and any algorithm tends to get stuck in local minima. In the following, we distinguish between three different choices of parameters, which require different numerical treatments.

Single phase field and no diffuse mass flux ($N = 1$, $a_0 = \infty$)

In case of a single phase field and $a_0 = \infty$, the phase field cost functional reads

$$\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma,\varphi) = \int_\Omega \frac{\gamma_\varepsilon(x)}{\varepsilon}\frac{|\sigma(x)|^2}{2} + \frac{b_1}{2}\left(\varepsilon|\nabla\varphi(x)|^2 + \frac{(\varphi(x)-1)^2}{\varepsilon}\right)\mathrm{d}x$$

with $\gamma_\varepsilon(x) = \varphi(x)^2 + a_1^2\varepsilon^2/b_1$. This energy functional is similar to the Steiner tree problem as presented in [24], hence the employed optimization method is the same as well.
We update the variables σ and φ alternatingly. For the minimization with respect to σ, we turn to an unconstrained formulation applying a dual variable $\lambda \in X_h^1$, such that

$$\begin{aligned}\min_{\substack{\sigma\in X_h^0\\ \int_\Omega \sigma\cdot\nabla\lambda+\lambda f_\varepsilon\,dx=0\,\forall\lambda\in X_h^1}} \int_\Omega \frac{\gamma_\varepsilon}{\varepsilon}\frac{|\sigma|^2}{2}\,\mathrm{d}x &= \min_{\sigma\in X_h^0}\max_{\lambda\in X_h^1}\int_\Omega \frac{\gamma_\varepsilon}{\varepsilon}\frac{|\sigma|^2}{2} - \sigma\cdot\nabla\lambda - \lambda f_\varepsilon\,\mathrm{d}x\\ &= \max_{\lambda\in X_h^1}\min_{\sigma\in X_h^0}\int_\Omega \frac{\gamma_\varepsilon}{\varepsilon}\frac{|\sigma|^2}{2} - \sigma\cdot\nabla\lambda - \lambda f_\varepsilon\,\mathrm{d}x,\end{aligned}$$

where in the last step, we are allowed to exchange the maximum and minimum due to standard convex duality arguments. For fixed λ, the inner minimization problem with respect to σ can be performed explicitly by computing the optimality condition

$$\int_\Omega \frac{\gamma_\varepsilon}{\varepsilon}\sigma\cdot\theta - \theta\cdot\nabla\lambda\,\mathrm{d}x = 0 \;\;\forall\,\theta\in X_h^0,$$

which yields $\sigma = \frac{\varepsilon\nabla\lambda}{\gamma_\varepsilon}$. Inserting this into the maximization problem with respect to λ leads to

$$\max_{\lambda\in X_h^1}\int_\Omega -\frac{\varepsilon|\nabla\lambda|^2}{2\gamma_\varepsilon} - \lambda f_\varepsilon\,\mathrm{d}x = \min_{\lambda\in X_h^1}\int_\Omega \frac{\varepsilon|\nabla\lambda|^2}{2\gamma_\varepsilon} + \lambda f_\varepsilon\,\mathrm{d}x.$$

The optimality condition reads

$$\int_\Omega \frac{\varepsilon\nabla\lambda\cdot\nabla\mu}{\gamma_\varepsilon}\,\mathrm{d}x = -\int_\Omega \mu f_\varepsilon\,\mathrm{d}x \;\;\forall\,\mu\in X_h^1. \tag{5.2}$$

Given the piecewise linear basis functions, this conditions reduces to a linear system of equations for the coefficient vector of λ. Subsequently, for given λ, the optimal σ can be computed from the above equation.
The minimization with respect to φ can be performed in a similar way. For fixed σ, we can compute the optimality condition for φ as

$$\int_\Omega \frac{|\sigma|^2\varphi\psi}{\varepsilon} + b_1\varepsilon\nabla\varphi\cdot\nabla\psi + \frac{b_1}{\varepsilon}(\varphi-1)\psi \,\mathrm{d}x = 0 \;\; \forall\, \psi\in X_h^1 \text{ with } \psi|_{\partial\Omega}=0. \tag{5.3}$$

Again, this reduces to a linear system of equations for the coefficient vector of φ, such that the optimal φ can be obtained via a linear system solver.
In order to obtain a good approximation of the original generalized urban planning energy, we start the alternating minimization process with a relatively large value of $\varepsilon_{\text{start}}$ and decrease the phase field parameter up to the desired accuracy. Note that for larger ε, the energy landscape is closer to being convex, hence the optimization process is less likely to get stuck in local minima. The complete method is summarized in Algorithm 6.

Algorithm 6 Minimization for $N=1$, $a_0=\infty$

```
function SPFS($\varepsilon_{\text{start}}, \varepsilon_{\text{end}}, N_{\text{iter}}, a_1, b_1, \mu_+, \mu_-, \rho_{\varepsilon_{\text{end}}}$)
    Set $f_\varepsilon = (\mu_+ - \mu_-) * \rho_{\varepsilon_{\text{end}}}$, $\sigma^0 = 0$
    for $j = 1,\dots,N_{\text{iter}}$ do
        Set $\varepsilon_j = \varepsilon_{\text{start}} - (j-1)\frac{\varepsilon_{\text{start}}-\varepsilon_{\text{end}}}{N_{\text{iter}}-1}$
        Set $\varphi^j$ as the solution of (5.3) for given fixed $\sigma = \sigma^{j-1}$
        Set $\gamma_\varepsilon^j = (\varphi^j)^2 + a_1^2\varepsilon_j^2/b_1$
        Set $\lambda^j$ as the solution of (5.2) for given fixed $\gamma_\varepsilon = \gamma_\varepsilon^j$
        Set $\sigma^j = \frac{\varepsilon_j\nabla\lambda^j}{2\gamma_\varepsilon^j}$
    end for
end function
return $\sigma^{N_{\text{iter}}}, \varphi^{N_{\text{iter}}}, \lambda^{N_{\text{iter}}}$
```

Multiple phase fields and no diffuse mass flux ($N>1$, $a_0=\infty$)

In case of multiple phase fields, the minimization step with respect to σ does not change, since only γ_ε is different. However, the minimization with respect to $\varphi_1,\dots,\varphi_N$ becomes more challenging due to the lack of convexity of the minimum term in γ_ε. Therefore, this part requires a more careful treatment and a suitable initialization to prevent the method from getting stuck in local minima.
In order to avoid minimization of γ_ε with respect to φ_i, we replace the term by a slightly simpler version. First, we define the regions

$$R_i^\varepsilon = \{x\in\Omega \,:\, \gamma_\varepsilon(x) = \varphi_i(x)^2 + \alpha_i^2\varepsilon^2/\beta_i\}\,, \quad i=1,\dots,N, \tag{5.4}$$

specifying those parts of Ω, where the i-th part of γ_ε is the minimum. Assuming that each region R_i^ε is fixed, we replace the energy $\mathcal{E}_\varepsilon^{a,b,\mu_+,\mu_-}$ by the functional

$$\hat{\mathcal{E}}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma,\varphi_1,\dots,\varphi_N) = \sum_{i=1}^N \int_\Omega \frac{\varphi_i(x)^2 + a_i^2\varepsilon^2/b_i}{\varepsilon}\frac{|\sigma(x)|^2}{2}\chi_{R_i^\varepsilon}(x) + \frac{b_i}{2}\varepsilon|\nabla\varphi_i(x)|^2 \\ + \frac{b_i}{2}\frac{(\varphi_i(x)-1)^2}{\varepsilon}\,\mathrm{d}x,$$

where $\chi_{R_i^\varepsilon}$ is the characteristic function of the region R_i^ε. Minimizing with respect to every φ_i separately gives the optimality conditions

$$\int_\Omega \frac{\varphi_i\psi}{\varepsilon}|\sigma|^2\chi_{R_i^\varepsilon} + b_i\varepsilon\nabla\varphi_i\cdot\nabla\psi + \frac{b_i}{\varepsilon}(\varphi_i - 1)\psi\,\mathrm{d}x = 0 \quad \forall\psi\in X_h^1 \text{ with } \psi|_{\partial\Omega} = 0 \tag{5.5}$$

with $\varphi_i|_{\partial\Omega} = 1$, which as in the case of $N = 1$ can be solved by some linear system solver. Since in each iteration step, the regions R_i^ε are fixed, the performance of the algorithm strongly depends on the quality of the initial guess for the $\varphi_1,\dots,\varphi_N$. Although the regions are allowed to change in between the iteration steps, if for example $R_1^\varepsilon = \Omega$ and $R_2^\varepsilon = \dots = R_N^\varepsilon = \emptyset$, in the next step $\varphi_2 = \dots = \varphi_N$ will be equal to 1, such that the regions R_i^ε will not change throughout the whole minimization process.
To avoid this, we apply an additional method to produce a suitable initial guess for all phase fields. First, we construct an initial network σ^0 (not necessarily the optimal one) which performs the transport between the given measures μ_+ and μ_-. In simple examples, this can be done by hand, however, in our experiments we simply compute the optimal σ as if only phase field φ_1 was active via Algorithm 6, ignoring the $\varphi_2,\dots,\varphi_N$. Having an initial guess for σ^0, we can compute the mass $m(x)$ flowing through each point of the optimal network. To this end, we consider

$$\left(\chi_{B_r(0)} * |\sigma^0|\right)(x) = \int_{B_r(x)} |\sigma^0|(y)\,\mathrm{d}y \approx 2rm(x) \tag{5.6}$$

if r is sufficiently large compared to the width of the support of σ^0. Hence, we can compute the regions R_i^ε via

$$R_i^\varepsilon = \{x\in\Omega \mid i = \operatorname*{argmin}_{j=1,\dots,N}(a_j m(x) + b_j)\} \tag{5.7}$$

and the initial phase fields are defined as

$$\varphi_i^0(x) = \begin{cases} 0 & \text{if } x\in R_i^\varepsilon, \\ 1 & \text{otherwise.}\end{cases}$$

Note that the width of the support of the optimal measure σ does not only depend on the transported mass $m(x)$, but also on the parameters a_i, b_i of the phase field i which is active in the point x (for a detailed investigation of, see the construction of a recovery sequence for the Γ-convergence proof in [33]). This problem can be bypassed by a simple energy rescaling. Instead of one phase field parameter ε, we set $\varepsilon_i = b_i\varepsilon/a_i$ to be the parameter associated with the phase field φ_i, thus we replace the original energy functional by a rescaled version

$$\hat{\mathcal{E}}_\varepsilon^{a,b,\mu_+,\mu_-}(\sigma,\varphi_1,\ldots,\varphi_N) = \int_\Omega \omega_\varepsilon\left(a_0, \frac{\hat{\gamma}_\varepsilon(x)}{\varepsilon}, |\sigma(x)|\right)\,\mathrm{d}x + \sum_{i=1}^N b_i\mathcal{L}_{\varepsilon_i}(\varphi_i) \tag{5.8}$$

with $\hat{\gamma}_\varepsilon(x) = \min_{i=1,\ldots,N}\{\varphi_i(x)^2 + a_i^2\varepsilon\varepsilon_i/b_i\} = \min_{i=1,\ldots,N}\{\varphi_i(x)^2 + a_i\varepsilon^2\}$. By doing so, one obtains that the support of σ becomes $m\varepsilon$ instead of $a_i m\varepsilon/b_i$, thus the mass flowing through a segment can be readily computed via (5.6). Note that in the original formulation, the phase field width depended on the parameters a_i, b_i, while the profile of the phase field, namely the steepness of the gradient, was independent of them. The rescaling reverts this relation, thus the slope of φ_i now depends on the choice of the corresponding a_i, b_i, which will affect the numerical solution slightly. Beyond that, the rescaling does not affect any of the previous analytical results and is accomplished for numerical purposes only. The whole method is summarized in Algorithm 7.

Algorithm 7 Minimization for $N > 1$, $a_0 = \infty$

```
function MPFS($\varepsilon_{\text{start}}, \varepsilon_{\text{end}}, N_{\text{iter}}, a_1,\ldots,a_N, b_1,\ldots,b_N, \mu_+, \mu_-, \rho_{\varepsilon_{\text{end}}}$)
    Set $f_\varepsilon = (\mu_+ - \mu_-) * \rho_{\varepsilon_{\text{end}}}$
    Set $(\sigma^0, \cdot, \cdot) = SPFS(\varepsilon_{\text{start}}, \varepsilon_{\text{end}}, N_{\text{iter}}, a_1, b_1, \mu_+, \mu_-, \rho_{\varepsilon_{\text{end}}})$
    Compute regions $R_1^\varepsilon,\ldots,R_N^\varepsilon$ via (5.7)
    for $j = 1,\ldots,N_{\text{iter}}$ do
        Set $\varepsilon_j = \varepsilon_{\text{start}} - (j-1)\frac{\varepsilon_{\text{start}}-\varepsilon_{\text{end}}}{N_{\text{iter}}-1}$
        Set $\varphi_i^j$ as the solution of (5.5) for given fixed $\sigma = \sigma^{j-1}$, $i = 1,\ldots,N$
        Update regions $R_1^\varepsilon,\ldots,R_N^\varepsilon$ via (5.4)
        Set $\gamma_\varepsilon^j = \min_{i=1,\ldots,N}\left((\varphi_i^j)^2 + a_i^2\varepsilon^2/b_i\right)$
        Set $\lambda^j$ as the solution of (5.2) for given fixed $\gamma_\varepsilon = \gamma_\varepsilon^j$
        Set $\sigma^j = \frac{\varepsilon_j\nabla\lambda^j}{2\gamma_\varepsilon^j}$
    end for
end function
return $\sigma^{N_{\text{iter}}}, \varphi_1^{N_{\text{iter}}},\ldots,\varphi_N^{N_{\text{iter}}}$
```

Multiple phase fields and diffuse mass flux ($N > 1$, $a_0 < \infty$)

The difference to the previously handled case is the possible occurrence of a diffuse mass flux, accompanied by a region where no phase field is active. The energy functional changes slightly, since ω_ε consists of two parts depending on the mass flowing. This means that we may introduce, in addition to the regions R_i^ε, a set

$$R_0^\varepsilon = \left\{ x \in \Omega \ : \ |\sigma(x)| > \frac{a_0}{\gamma_\varepsilon/\varepsilon} \right\},$$

where no phase field is active and ω_ε attains its second part. Since there might be regions where $\gamma_\varepsilon(x) = \varphi_i(x)^2 + a_i^2\varepsilon^2/b_i$ but also $|\sigma(x)| > \frac{a_0}{\gamma_\varepsilon/\varepsilon}$ (hence $R_i^\varepsilon \cap R_0^\varepsilon$ for some i), we define

$$\tilde{R}_i^\varepsilon = R_i^\varepsilon \setminus R_0^\varepsilon \tag{5.9}$$

to separate the active regions from each other.
The optimal phase fields φ_i, as before, are obtained by minimizing the energy

$$\sum_{i=1}^N \int_\Omega \frac{\varphi_i(x)^2 + a_i^2\varepsilon^2/b_i}{\varepsilon} \frac{|\sigma(x)|^2}{2} \chi_{\tilde{R}_i^\varepsilon}(x) + \frac{b_i}{2}\varepsilon|\nabla\varphi_i(x)|^2 + \frac{b_i}{2\varepsilon}(1-\varphi_i(x))^2 \,\mathrm{d}x. \tag{5.10}$$

The minimization with respect to σ requires a little more care due to the changes in the term ω_ε. The optimization problem in σ reads

$$\min_{\substack{\sigma \in X_h^0 \\ \int_\Omega \sigma\cdot\nabla\lambda + \lambda f_\varepsilon \,\mathrm{d}x = 0 \,\forall \lambda \in X_h^1}} \int_\Omega \omega_\varepsilon\left(a_0, \frac{\gamma_\varepsilon(x)}{\varepsilon}, |\sigma(x)|\right) \mathrm{d}x\,.$$

The optimality conditions read

$$\begin{aligned} 0 &= \int_\Omega \frac{\xi(|\sigma|)}{|\sigma|}\sigma\cdot\psi - \nabla\lambda\cdot\psi \,\mathrm{d}x & \forall\, \psi \in X_h^0, \\ 0 &= \int_\Omega \sigma\cdot\nabla\mu + \mu f_\varepsilon \,\mathrm{d}x & \forall\, \mu \in X_h^1, \end{aligned}$$

where $\xi(|\sigma|)$ denotes the partial derivative of ω_ε with respect to the third component, namely

$$\xi(|\sigma|) = \partial_3\omega_\varepsilon\left(a_0, \frac{\gamma_\varepsilon}{\varepsilon}, |\sigma|\right) = \min\left\{\frac{\gamma_\varepsilon}{\varepsilon}|\sigma|, a_0\right\} + 2\varepsilon^p|\sigma|.$$

Due to the non-linearity, the optimality conditions do not reduce to a linear system as before. Instead, by introducing basis functions $\{b_i^0\}_i$ of the space X_h^0 and $\{b_i^1\}_i$ of X_h^1, the optimality conditions with respect to the coefficient vectors $\hat{\sigma}, \hat{\lambda}$ of the corresponding variables can be written as

$$0 = R(\hat{\sigma}, \hat{\lambda}) = \begin{pmatrix} M\left[\frac{\xi(|\sigma|)}{|\sigma|}\right] & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \hat{\sigma} \\ \hat{\lambda} \end{pmatrix} + \begin{pmatrix} 0 \\ F \end{pmatrix},$$

where $M\left[\frac{\xi(|\sigma|)}{|\sigma|}\right]$, B are the finite element matrices and F is a vector defined by

$$M\left[\frac{\xi(\sigma|)}{|\sigma|}\right]_{ij} = \int_\Omega \frac{\xi(|\sigma|)}{|\sigma|} b_i^0 b_j^0 \,\mathrm{d}x,\ B_{ij} = \int_\Omega b_i^0 \cdot \nabla b_j^1 \,\mathrm{d}x,\ F_i = \int_\Omega b_i^1 f_\varepsilon \,\mathrm{d}x.$$

The task of finding an optimal σ reduces now to solving the non-linear system $R(\hat\sigma, \hat\lambda) = 0$. Since the solution of this equation is computationally expensive, we perform one step of Newton's method in each alternating iteration.

As before, the definition of the active regions R_i^ε requires a suitable initial guess for the active phase fields in each point. To this end, we apply the same routine as before, precomputing an initial mass flux σ^0 and defining the initial regions via (5.9). The procedure is summarized in Algorithm 8.

Algorithm 8 Minimization for $N > 1$, $a_0 < \infty$

```
function MPFSD($\varepsilon_{\text{start}}, \varepsilon_{\text{end}}, N_{\text{iter}}, a_1, \dots, a_N, b_1, \dots, b_N, \mu_+, \mu_-, \rho_{\varepsilon_{\text{end}}}$)
    Set $f_\varepsilon = (\mu_+ - \mu_-) * \rho_{\varepsilon_{\text{end}}}$
    Set $(\sigma^0, \cdot, \lambda^0) = SPFS(\varepsilon_{\text{start}}, \varepsilon_{\text{end}}, N_{\text{iter}}, a_1, b_1, \mu_+, \mu_-, \rho_{\varepsilon_{\text{end}}})$
    Compute regions $R_0^\varepsilon, \tilde R_1^\varepsilon, \dots, \tilde R_N^\varepsilon$ via (5.9)
    for $j = 1, \dots, N_{\text{iter}}$ do
        Set $\varepsilon_j = \varepsilon_{\text{start}} - (j-1)\frac{\varepsilon_{\text{start}} - \varepsilon_{\text{end}}}{N_{\text{iter}} - 1}$
        Set $\varphi_i^j$ as the minimizer of (5.10) for given fixed $\sigma = \sigma^{j-1}$, $i = 1, \dots, N$
        Update regions $R_0^\varepsilon, \tilde R_1^\varepsilon, \dots, \tilde R_N^\varepsilon$ via (5.4) and (5.9)
        Set $\gamma_\varepsilon^j = \min_{i=1,\dots,N}\left((\varphi_i^j)^2 + a_i^2\varepsilon^2/b_i\right)$
        Set $(\hat\sigma^j, \hat\lambda^j) = (\hat\sigma^{j-1}, \hat\lambda^{j-1}) - DR(\hat\sigma^{j-1}, \hat\lambda^{j-1})^{-1} R(\hat\sigma^{j-1}, \hat\lambda^{j-1})$ for $\gamma_\varepsilon = \gamma_\varepsilon^{j-1}$
    end for
end function
return $\sigma^{N_{\text{iter}}}, \varphi_1^{N_{\text{iter}}}, \dots, \varphi_N^{N_{\text{iter}}}$
```

5.3.3. Discrete Γ-convergence

By employing the phase field model (5.1) in order to gain an approximation of the branched transport and urban planning functional, we have indirectly assumed that the proposed Γ-convergence result also holds for the discretized energy. However, this is not clear a priori. It is a well-known fact that the Γ-limit of the Ambrosio–Tortorelli functional depends on the choice of the discrete grid size. Even more, in [6] the authors developed a quantitative result for the asymptotic behaviour of a finite difference discretization of the latter, if both the Γ-convergence parameter ε and the mesh size δ tend to zero. Roughly spoken, their main result states that the Γ-convergence result can be maintained if and only if $\frac{\delta(\varepsilon)}{\varepsilon} \to 0$ if $\varepsilon \to 0$, where $\delta(\varepsilon)$ describes the mesh size as a function of ε with $\delta(\varepsilon) \to 0$ if $\varepsilon \to 0$. By replacing the regular grid by a discretization on random point sets, in [7] the authors could extend the Γ-convergence result of Ambrosio–Tortorelli to the Mumford–Shah functional even in the case $\frac{\delta(\varepsilon)}{\varepsilon} \to C > 0$ for $\varepsilon \to 0$. For the case of finite elements with piecewise linear basis functions, a similar result has been proposed in [8].

5.3.4. Results

All algorithms were implemented in MATLAB©. We first computed a triangulation $\mathcal{T}_h$ of Ω by defining a quadrilateral grid with mesh size h and subdividing every square into two triangles. Then we tested the proposed methods for different parameter sets (where we used the rescaled version (5.8) of the phase field energy).

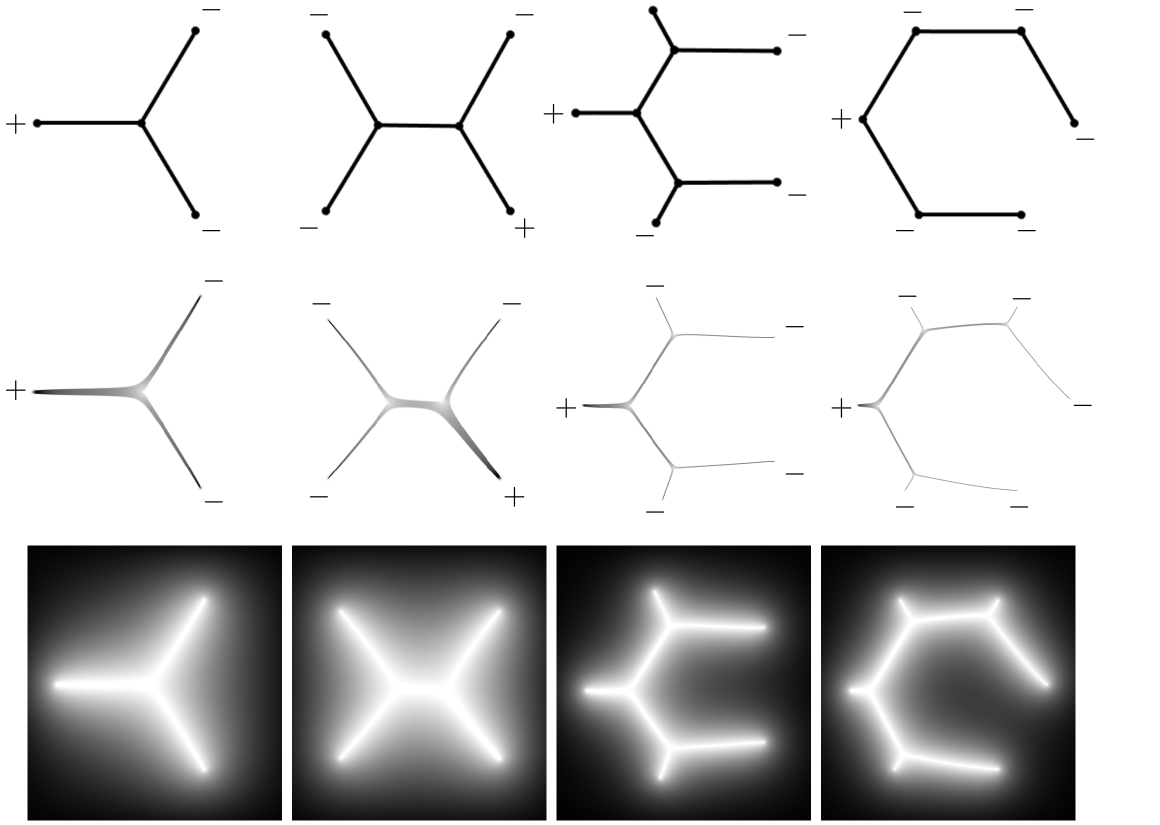


Figure 5.2.: Optimal transportation networks from a single source to a number of identical sinks at the corners of a regular polygon for parameters $N = 1$, $a_1 = 0.05$, $b_1 = 1$, $\varepsilon = 0.005$. Top row: Exact solutions obtained via optimization of the branching points. Middle row: Support of the numerically computed mass flux σ. Bottom row: Numerically computed phase field φ.

In order to compare the method with the one previously described in [24] for the Steiner case, as a first test we computed the optimal network in case of $N = 1$ and $a_0 = \infty$ for a number of three, four, five or six evenly distributed points on a circle. In this case, the applied method should yield the same results as presented in [24]. Figure 5.2 shows that the results match the exact solutions of the minimal Steiner tree problem in case of $a_1 = 0.05$ and $b = 1$. For larger values of a_1, the network costs depend more on the amount of transported mass, thus the results become more asymmetric, which is shown in Figure 5.3. Notice that due to the energy rescaling, the slope of the optimal phase field φ does now depend on the choice of a_i and b_i, which describes the difference between the obtained phase fields for $a_1 = 0.05$ and $a_1 = 1$.

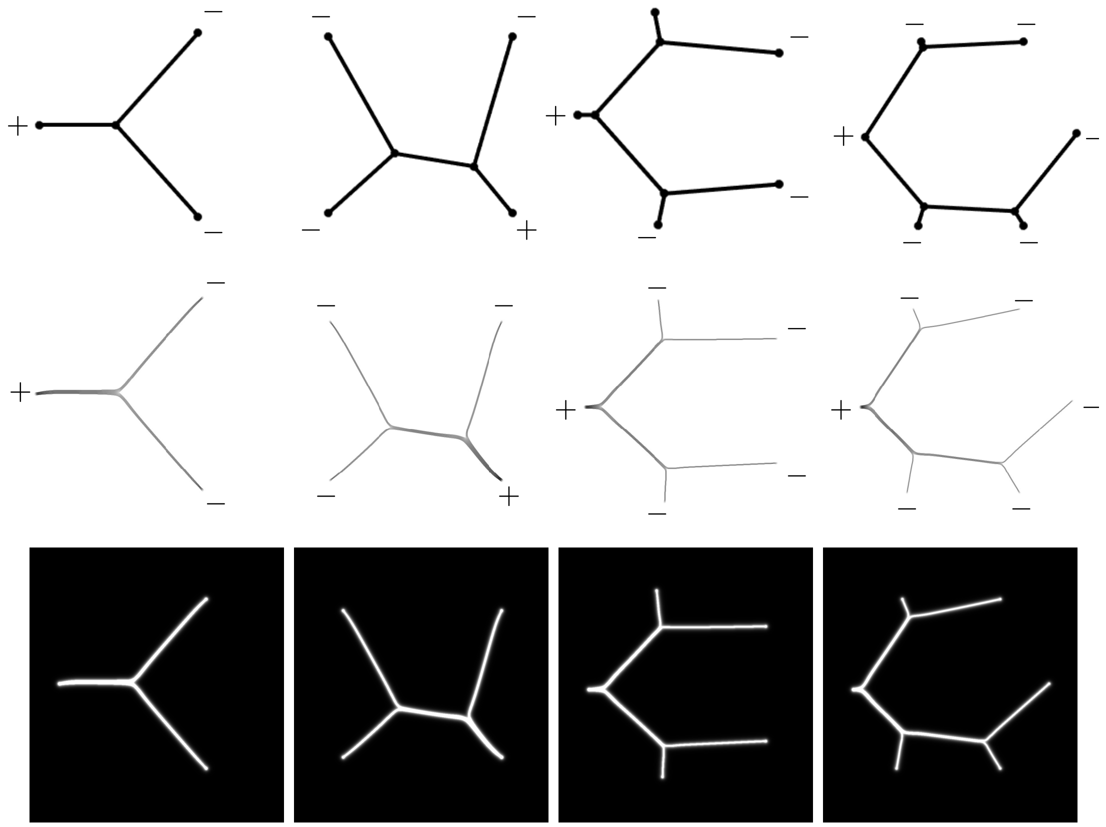


Figure 5.3.: Optimal transportation networks from a single source to a number of identical sinks at the corners of a regular polygon for parameters $N = 1$, $a_1 = 1$, $b_1 = 1$, $\varepsilon = 0.005$. Top row: Exact solutions obtained via optimization of the branching points. Middle row: Support of the numerically computed mass flux σ. Bottom row: Numerically computed phase field φ.

Next, we tested two more complex settings, where different phase fields can be active. The first one consists of four evenly spaced sources and four evenly spaced sinks, such that the optimal network topology may vary between four straight lines and one single tree (compare the results of the functional lifting approach from Section 4.3.4). The second one approximates the transport from a single source in the middle and a number of evenly distributed sinks on the boundary of a circle. We first computed the optimal network in case of $N = 1$ (Figure 5.4). Afterwards, we set $N = 3$ and allowed three different phase

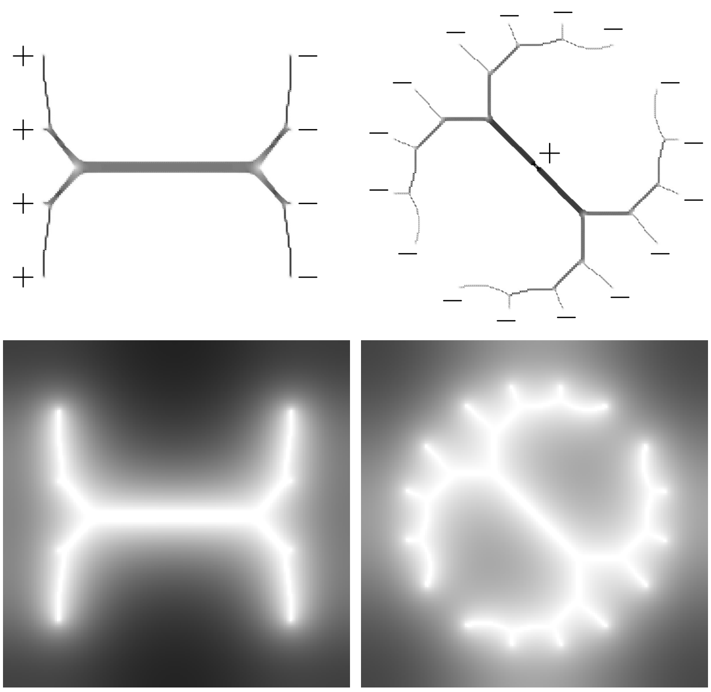


Figure 5.4.: Numerically computed mass flux and phase field for parameters $N = 1$, $a_1 = 0.05$, $b_1 = 1$, $\varepsilon = 0.005$. Left column: Transport from 4 sources to 4 sinks with equal distance. Right column: Transport from a single source in the middle to 16 evenly spaced sinks on the boundary of a circle.

fields to be active along the network (Figure 5.5).
The same settings can be applied to the case $a_0 < \infty$ such that in some parts of the network, no phase field will be active. We tested the case $N = 2$ and $a_0 < \infty$ with the same sources and sinks as before. The result is shown in Figure 5.6.
Finally, we simulated the situation where $a_0 < \infty$ and the sources and sinks are not concentrated on a finite number of points, but include a continuous part. In our example, the transport takes place between a single source in the middle and a spatially uniform sink on the boundary of a circle. Figure 5.7 shows that indeed only a part of the transport network is covered by a phase field, whereas the rest is transported off-network by travelling expenses of a_0 per unit mass.

5.3.5. Phase field locking

As pointed out in the description of the optimization strategy, for small ε the energy functional is quite far from being convex and as a consequence, the numerical method easily gets stuck in local minima. Additionally, the ability of the phase field to “move” within the image domain is influenced also by the relation between the Γ-convergence parameter ε and the mesh size δ. Roughly spoken, for ε too small compared to δ, the

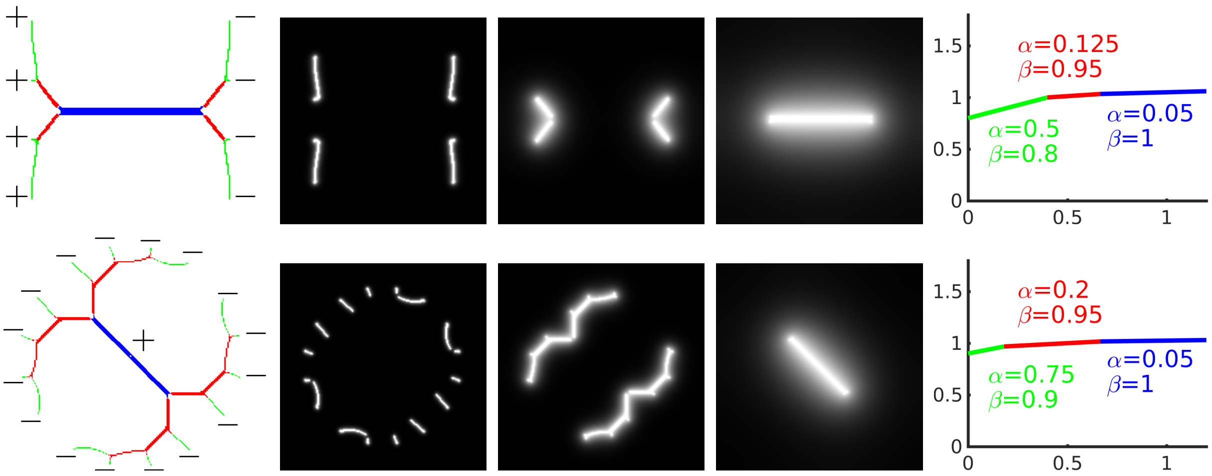


Figure 5.5.: Numerically computed mass flux and phase fields for μ_+, μ_- as in Figure 5.4 for $N = 3$, $\varepsilon = 0.005$ and for the cost functions shown on the right. Left column: Optimal mass flux σ, where the colour indicates which phase field is active. Columns 2 to 4: Optimal phase fields $\varphi_1, \varphi_2, \varphi_3$. Right column: Cost function associated with the displayed mass flux.

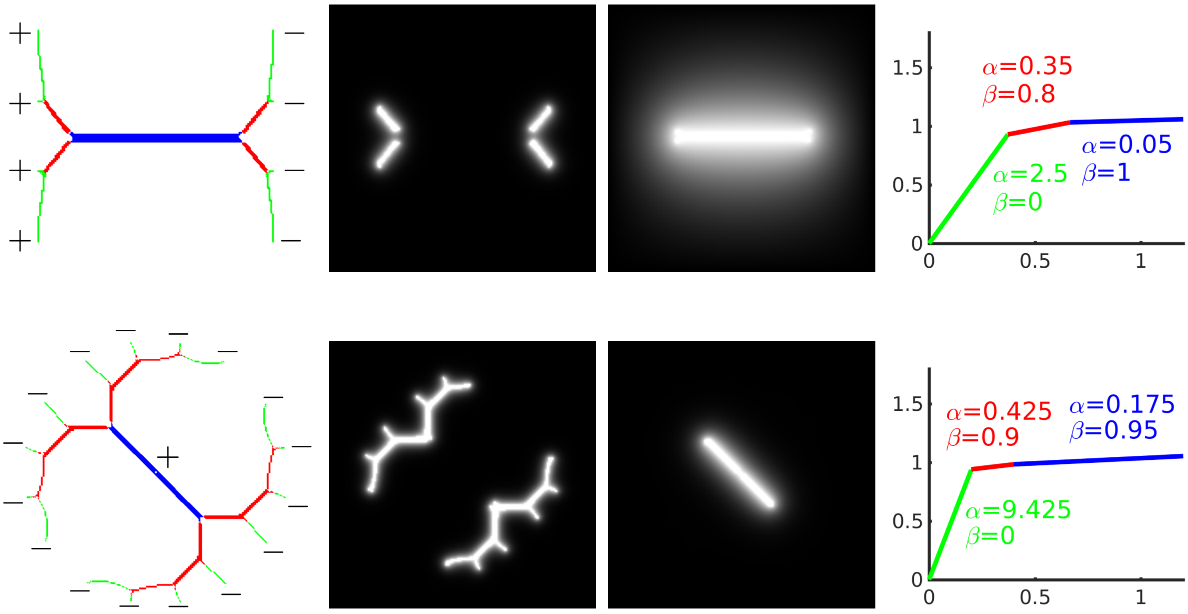


Figure 5.6.: Numerically computed mass flux and phase fields for μ_+, μ_- as in Figure 5.4 for $N = 2$, $a_0 < \infty$, $\varepsilon = 0.005$ and for the cost functions shown on the right. Left column: Optimal mass flux σ, where the colour indicates which phase field is active (green corresponds to the region where no phase field is active). Columns 2 to 3: Optimal phase fields φ_1, φ_2. Right column: Cost function associated with the displayed mass flux.

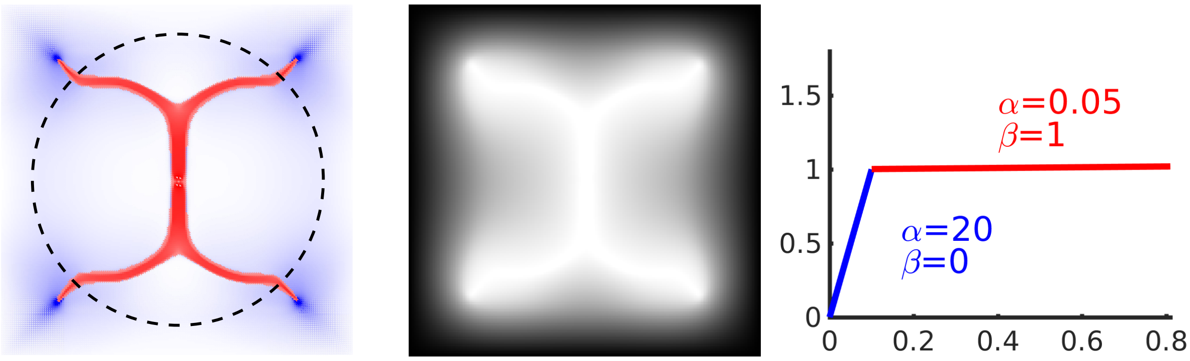


Figure 5.7.: Numerically computed mass flux and phase field for transport from a single source in the middle to a spatially uniform sink on the boundary of a circle for $N = 1$, $a_0 < \infty$, $\varepsilon = 0.005$ and the cost function shown on the right. Left: Optimal mass flux σ. The circle indicates the local of the sink. Middle: Optimal phase field φ. Right: Cost function associated with the displayed mass flux.

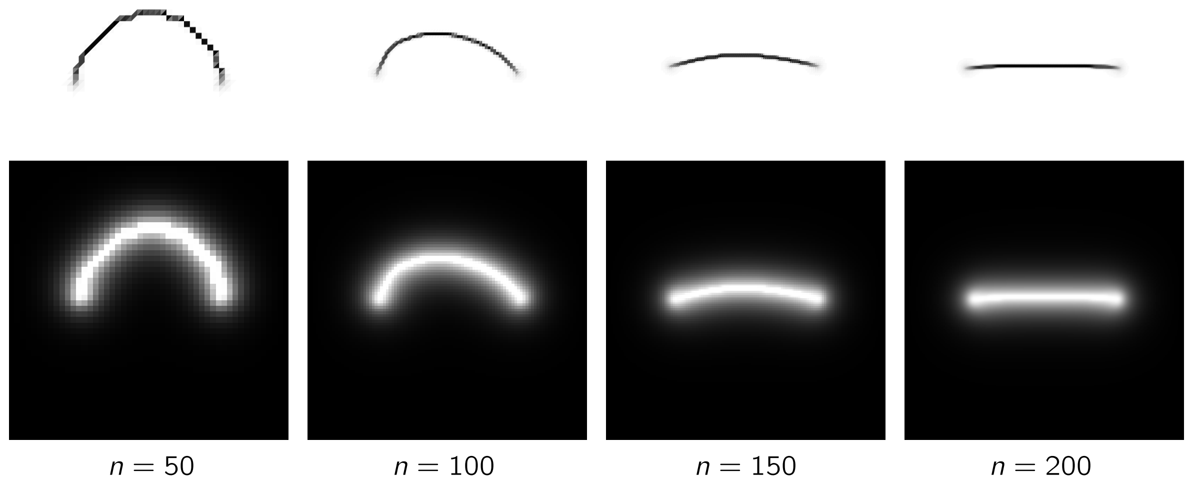
n	50	100	150	200
iterations	27	2438	3414	4915
energy diff.	$3.65 \cdot 10^{-9}$	$9.52 \cdot 10^{-9}$	$8.88 \cdot 10^{-9}$	$4.18 \cdot 10^{-9}$

Table 5.1.: Number of iterations and energy difference for transport between two points for $\varepsilon = 0.05$, where the iteration was stopped after the energy difference fell below 10^{-8}.

algorithm terminates at a point where φ is far from being optimal. This behaviour is denoted as *phase field locking*: The phase field is unable to relocate on the underlying grid and therefore "locked" to its current position.

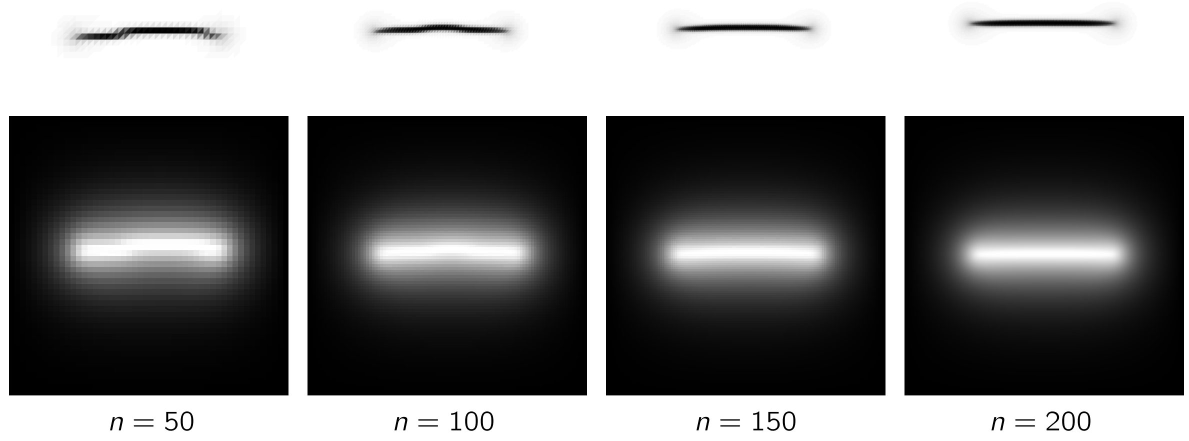
In order to investigate this problem numerically, we performed some tests simulating transport between two adjacent points with the same mass located on a horizontal line with $N = 1$ and $a_0 = \infty$. Obviously, the optimal transport network consists of a single line segment connecting the two points, consequently, for a fixed ε, the global minimizer is represented by a smoothed approximation of this line. Instead, we initialize the iteration process with a phase field and corresponding mass flux satisfying $\operatorname{div} \sigma = f_\varepsilon$ and consisting of a smoothed half-circle in the upper half of the image domain. With this as a starting point, we performed simulations with fixed $\varepsilon = 0.05$ and $\varepsilon = 0.1$ respectively as well as different image resolutions by creating a regular triangulation of the image domain $(0, 1)^2$ containing n^2 grid nodes with $n = 50, 100, 150, 200$. The results are displayed in Figures 5.8 and 5.9.

In the above experiment, the iteration process was stopped if the energy difference between two consecutive iterations fell below 10^{-8}, such that as a consequence, the energy and the corresponding variables did not change any more. Tables 5.1 and 5.2 show the number of iterations up to this point as well as the final energy difference. The results clearly show the effect of a different relation between ε and the grid size $\delta = \frac{1}{n-1}$. While for a small ε,



$n = 50$ $n = 100$ $n = 150$ $n = 200$

Figure 5.8.: Numerically computed mass flux and phase field for transport between two points located on a horizontal line with $\varepsilon = 0.05$ for different grid resolutions. The initial phase field was set to a half circle in the upper half of the image domain.



$n = 50$ $n = 100$ $n = 150$ $n = 200$

Figure 5.9.: Numerically computed mass flux and phase field for transport between two points located on a horizontal line with $\varepsilon = 0.1$ for different grid resolutions. The initial phase field was set to a half circle in the upper half of the image domain.

a smaller grid size is needed to prevent phase field locking, a relatively larger ε allows for simulations on a coarser grid. This effect could be largely avoided in various ways: By starting with a large ε and slowly decreasing its value during the iterations process, one can slightly relax the problem and still benefit from a sharp phase transition in the final

n	50	100	150	200
iterations	865	389	553	579
energy diff.	$1.67 \cdot 10^{-9}$	$7.76 \cdot 10^{-9}$	$8.09 \cdot 10^{-9}$	$8.41 \cdot 10^{-9}$

Table 5.2.: Number of iterations and energy difference for transport between two points for $\varepsilon = 0.1$, where the iteration was stopped after the energy difference fell below 10^{-8}.

result. Other approaches could be an adaptive grid refinement of regions where at least one phase field is away from 1 or the construction of a suitable initialization. Note that the problem of phase field locking is also present in the Ambrosio–Tortorelli approximation of the Mumford–Shah image segmentation functional. However, its effect is mostly prevented by initializing the phase field with the gradient of the given image, which makes too much “movement” most commonly unnecessary.

5.3.6. Discussion and outlook

We have proposed a phase field approximation of the generalized urban planning functional, which admits some useful properties such as a Γ-convergence result and is therefore of sustainable use for numerically computing optimal network structures. The functional does not only cover the classical urban planning case (and thus represents one of the first methods capable of providing optimal urban planning networks to the best of our knowledge), by approximating any other concave cost functional by piecewise linear functions one could also tackle more general optimal network problems.

Although phase field models are known to get stuck in local minima due to the non-convexity of the energy landscape, for large ε, the phase field energy is closer to being convex. Thus, by a slow decrease of the relaxation parameter ε, the algorithm hopefully avoids most of them and finally reaches the global optimum, provided that the grid size is small enough in the limit. This automatically suggests the usage of an adaptive grid refinement, which would ensure the maintenance of the Γ-convergence result in the discrete case.

6 Conclusion and outlook

The major task of this work was to help establish a deeper understanding of the different features and challenges of numerical methods for optimal transportation network problems as well as to provide some novel approaches. Starting with a short review of a variety of existing methods from the literature, we described and discussed two different treatments which both proved to be of practical use for numerical simulations.

In Chapter 4, we investigated an image-based reformulation of the branched transport and urban planning energies. As a preliminary step, we analysed the effect of a convex relaxation approach via functional lifting of the Mumford–Shah-type functionals. Afterwards we presented two different discretization approaches. A straightforward initial concept was built on a finite difference discretization scheme of the convex saddle point problem. To simulate the transport in a variety of settings, we implemented a primal-dual algorithm combined with an intrinsic iterative projection to handle the large set of involved non-local constraints. In most cases, the results were able to recover the optimal transportation network, which was shown by a study of the influence of the branching parameters on the network topology. However, the simple implementation comes along with the drawback of computational inefficiency, which in particular becomes noticeable in the simulation of complex network structures. For this reason, we developed a novel advanced discretization scheme based on adaptive finite elements of triangular prism grid. While the adaptivity was capable of handling the high dimensionality of the problem, the triangular grid structure enabled an efficient treatment of the non-local constraint set. The benefit was underlined by the simulations, where a high grid resolution could be achieved without an unsustainable increase of the computation time.

Another numerical concept covering the generalized urban planning problem was presented in Chapter 5 as an extension to the phase field approximations of the branched transport as well as the Steiner tree problem. The approach enhances the existing models by introducing multiple phase fields corresponding to each affine segment of the generalized urban planning cost. We developed three different algorithmic frameworks distinguishing between the cases of a single phase field, multiple phase fields as well as the appearance of a diffuse component. The simulations yielded adequate approximations of the optimal network structures. To finalize the chapter, we discussed some problematic aspects of a phase field relaxation approach in the context of numerical simulations.

Although the results of this thesis prove that there exists a multitude of suitable numerical treatments of transportation network problems, all existing methods certainly leave some room for improvement. The handling of real data sets such as a population and workplace density of a city for the purpose of designing a public transportation network is still far from being satisfactory. Furthermore, most methods, although providing a good approximation, cannot claim to end up with the globally optimal solution. Due to a lack of convexity of the investigated models, any numerical method necessarily either suffers from a variety of local minimizers or from a loss of accuracy by a convex relaxation. For this reason, the research concerning numerical methods for transportation networks is still of high relevance. As part of this ongoing research, the ideas presented in this thesis could be further improved in several points. In the following, we want to outline some topics which seem to be of particular importance to us.

An interesting aspect of the functional lifting approach is given by the loss of accuracy of the convex relaxation, which we investigated within this work. In order to answer the question whether the relaxation yields the convex envelope of the original non-convex energy, a rigorous proof of certain necessary criteria is still a subject of future work. Another improvement would be a fully optimized, efficiency-based implementation of the adaptive grid approach in order to maximally exploit the benefits of adaptivity. In this spirit, one could also consider a GPU-based implementation within the CUDA framework, which could further decrease the runtime and therefore enable even more complex network situations.

The presented phase field approximation result comes along with similar problems as other methods of this type. Here, an interesting feature also present in case of the famous Ambrosio–Tortorelli model is the dependency on the relation between the grid size and the Γ-convergence parameter. While we broached the problem of the numerically computed phase field being "locked" at some local minimum within the iteration process, the development of a criterion for the relation between the involved parameters is still missing. Besides, a more efficient implementation could be achieved by applying an adaptive refinement approach. Finally, a more elaborate algorithmic framework capable of avoiding local minima combined with a rigorous proof of convergence could possibly advance the numerical results.

A
Construction of a vector field

In this chapter, we aim at completing the considerations concerning the tightness of the convex relaxation of the lifted branched transport problem in Section 4.2.1, Example 4.2.1, by constructing an almost optimal divergence-free vector field $\hat{\phi} \in \mathcal{K}_1$. To be precise, let us fix some notation and values for this particular example.
Let $\Omega = [0,1]^2$ be the image domain and set the initial and final measures as $\mu_+ = \frac{1}{2}(\delta_{P_1} + \delta_{P_2})$, $\mu_- = \frac{1}{2}(\delta_{Q_1} + \delta_{Q_2})$ with

$$P_1 = (\tfrac{1}{4}, 0)^T, P_2 = (\tfrac{3}{4}, 0)^T, Q_1 = (\tfrac{1}{4}, 1)^T, Q_2 = (\tfrac{3}{4}, 1)^T.$$

We recall the definition of the image-related cost functional as well as the relaxed version as

$$\tilde{\mathcal{M}}^{\alpha}(u) = \int_{S_u \cap \Omega} [u]^{\alpha} \mathrm{d}\mathcal{H}^1(x) + \iota_0((\nabla u \mathcal{L}^2 + D^c u) \llcorner \Omega), \quad J(v) = \sup_{\phi \in \mathcal{K}_1} \int_{\Omega \times \mathbb{R}} \phi \cdot \mathrm{d}Dv + \iota_{\mathcal{C}}(v)$$

with $\mathcal{C}, \mathcal{K}_1$ as in (4.1), (4.2) respectively. Now let $\hat{\alpha}$ be chosen such that

$$\min_{u \in \mathcal{A}_u(\mu_+, \mu_-)} \tilde{\mathcal{M}}^{\hat{\alpha}}(u) = \tilde{\mathcal{M}}^{\hat{\alpha}}(u_1) = \tilde{\mathcal{M}}^{\hat{\alpha}}(u_2^{\hat{\alpha}})$$

for $u_1, u_2^{\hat{\alpha}} \in \mathcal{A}_u(\mu_+, \mu_-)$ as constructed in Example 4.2.1 (note that $u_2^{\hat{\alpha}}$ depends on the branching parameter $\hat{\alpha}$). The aim of this chapter is to construct a divergence-free $\hat{\phi}_{\hat{\alpha}} \in \mathcal{K}_1^{\hat{\alpha}}$ such that

$$\sup_{\phi \in \mathcal{K}_1^{\hat{\alpha}}} \int_{\Omega \times \mathbb{R}} \phi \cdot \mathrm{d}D1_{u_1} = \int_{\Omega \times \mathbb{R}} \hat{\phi}_{\hat{\alpha}} \cdot \mathrm{d}D1_{u_1}, \tag{A.1}$$

where we added the branching parameter $\hat{\alpha}$ to the definitions of $\mathcal{K}_1$, $\hat{\phi}$ and u_2 to emphasize their dependence on $\hat{\alpha}$. For the sake of readability, we set $1_{u_2}^{\hat{\alpha}}$ as the characteristic function of the subgraph of $u_2^{\hat{\alpha}}$. In the following, we are going to drop the requirement of smoothness in the constraint set $\mathcal{K}_1^\alpha$ and define

$$\mathcal{K}_1^\alpha = \{\phi = (\phi^x, \phi^s) \in L^\infty(\Omega \times \mathbb{R}, \mathbb{R}^2 \times \mathbb{R}) \ : \ \text{div } \phi \in L^\infty(\Omega \times \mathbb{R}, \mathbb{R}^2 \times \mathbb{R}),\ \phi^s \geq 0,$$
$$\Big| \int_{s_1}^{s_2} \phi^x \ \mathrm{d}s \Big| \leq |s_2 - s_1|^\alpha \ \forall\ x \in \Omega, s_1, s_2 \in \mathbb{R}, s_1 < s_2\}. \tag{A.2}$$

Since the functional $F(\phi) := \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}D1_{u_1}$ is continuous with respect to the norm $\|\phi\|_{L^\infty} + \|\text{div } \phi\|_{L^\infty}$, the supremum is not changed by replacing the definition of $\mathcal{K}_1^\alpha$ in (4.2) by (A.2). Let us first state a preliminary result.

Lemma A.0.1. *The divergence-free $\hat{\phi}_{\hat{\alpha}} \in \mathcal{K}_1^{\hat{\alpha}}$ satisfying equation* (A.1) *also satisfies*

$$\sup_{\phi \in \mathcal{K}_1^{\hat{\alpha}}} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}D1_{u_2}^{\hat{\alpha}} = \int_{\Omega\times\mathbb{R}} \hat{\phi}_{\hat{\alpha}} \cdot \mathrm{d}D1_{u_2}^{\hat{\alpha}}.$$

Proof. By abbreviating $\hat{\phi} = \hat{\phi}_{\hat{\alpha}}$, $u_2 = u_2^{\hat{\alpha}}$ and $1_{u_2} = 1_{u_2}^{\hat{\alpha}}$, we have

$$\begin{aligned}
\int_{\Omega\times\mathbb{R}} \hat{\phi} \cdot \mathrm{d}D1_{u_2} &= \int_{R_2} \hat{\phi} \cdot \nu \ \mathrm{d}\mathcal{H}^2 - \underbrace{\int_{\Omega\times\mathbb{R}} 1_{u_2} \ \text{div}\hat{\phi} \ \mathrm{d}x\mathrm{d}s}_{=0} \\
&= \int_{R_2} \hat{\phi} \cdot \nu \ \mathrm{d}\mathcal{H}^2 = \int_{R_1} \hat{\phi} \cdot \nu \ \mathrm{d}\mathcal{H}^2 \\
&= \int_{R_1} \hat{\phi} \cdot \nu \ \mathrm{d}\mathcal{H}^2 - \underbrace{\int_{\Omega\times\mathbb{R}} 1_{u_1} \ \text{div}\hat{\phi} \ \mathrm{d}x\mathrm{d}s}_{=0} \\
&= \int_{\Omega\times\mathbb{R}} \hat{\phi} \cdot \mathrm{d}D1_{u_1} = \sup_{\phi\in\mathcal{K}_1} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}D1_{u_1} \\
&= J(1_{u_1}) = J(1_{u_2}) = \sup_{\phi\in\mathcal{K}_1} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}D1_{u_2},
\end{aligned}$$

where

$$R_1 = \{(x,s) \in \partial\Omega \times \mathbb{R} \ : \ 1_{u_1}(x,s) = 1\} = \{(x,s) \in \partial\Omega \times \mathbb{R} \ : \ 1_{u_2}(x,s) = 1\} = R_2$$

and ν denotes the outer unit normal to $R_1 = R_2$. □

In other words, Lemma A.0.1 states that the constructed $\hat{\phi}_{\hat{\alpha}}$ also realizes the supremum with respect to $1_{u_2}^{\hat{\alpha}}$. As a consequence, $\hat{\phi}_{\hat{\alpha}}$ needs to "fit" both topologies given by the discontinuity sets of the minimizing images u_1 and $u_2^{\hat{\alpha}}$.
Let us investigate the structure of an optimal ϕ. For the particular example of μ_+ and μ_-

as defined above, we have

$$u(\mu_+, \mu_-)(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 \leq \frac{1}{4}, \\ \frac{1}{2} & \text{if } x_1 > \frac{1}{4}, x_1 \leq \frac{3}{4}, \\ 0 & \text{otherwise} \end{cases}$$

for $(x_1, x_2) \in \Omega$ and $u_1(x_1, x_2) = u(\mu_+, \mu_-)(x_1, x_2)$. Noticing that u_1 and 1_{u_1} are piecewise constant and in particular $1_{u_1} \in SBV(\Omega \times \mathbb{R})$, we have

$$D1_{u_1} = \nu_{\Gamma_{u_1}} \cdot \mathcal{H}^2 \llcorner \Gamma_{u_1}, \;\nu_{\Gamma_{u_1}}(x, s) = \begin{cases} (0, -1)^T & \text{for } x \in \Omega \setminus S_{u_1}, \\ (\nu_{u_1}(x), 0)^T & \text{for } x \in S_{u_1}, \end{cases}$$

where Γ_{u_1} is the singular set of 1_{u_1} and $\nu_{\Gamma_{u_1}}$ its outer unit normal (extended to $\Omega \times \mathbb{R}$). In addition, we have $S_{u_1} = \left(\{\frac{1}{4}\} \times [0, 1]\right) \cup \left(\{\frac{3}{4}\} \times [0, 1]\right)$ and $\nu_{u_1} = (-1, 0)^T$ on S_{u_1}, thus we obtain for $\phi = (\phi^x, \phi^s)^T$

$$\begin{aligned} &\int_{\Omega\times\mathbb{R}} \phi(x, s) \cdot \mathrm{d}D1_{u_1}(x, s) = \int_{\Gamma_{u_1}} \phi(x, u_1(x)) \cdot \nu_{\Gamma_{u_1}}(x) \mathrm{d}\mathcal{H}^2(x) \\ = &\int_{\Omega\setminus S_{u_1}} -\phi^s(x, u_1(x)) \, \mathrm{d}x + \int_{S_{u_1}} \left(\int_{u_1^-(x)}^{u_1^+(x)} \phi^x(x, s) \, \mathrm{d}s \right) \cdot \nu_{u_1}(x) \, \mathrm{d}\mathcal{H}^1(x) \\ = &\int_{\Omega\setminus S_{u_1}} -\phi^s(x, u_1(x)) \, \mathrm{d}x + \int_0^1 \left(-\int_{\frac{1}{2}}^1 \phi^{x_1}(\tfrac{1}{4}, x_2, s) \, \mathrm{d}s - \int_0^{\frac{1}{2}} \phi^{x_1}(\tfrac{3}{4}, x_2, s) \, \mathrm{d}s \right) \mathrm{d}x_2 \end{aligned} \quad (A.3)$$

with $\phi^x = (\phi^{x_1}, \phi^{x_2})^T$. The remaining task is now to find a divergence-free $\hat{\phi}_{\hat{\alpha}} \in \mathcal{K}_1^{\hat{\alpha}}$ which maximizes the last expression in (A.3). Unfortunately, we were not able to construct this $\hat{\phi}_{\hat{\alpha}}$ explicitly, in particular, we could not find any ϕ which satisfies all constraints and yields the correct value of (A.3) at the same time. Instead, in the following we are going to construct a ϕ which leads to the correct value of (A.3) and incorporates all necessary information about the optimal ϕ as derived above, but slightly violates the constraints in $\mathcal{K}_1^{\hat{\alpha}}$. Although this ϕ is not the truly optimal choice, it can nevertheless be used to obtain an estimate on the quality of the solutions u_1 and $u_2^{\hat{\alpha}}$. On the one hand, it can be shown to satisfy the constraints in $\mathcal{K}_1^{\alpha}$ for a slightly different $\alpha = \hat{\alpha} + \delta$ (which will be specified later). On the other hand, it is still close to the truly optimal $\hat{\phi}_{\hat{\alpha}}$ and therefore allows an estimate on the primal-dual gap for the solutions 1_{u_1} and $1_{u_2}^{\hat{\alpha}}$.

Let $\alpha = \hat{\alpha} + \delta$ for some $\delta > 0$. We aim at constructing $\hat{\phi}_\alpha = (\hat{\phi}_\alpha^x, \hat{\phi}_\alpha^s)$ which maximizes (A.3) and satisfies all inequality constraints in $\mathcal{K}_1^\alpha$. To this end, we set $\hat{\phi}_\alpha^s = 0$. $\hat{\phi}_\alpha^x$ is bounded by the convex constraints $|\int_{s_1}^{s_2} \hat{\phi}_\alpha^x \, \mathrm{d}s| \leq |s_2 - s_1|^\alpha$ for all $s_1 < s_2$ and $x \in \Omega$. We define

$$\hat{\phi}^x_\alpha(x_1, x_2, s) = \begin{cases} 0 & \text{if } s \notin [0, 1], \\ \varphi_1(x_1, x_2) & \text{if } s \in [\frac{1}{2}, 1], \\ \varphi_2(x_1, x_2) & \text{if } s \in [0, \frac{1}{2}) \end{cases}$$

for two functions $\varphi_1, \varphi_2 : \Omega \to \mathbb{R}^2$. For symmetry reasons, we can restrict ourselves to the construction of φ_1 and set $\varphi_2(x_1, x_2) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right) \varphi_1(1 - x_1, x_2)$. The additional constraint of φ_1 being divergence-free can be achieved by applying a trick: Instead of constructing φ_1 directly, we define $v \in C^{0,1}(\Omega)$ and set $\varphi_1 = Dv^\perp$. Then by Rademacher's theorem (see for instance [32], Theorem 3.1.6) v is differentiable almost everywhere and we have

$$\operatorname{div} \varphi_1 = \operatorname{div}\left(Dv^\perp\right) = \operatorname{div}\begin{pmatrix} -\frac{\partial v}{\partial x_2} \\ \frac{\partial v}{\partial x_1} \end{pmatrix} = 0$$

and $\hat{\phi}_\alpha \in \mathcal{K}^\alpha_1$ as requested.
We can isolate the region where v needs to be constructed explicitly even further by the following considerations. Again, due to symmetry reasons, we set $\varphi_1(x_1, x_2) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right) \varphi_1(x_1, 1 - x_2)$ for all $x \in [0, 1] \times (\frac{1}{2}, 1]$. Additionally, we define

$$\varphi_1(x_1, x_2) = \begin{cases} \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right) \varphi_1(\frac{1}{2} - x_1, x_2) & \text{if } x_1 < \frac{1}{4}, \\ \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right) \varphi_1(\frac{3}{4} - x_1, x_2) & \text{if } x_1 > \frac{3}{4} \end{cases}$$

and thus restrict ourselves to the subregion $\tilde{\Omega} = [\frac{1}{4}, \frac{3}{4}] \times [0, \frac{1}{2}]$. Now let $\hat{x}^\alpha = (\frac{1}{2}, \hat{x}^\alpha_2)^T$ be the lower branching point in the graph G^α_2 (also depending on the branching parameter α). Although G^α_2 is not the optimal topology for the given α, one can compute $\hat{x}^\alpha$ such that the cost of G^α_2 become minimal with respect to $\hat{x}^\alpha_2$. This leads to

$$\hat{x}^\alpha_2 = \frac{2^{\alpha-1}}{\sqrt{16 - 2^{2\alpha+2}}}.$$

Let $R = \sqrt{\frac{1}{16} + (\hat{x}^\alpha_2)^2}$ be the distance between the leftmost point P_1 of the initial measure μ_+ and $\hat{x}^\alpha$ (see Figure A.1).

To maximize (A.3) under the given constraints, φ_1 must be orthogonal to the line given by $x_1 = \frac{1}{4}$. Additionally, for $\hat{\phi}_\alpha$ to fit the topology given by u^α_2, φ_1 necessarily needs to be orthogonal to the connection between the points P_1 and $\hat{x}^\alpha$. Thus, we define

$$v(x_1, x_2) = \begin{cases} \beta\sqrt{\left(x_1 - \frac{1}{4}\right)^2 + x_2^2} & \text{if } \quad \sqrt{\left(x_1 - \frac{1}{4}\right)^2 + x_2^2} \leq R, \\ \beta b(x_1, x_2) & \text{otherwise}, \end{cases}$$

for all $(x_1, x_2) \in \tilde{\Omega}$, where $\beta = (\frac{1}{2})^{\alpha-1}$ is chosen such that the constraints for $\hat{\phi}_\alpha \in \mathcal{K}^\alpha_1$ are satisfied with equality along the lines $x_1 = \frac{1}{4}$ and between P_1 and $\hat{x}^\alpha$. Outside of the circle,

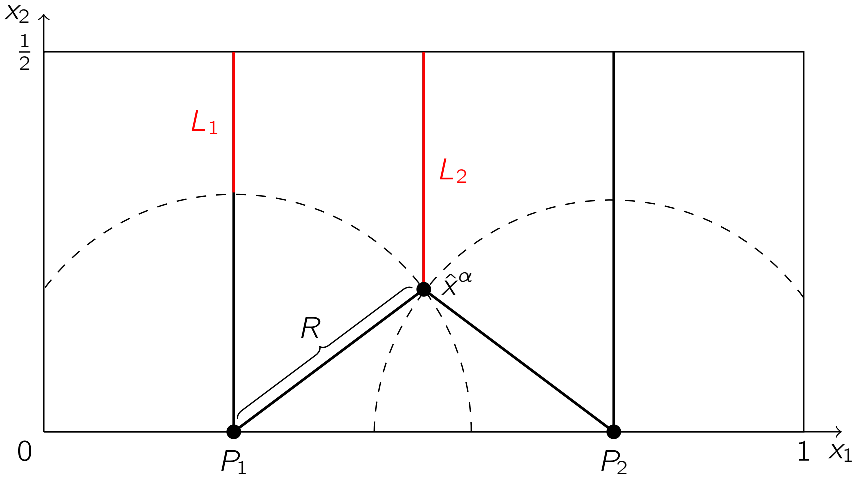


Figure A.1.: Construction of the function v on the domain $[\frac{1}{4}, \frac{3}{4}] \times [0, \frac{1}{2}]$.

we define v to be constant along ellipses with increasing semi-minor and semi-major axis $b \to 1$, $a \to \infty$ as $x_2 \to \frac{1}{2}$, which will be specified in the following. For every point (x_1, x_2) with $\sqrt{(x_1 - \frac{1}{4})^2 + x_2^2} > R$, we define $a(x_1, x_2), b(x_1, x_2)$ such that

$$\frac{(x_1 - \frac{1}{4})^2}{a(x_1, x_2)^2} + \frac{x_2^2}{b(x_1, x_2)^2} = 1. \tag{A.4}$$

Moreover, the amount of mass flowing through the line L_1 needs to be uniformly distributed along the line L_2 in order to maximize $\int_{\Omega \times \mathbb{R}} \phi \cdot \mathrm{d}D1^\alpha_{u_2}$ under the constraint set (which can be seen by repeating the computations in equation (A.3) with u_2^α instead of u_1). In other words, if μ (λ respectively) denotes the x_2-coordinate of the intersection of an ellipse with the line L_1 (L_2 respectively), the proportion of $l([0, \mu] \cap L_1)/l(L_1)$ must be the same as $l([0, \lambda] \cap L_2)/l(L_2)$, where $l(\cdot)$ denotes the length of a line segment. With $\mu = b(x_1, x_2)$ and λ given by the equation

$$\frac{1}{16a(x_1, x_2)^2} + \frac{\lambda^2}{b(x_1, x_2)^2} = 1,$$

this leads to the condition

$$\frac{b(x_1, x_2) - R}{\frac{1}{2} - R} = \frac{\lambda - \hat{x}_2^\alpha}{\frac{1}{2} - \hat{x}_2^\alpha} \quad \Leftrightarrow \quad \lambda = \hat{x}_2^\alpha + \frac{\frac{1}{2} - \hat{x}_2^\alpha}{\frac{1}{2} - R} \left(b(x_1, x_2) - R \right).$$

Together, this implies

$$a(x_1, x_2)^2 = \frac{1}{16} \cdot \frac{b(x_1, x_2)^2}{b(x_1, x_2)^2 - \left(\hat{x}_2^\alpha + C \left(b(x_1, x_2) - R \right) \right)^2},$$

where we abbreviated $C = \frac{1/2 - \hat{x}_2^\alpha}{1/2 - R}$. Inserting this into equation (A.4) yields a formula for $b(x_1, x_2)$,

$$b(x_1,x_2)^2 = x_2^2 + 16(x_1 - \tfrac{1}{4})^2 \left(b(x_1,x_2)^2 - (\hat{x}_2^\alpha + C(b(x_1,x_2) - R))^2 \right).$$

Solving the quadratic equation for b leads to

$$b(x_1,x_2) = -\frac{p(x_1)}{2c(x_1)} + \sqrt{\left(\frac{p(x_1)}{2c(x_1)}\right)^2 - \frac{q(x_1,x_2)}{c(x_1)}}, \tag{A.5}$$

where

$$\begin{aligned} c(x_1) &= 16(1 - C^2)(x_1 - \tfrac{1}{4})^2 - 1, \\ p(x_1) &= 16C(C-1)(x_1 - \tfrac{1}{4})^2, \\ q(x_1,x_2) &= -4(C-1)^2(x_1 - \tfrac{1}{4})^2 + x_2^2. \end{aligned}$$

It is straightforward to verify that v is continuous. Additionally, φ_1 is divergence-free by definition and maximizes (A.3) under the given constraints. Thus it remains to check whether $\hat{\phi}_\alpha$ satisfies all constraints in $\mathcal{K}_1^\alpha$. To this end, we prove the following result.

Theorem A.0.2. *Let $\tilde{\alpha} > 0$ be the real root of the function $f(\alpha) := (6 - 2^{\alpha+1} - 2^{2-\alpha} + 2^{-2\alpha})^2 - 2^{2-4\alpha} + 2^{-2\alpha}$. Then $\hat{\phi}_\alpha \in \mathcal{K}_1^\alpha$ for all $\alpha \geq \tilde{\alpha}$.*

Proof. The proof consists in evaluating the inequality constraints for the constructed $\hat{\phi}_\alpha$, which leads to a condition for α. For the inequality constraints, due to Theorem 4.2.3 it suffices to verify

$$\tfrac{1}{2}|\varphi_1(x_1,x_2)| \leq \left(\tfrac{1}{2}\right)^\alpha \qquad \forall\ (x_1,x_2) \in [\tfrac{1}{4}, \tfrac{3}{4}] \times [0, \tfrac{1}{2}], \tag{A.6}$$

$$\tfrac{1}{2}|\varphi_1(x_1,x_2) + \varphi_2(x_1,x_2)| \leq 1 \qquad \forall\ (x_1,x_2) \in [\tfrac{1}{4}, \tfrac{1}{2}] \times [0, \tfrac{1}{2}]. \tag{A.7}$$

The constraint evaluation involves several technical but straightforward computations, thus we only provide the main ideas of the proof. We define the region $B := \{(x_1,x_2) \in [\frac{1}{4}, \frac{3}{4}] \times [0, \frac{1}{2}] \ :\ \sqrt{(x_1 - \frac{1}{4})^2 + x_2^2} \leq R\}$. Let us regard the two constraint sets separately.

First constraint (A.6):

We distinguish between the following cases according to the definition of v:

$$\text{(a) } (x_1,x_2) \in B, \ \text{ (b) } (x_1,x_2) \notin B.$$

For $(x_1,x_2) \in B$, the constraint (A.6) is satisfied with equality by construction. For $(x_1,x_2) \notin B$, one can easily verify that the left-hand side attains its maximum in $x_1 = \frac{1}{4}$, thus we have

$$|\varphi_1(x_1,x_2)| \le |\varphi_1(\tfrac{1}{4},x_2)| = \beta\sqrt{\left(\tfrac{\partial b}{\partial x_1}(\tfrac{1}{4},x_2)\right)^2 + \left(\tfrac{\partial b}{\partial x_2}(\tfrac{1}{4},x_2)\right)^2} = \beta = \left(\frac{1}{2}\right)^{\alpha-1},$$

thus (A.6) is satisfied for all $\alpha \in (0,1)$.

Second constraint (A.7):

From the definition of φ_1, φ_2 in the construction above, we obtain

$$\begin{aligned} & \tfrac{1}{2}|\varphi_1(x_1,x_2)+\varphi_2(x_1,x_2)| \le 1 \\ \Leftrightarrow \quad & \sqrt{\left(\tfrac{\partial v}{\partial x_2}(x_1,x_2)+\tfrac{\partial v}{\partial x_2}(1-x_1,x_2)\right)^2 + \left(\tfrac{\partial v}{\partial x_1}(x_1,x_2)-\tfrac{\partial v}{\partial x_1}(1-x_1,x_2)\right)^2} \le 2 \end{aligned} \tag{A.8}$$

As before, we distinguish between the following cases:

(a) $(x_1,x_2),(1-x_1,x_2)\in B$, (b) $(x_1,x_2)\in B,(1-x_1,x_2)\notin B$, (c) $(x_1,x_2),(1-x_1,x_2)\notin B$

It is easy to verify that in case (a), (A.8) is satisfied for every $\alpha \in (0,1)$. Moreover, one can show that for α relatively close to $\hat{\alpha}$ (in particular, for $\hat{\alpha} \le \alpha \le \tilde{\alpha}$), the left-hand side of (A.8) attains its maximum within the region defined in (c). Thus, it remains to derive a condition for α from the constraint for all $(x_1,x_2),(1-x_1,x_2)\notin B$. In this region, we have $v(x_1,x_2) = \beta b(x_1,x_2)$ (same holds for $(1-x_1,x_2)$), thus by inserting the definition of b, we obtain the constraint

$$\beta\sqrt{\left(\tfrac{\partial b}{\partial x_2}(x_1,x_2)+\tfrac{\partial b}{\partial x_2}(1-x_1,x_2)\right)^2 + \left(\tfrac{\partial b}{\partial x_1}(x_1,x_2)-\tfrac{\partial b}{\partial x_1}(1-x_1,x_2)\right)^2} \le 2. \tag{A.9}$$

The left-hand side attains its maximum in $x_2 = \frac{1}{2}$, where $\frac{\partial b}{\partial x_1}(x_1,\frac{1}{2}) = \frac{\partial b}{\partial x_1}(1-x_1,\frac{1}{2})$, thus (A.9) is equivalent to

$$\begin{aligned} & \beta\left(\tfrac{\partial b}{\partial x_2}(x_1,\tfrac{1}{2})+\tfrac{\partial b}{\partial x_2}(1-x_1,\tfrac{1}{2})\right) \\ = & -\beta\left(\frac{16(1-C)\left((x_1-\frac{1}{4})^2+(x_1-\frac{3}{4})^2\right)-2}{256(1-C)^2(x_1-\frac{1}{4})^2(x_1-\frac{3}{4})^2-16(1-C)\left((x_1-\frac{1}{4})^2+(x_1-\frac{3}{4})^2\right)+1}\right) \le 2, \end{aligned} \tag{A.10}$$

where we again abbreviated $C = (\frac{1}{2}-\hat{x}_2^\alpha)/(\frac{1}{2}-R)$, $R = \sqrt{\frac{1}{16}+(\hat{x}_2^\alpha)^2}$. Again, one can derive that the left-hand side of (A.10) attains its maximum in

$$x_1 = -\frac{\sqrt{\frac{C}{4}-\frac{C^2}{4}+\frac{1}{2}\sqrt{C(C-1)^3}}-C+1}{2(C-1)},$$

and inserting this into (A.10) yields

$$\frac{\sqrt{C(C-1)}}{\left(\sqrt{C(C-1)}-C\right)(1-C)} \leq \frac{4}{\beta} \Leftrightarrow 0 \leq \left(6-2^{\alpha+1}-2^{2-\alpha}+2^{-2\alpha}\right)^2 - 2^{2-4\alpha}+2^{-2\alpha} = f(\alpha). \tag{A.11}$$

One can additionally show that the function f admits only complex roots except for one. Consequently, the constraints are satisfied if $\alpha \geq \tilde{\alpha}$ for $\tilde{\alpha}$ being the real root of f. □

Remark A.0.3. One can compute the approximate value of $\tilde{\alpha} \approx 0.366006$ numerically. The value of the critical $\hat{\alpha}$ can be obtained (by setting $\mathcal{M}^{\hat{\alpha}}(G_1) = \mathcal{M}^{\hat{\alpha}}(G_2^{\hat{\alpha}})$) as the real root of the function $g(\alpha) := 2^{4-2\alpha} - 2^{6-2\alpha} + 2^{6-\alpha} + 2^{2\alpha} - 2^{\alpha+4} + 2^{2\alpha+2} - 8$ in the interval $(0,1)$, which approximately satisfies $\hat{\alpha} \approx 0.263034$. Setting $\alpha = \hat{\alpha} + \delta$, this means that we have a minimal approximate error of $\delta \approx 0.102971$, which is also reflected by the numerical experiments.

The constructed image and corresponding vector field as well as a numerically obtained solution are shown in Figure A.2. Furthermore, for the critical $\hat{\alpha}$ we obtain an upper bound on the primal-dual gap for the $\tilde{\mathcal{M}}^{\hat{\alpha}}$-minimizers u_1 and $u_2^{\hat{\alpha}}$.

Theorem A.0.4. *Let Δ_α denote the primal-dual gap for the branched transport functional lifting problem, i.e.*

$$\Delta_\alpha(\tilde{v}, \tilde{\phi}) := \sup_{\phi \in \mathcal{K}_1^\alpha} \int_{\Omega\times\mathbb{R}} \phi \cdot \mathrm{d}D\tilde{v} - \inf_{v \in \mathcal{C}} \int_{\Omega\times\mathbb{R}} \tilde{\phi} \cdot \mathrm{d}Dv + \iota_{\mathcal{C}}(\tilde{v}) + \iota_{\mathcal{K}_1^\alpha}(\tilde{\phi}).$$

For $\hat{\phi}_\alpha$ as constructed above and $u_1, u_2^{\hat{\alpha}}, u_2^\alpha$, we have

$$\begin{aligned}
\Delta_\alpha(1_{u_1}, \hat{\phi}_\alpha) &= 0,\\
\Delta_\alpha(1_{u_2}^{\hat{\alpha}}, \hat{\phi}_\alpha) &\leq 1 - 2^{1-\alpha} + \frac{2^{2-\alpha} - 2^{\hat{\alpha}}}{4\sqrt{1-2^{2\hat{\alpha}-2}}},\\
\Delta_\alpha(1_{u_2}^{\alpha}, \hat{\phi}_\alpha) &\leq 1 - 2^{1-\alpha} + \frac{2^{2-\alpha} - 2^{\alpha}}{4\sqrt{1-2^{2\alpha-2}}}
\end{aligned}$$

for $\alpha = \hat{\alpha} + \delta$ for some $\delta > 0$.

Proof. The proof requires the evaluation of the first two parts of Δ_α for the given variables $1_{u_1}, 1_{u_2}^{\hat{\alpha}}, 1_{u_2}^\alpha$ and $\hat{\phi}_\alpha$ (the third and fourth part vanish due to $1_{u_1}, 1_{u_2}^{\hat{\alpha}}, 1_{u_2}^\alpha \in \mathcal{C}$ and $\hat{\phi}_\alpha \in \mathcal{K}_1^\alpha$). Let us start with the infimum, for which we have

$$\inf_{v\in\mathcal{C}} \int_{\Omega\times\mathbb{R}} \hat{\phi}_\alpha \cdot \mathrm{d}Dv = \inf_{v\in\mathcal{C}} \int_S \hat{\phi}_\alpha \cdot \nu \;\mathrm{d}\mathcal{H}^2 - \underbrace{\int_{\Omega\times\mathbb{R}} v \;\mathrm{div}\hat{\phi}_\alpha \;\mathrm{d}x\mathrm{d}s}_{=0}$$

with $S = \{(x,s) \in \partial\Omega \times \mathbb{R} \ : \ 1_{u(\mu_+,\mu_-)} = 1, \ \hat{\phi}_\alpha \neq 0\}$, where the second part vanishes due to $\hat{\phi}_\alpha$ being divergence-free by construction. For the remaining part, we have

$$\begin{aligned}
\inf_{v\in\mathcal{C}} \int_S \hat{\phi}_\alpha \cdot \nu \ \mathrm{d}\mathcal{H}^2 &= 2\left(\int_0^{\frac{1}{2}}\int_{S_1} \hat{\phi}_\alpha^x \cdot n \ \mathrm{d}\mathcal{H}^1 \mathrm{d}s + \int_{\frac{1}{2}}^1 \int_{S_2} \hat{\phi}_\alpha^x \cdot n \ \mathrm{d}\mathcal{H}^1 \mathrm{d}s\right)\\
&= \int_{S_1} \varphi_2 \cdot n \ \mathrm{d}\mathcal{H}^1 + \int_{S_2} \varphi_1 \cdot n \ \mathrm{d}\mathcal{H}^1\\
&= \int_0^{\frac{3}{4}} \varphi_2(x_1,0)\cdot(0,-1)^T \ \mathrm{d}x_1 + \int_0^{\frac{1}{2}} \varphi_2(0,x_2)\cdot(-1,0)^T \ \mathrm{d}x_2\\
&\quad + \int_0^{\frac{1}{4}} \varphi_1(x_1,0)\cdot(0,-1)^T \ \mathrm{d}x_1 + \int_0^{\frac{1}{2}} \varphi_1(0,x_2)\cdot(-1,0)^T \ \mathrm{d}x_2\\
&= \int_{\frac{1}{4}}^1 \varphi_{12}(x_1,0) \ \mathrm{d}x_1 - \int_0^{\frac{1}{2}} \varphi_{11}(1,x_2) \ \mathrm{d}x_2\\
&\quad - \int_0^{\frac{1}{4}} \varphi_{12}(x_1,0) \ \mathrm{d}x_1 - \int_0^{\frac{1}{2}} \varphi_{11}(0,x_2) \ \mathrm{d}x_2, \qquad \text{(A.12)}
\end{aligned}$$

where $S = (S_1 \times [0,\frac{1}{2}]) \cup (S_2 \times (\frac{1}{2},1])$ and

$$\begin{aligned}
S_1 &= \{x \in \partial\Omega \ : \ x_1 \leq \tfrac{3}{4}, \ x_2 \leq \tfrac{1}{2}\},\\
S_2 &= \{x \in \partial\Omega \ : \ x_1 \leq \tfrac{1}{4}, \ x_2 \leq \tfrac{1}{2}\}.
\end{aligned}$$

In the last equation in (A.12) we inserted $\varphi_2(x_1,x_2) = \left(\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right)\varphi_1(1-x_1,x_2)$ and denote by $\varphi_1 = (\varphi_{11},\varphi_{12})^T$ the two components of φ_1. Further, with $\varphi_1 = Dv^\perp$, we obtain $\varphi_{11} = -\frac{\partial v}{\partial x_2}$ and $\varphi_{12} = \frac{\partial v}{\partial x_1}$, consequently with the definition of v as above

$$\inf_{v\in\mathcal{C}} \int_{\Omega\times\mathbb{R}} \hat{\phi}_\alpha \cdot \ \mathrm{d}Dv = \beta\left(b(1,\tfrac{1}{2}) + b(0,\tfrac{1}{2})\right) = \beta = \left(\tfrac{1}{2}\right)^{\alpha-1}. \qquad \text{(A.13)}$$

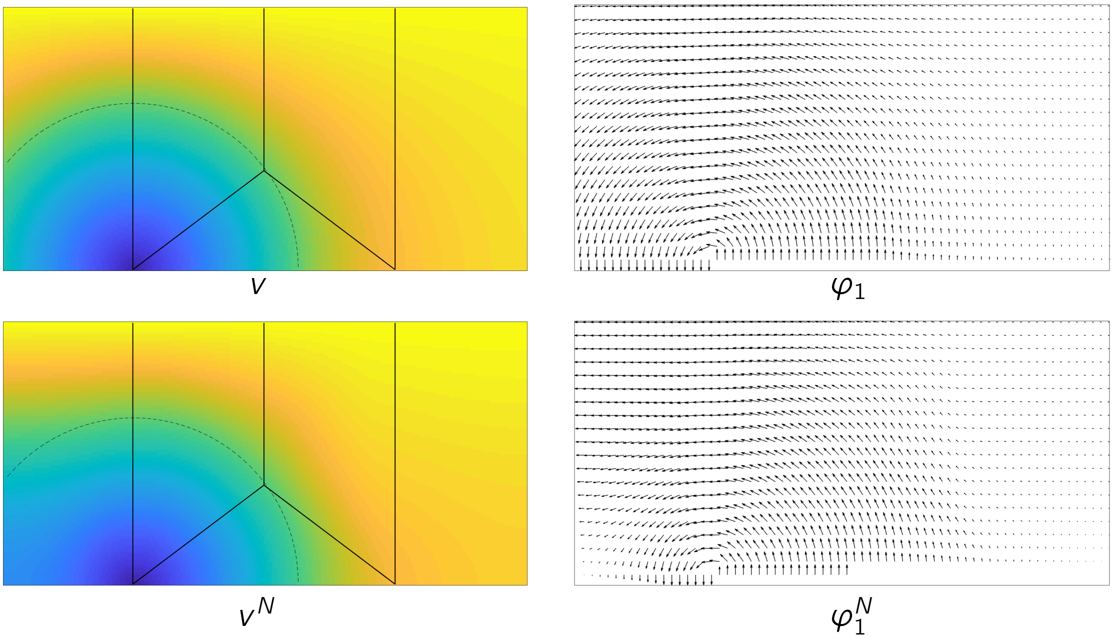
For the first part of Δ_α, we have

$$\sup_{\phi\in\mathcal{K}_1^\alpha} \int_{\Omega\times\mathbb{R}} \phi\cdot \mathrm{d}D1_{u_1} \leq \tilde{\mathcal{M}}^\alpha(u_1) = \left(\tfrac{1}{2}\right)^{\alpha-1}, \qquad \text{(A.14)}$$

$$\sup_{\phi\in\mathcal{K}_1^\alpha} \int_{\Omega\times\mathbb{R}} \phi\cdot \mathrm{d}D1_{u_2}^{\hat{\alpha}} \leq \tilde{\mathcal{M}}^\alpha(u_2^{\hat{\alpha}}) = 1 + \frac{2^{2-\alpha}-2^{\hat{\alpha}}}{4\sqrt{1-2^{2\hat{\alpha}-2}}}, \qquad \text{(A.15)}$$

$$\sup_{\phi\in\mathcal{K}_1^\alpha} \int_{\Omega\times\mathbb{R}} \phi\cdot \mathrm{d}D1_{u_2}^{\alpha} \leq \tilde{\mathcal{M}}^\alpha(u_2^{\alpha}) = 1 + \frac{2^{2-\alpha}-2^{\alpha}}{4\sqrt{1-2^{2\alpha-2}}}. \qquad \text{(A.16)}$$

For the primal-dual gap for $u_2^{\hat{\alpha}}$ and u_2^α, we have used that $u_2^{\hat{\alpha}}$ (u_2^α respectively) was chosen to be the optimal topology for $\hat{\alpha}$ (α respectively), thus the optimal branching point is $\hat{x}^{\hat{\alpha}}$



v φ_1

v^N φ_1^N

Figure A.2.: Comparison between the constructed image v and corresponding vector field φ_1 (upper row) and a numerically computed counterpart v^N and φ_1^N (lower row). The constructed v fits quite well to the numerical solution within the left circle and the upper left side of the image domain, whereas in the right part, one can see the differences leading to a violation of the constraints in $\mathcal{K}_1^{\hat{\alpha}}$. To obtain the numerical image v^N, we solved the branched transport problem numerically and extracted the image from the computed vector field via solving the Poisson equation.

($\hat{x}^\alpha$ respectively). Subtracting (A.13) from (A.14), (A.15) and (A.16), we finally obtain

$$\begin{aligned}
\Delta_\alpha(1_{u_1}, \hat{\phi}_\alpha) &\leq \left(\tfrac{1}{2}\right)^{\alpha-1} - \left(\tfrac{1}{2}\right)^{\alpha-1} = 0,\\
\Delta_\alpha(1_{u_2}^{\hat{\alpha}}, \hat{\phi}_\alpha) &\leq 1 + \frac{2^{2-\alpha} - 2^{\hat{\alpha}}}{4\sqrt{1 - 2^{2\hat{\alpha}-2}}} - 2^{1-\alpha},\\
\Delta_\alpha(1_{u_2}^{\alpha}, \hat{\phi}_\alpha) &\leq 1 + \frac{2^{2-\alpha} - 2^{\alpha}}{4\sqrt{1 - 2^{2\alpha-2}}} - 2^{1-\alpha}.
\end{aligned}$$

□

Remark A.0.5. With the numerically computed values of $\hat{\alpha}$ and α from Remark A.0.3, we obtain

$$\Delta_\alpha(1_{u_2}^{\hat{\alpha}}, \hat{\phi}_\alpha) \leq C_1 \approx 0.043054, \quad \Delta_\alpha(1_{u_2}^{\alpha}, \hat{\phi}_\alpha) \leq C_2 \approx 0.041494.$$

Bibliography

[1] G. Alberti, G. Bouchitté, and G. Dal Maso. The calibration method for the Mumford–Shah functional. *Comptes Rendus de l'Académie des Sciences - Série 1,* 329(3):249–254, 1999. 15

[2] G. Alberti, G. Bouchitté, and G. Dal Maso. The calibration method for the Mumford–Shah functional and free-discontinuity problems. *Calculus of Variations and partial differential equations,* 16(3):299–333, 2003. 12, 15, 16, 17, 63

[3] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems.* Clarendon Press, Oxford, 2000. 7, 8, 9, 12, 14

[4] L. Ambrosio and V. M. Tortorelli. On the approximation of free discontinuity problems. *Bollettino dell'Unione Matematica Italiana B,* 6(7):105–123, 1992. 2, 14

[5] MOSEK ApS. *The MOSEK Fusion API for C++ manual. Version 9.0.88*, 2019. https://docs.mosek.com/9.0/cxxfusion/index.html. 96

[6] A. Bach, A. Braides, and C. I. Zeppieri. Quantitative analysis of finite-difference approximations of free-discontinuity problems. arXiv:1807.05346 [math.AP]. 14, 53, 109, 115

[7] A. Bach, M. Cicalese, and M. Ruf. Random finite-difference discretizations of the Ambrosio–Tortorelli functional with optimal mesh size. arXiv:1902.08437 [math.AP]. 115

[8] G. Bellettini and A. Coscia. Discrete approximation of a free discontinuity problem. *Numerical Functional Analysis and Optimization,* 15(3-4):201–224, 1994. 14, 116

[9] L. Bergamaschi, E. Facca, A. Martinez, and M. Putti. Spectral preconditioners for the efficient numerical solution of a continuous branched transport model. *Journal of Computational and Applied Mathematics,* 354:259–270, 2019. 52

[10] B. Berkels, A. Effland, and M. Rumpf. A posteriori error control for the binary Mumford–Shah model. *Mathematics of Computation,* 86:1769–1791, 2017. 86

[11] M. Bernot, V. Caselles, and J.-M. Morel. Traffic plans. *Publicacions Matemàtiques,* 49(2):417–451, 2005. 38

[12] M. Bernot, V. Caselles, and J.-M. Morel. The structure of branched transportation networks. *Calculus of Variations and Partial Differential Equations,* 32(3):279–317, 2008. 38, 39

[13] M. Bonafini, G. Orlandi, and É. Oudet. Variational approximation of functionals defined on 1-dimensional connected sets: the planar case. *arXiv:1610.03839 [math.OC]*, 2016. 50, 92

[14] M. Bonnivard, A. Lemenant, and F. Santambrogio. Approximation of length minimization problems among compact connected sets. *SIAM Journal on Mathematical Analysis,* 47(2):1489–1529, 2015. 50

[15] J. P. Boyle and R. L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. *Advances in Order Restricted Statistical Inference,* 37:28–47, 1986. 73, 84

[16] A. Braides. *Γ-convergence for beginners.* Volume 22 of *Oxford Lecture Series in Mathematics and its Applications.* Oxford University Press, Oxford, 2002. 10, 11

[17] A. Braides. *Handbook of Differential Equations: Stationary partial differential equations*, chapter *A handbook of* Γ*-convergence*, pages 101–213. Elsevier, 2006. 13

[18] A. Brancolini and G. Buttazzo. Optimal networks for mass transportation problems. *ESAIM Control, Optimisation and Calculus of Variations,* 11(1):88–101, 2005. 33, 39, 40

[19] A. Brancolini, C. Rossmanith, and B. Wirth. Optimal micropatterns in 2D transport networks and their relation to image inpainting. *Archive for Rational Mechanics and Analysis,* 228(1):279–308, 2017. 2, 47, 55, 56, 57, 58, 60, 63

[20] A. Brancolini and B. Wirth. Equivalent formulations for the branched transport and urban planning problems. *Journal de Mathématiques Pures et Appliquées,* 106(4):695–724, 2016. 33, 38, 39, 41, 43, 44, 45

[21] A. Brancolini and B. Wirth. General transport problems with branched minimizers as functionals of 1-currents with prescribed boundary. *Calculus of Variations and Partial Differential Equations,* 57(3):82, 2018. 33, 46

[22] G. Buttazzo, A. Pratelli, S. Solimini, and E. Stepanov. *Optimal urban networks via mass transportation.* Volume 1961 of *Lecture Notes in Mathematics.* Springer, Berlin, 2009. 41

[23] A. Chambolle. Convex representation for lower semicontinuous envelopes of funtionals in L^1. *Journal of Convex Analysis,* 8(1):149–170, 2001. 63, 65

[24] A. Chambolle, B. Merlet, and L. Ferrari. A simple phase-field approximation of the Steiner problem in dimension two. *Advances in Calculus of Variations,* 12(2), 2016. 50, 51, 103, 104, 105, 108, 110, 116, 117

[25] A. Chambolle and T. Pock. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. *2011 International Conference on Computer Vision, Barcelona:* 1762–1769, 2011. 95

[26] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision,* 40(1):120–145, 2011. 71, 74, 79, 84, 100

[27] T. F. Chan and J. Shen. Variational image inpainting. *Communications on pure and applied mathematics,* 58(5):579–619, 2005. 61

[28] L. C. Evans and W. Gangbo. Differential equations methods for the Monge–Kantorovich mass transfer problem. *Memoirs of the American Mathematical Society* 137, 1999. 52

[29] E. Facca, F. Cardin, and M. Putti. Towards a stationary Monge–Kantorovich dynamics: The Physarum Polycephalum experience. arXiv:1610.06325v1 [math.NA]. 52

[30] E. Facca, S. Daneri, F. Cardin, and M. Putti. Numerical solution of Monge–Kantorovich equations via a dynamic formulation. arXiv:1709.06765 [math.NA]. 52

[31] M. Fampa, J. Lee, and N. Maculan. An overview of exact algorithms for the Euclidean Steiner tree problem in n-space. *International Transactions in operational research,* 23:861–874, 2016. 48

[32] H. Federer. *Geometric measure theory.* Volume 153 of *Grundlehren der mathematischen Wissenschaften.* Springer, 1969. 128

[33] L. A. D. Ferrari, C. Rossmanith, and B. Wirth. Phase field approximations of branched transportation problems. arXiv:1805.11399 [math.OC]. 2, 47, 104, 105, 106, 108, 113

[34] R. Fonseca, M. Brazil, P. Winter, and M. Zachariasen. Faster exact algorithm for computing Steiner trees in higher dimensional Euclidean spaces. *Presented at the 11th DIMACS Implementation Challenge Workshop, Providence, RI*, 2014. http://dimacs11.cs.princeton.edu/workshop/FonsecaBrazilWinterZachariasen.pdf. 48

[35] E. Gilbert and H. Pollak. Steiner minimal trees. *SIAM Journal of Applied Mathematics,* 16(1):1–29, 1968. 48

[36] E. De Giorgi, M. Carriero, and A. Leaci. Existence theorem for a minimum problem with free discontinuity. *Archive for Rational Mechanics and Analysis,* 108(3):195–218, 1989. 12

[37] E. De Giorgi and T. Franzoni. Su un tipo di convergenza variazionale. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Serie* 8 (58):842–850, 1975. 10

[38] D. Juhl, D. M. Warme, P. Winter, and M. Zachariasen. The GeoSteiner software package for computing Steiner trees in the plane: an updated computational study. *Mathematical Programming Computation,* 10(4):487–532, 2018. 48

[39] L. Kantorovich. On the transfer of masses. *Doklady Akademii Nauk USSR,* 37:7–8, 1942. 32

[40] N. Maculan, P. Michelon, and A. Xavier. The Euclidean Steiner tree problem in $\mathbb{R}^n$: A mathematical programming formulation. *Annals of Operations Research,* 96(1):209–220, 2000. 48

[41] F. Maddalena and S. Solimini. Transport distances and irrigation models. *Journal of Convex Analysis,* 16(1):121–152, 2009. 39

[42] F. Maddalena, S. Solimini, and J.-M. Morel. A variational model of irrigation patterns. *Interfaces and free boundaries,* 5(4):391–415, 2003. 33, 38

[43] G. Dal Maso. *An introduction to* Γ*-convergence.* Birkhäuser, Basel, 1993. 10

[44] M. Matuszak, J. Miekisz, and T. Schreiber. Solving ramified optimal transport problems in the Bayesian influence diagram framework. In *International conference on artificial intelligence and soft computing, Lecture Notes in Computer Science:* 582–590, 2012. 48

[45] J. M. L. Maubach. Local bisection refinement for n-simplicial grids generated by reflection. *SIAM Journal on Scientific Computing,* 16(1):210–227, 1995. 19

[46] Z. A. Melzak. On the problem of Steiner. *Canadian Mathematical Bulletin,* 4(2):143–148, 1961. 48

[47] L. Modica and S. Mortola. Un esempio di Γ-convergenza. *Bollettino della Unione Matematica Italiana B,* 14(1):285–299, 1977. 2, 13, 50

[48] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris:* 666–704, 1781. 32

[49] A. Monteil. Uniform estimates for a Modica–Mortola type approximation of branched transportation. *ESAIM Control Optimisation and Calculus of Variations,* 23(1):309–335, 2017. 48, 49, 50

[50] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics,* 17:577–685, 1989. 2, 11

[51] E. Oudet and F. Santambrogio. A Modica–Mortola approximation for branched transport and applications. *Archive for Rational Mechanics and Analysis,* 201(1):115–142, 2011. 48, 49, 50, 52, 103

[52] P. Pegon, F. Santambrogio, and Q. Xia. A fractal shape optimization problem in branched transport. *arXiv:1709.01415 [math.OC]*, 2017. 50

[53] J. Piersa. Ramification algorithm for transporting routes in $\mathbb{R}^2$. *2014 IEEE 26th International Conference on Tools with Artificial Intelligence, Limassol:* 657–664, 2014. 48

[54] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the Mumford–Shah functional. *2009 IEEE 12th International Conference on Computer Vision:* 1133–1140, 2009. 18, 63, 71

[55] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global solutions of variational models with convex regularization. *SIAM Journal on Imaging Sciences,* 3(4):1122–1145, 2010. 18

[56] H. J. Prömel and A. Steger. *The Steiner tree problem. A tour through graphs, algorithms, and complexity. Advanced Lectures in Mathematics.* Vieweg, 2002. 33

[57] QuocMesh. *Using and Programming the QuocMesh Library. QuocMesh Collective. Version 1.5*, 2014. https://archive.ins.uni-bonn.de/numod.ins.uni-bonn.de/software/quocmesh/1.5/doc/lib/index.html. 89

[58] J. Rasch. *Advanced convex analysis for improved variational image reconstruction.* PhD thesis, Westfälische Wilhelms-Universität Münster, 2018. 74, 75

[59] S. Repin. A posteriori error estimation for variational problems with uniformly convex functionals. *Mathematics of Computation,* 69(230):481–500, 1999. 86

[60] F. Santambrogio. *Optimal transport for applied mathematicians.* Birkhäuser, Basel, 2015. 32, 37

[61] W. D. Smith. How to find Steiner minimal trees in Euclidean d-space. *Algorithmica,* 7(1-6):137–177, 1992. 48

[62] R. Stevenson. The completion of locally refined simplicial partitions created by bisection. *Mathematics of Computation,* 77(261):227–241, 2007. 19

[63] C. T. Traxler. An algorithm for adaptive mesh refinement in n dimensions. *Computing,* 59:115–137, 1997. 19, 20, 22

[64] R. Verfürth. A posteriori error estimation and adaptive mesh-refinement techniques. *Journal of Computational and Applied Mathematics,* 50:67–83, 1994. 19, 86

[65] C. Villani. *Topics in optimal transportation.* Volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, 2003. 32

[66] C. Villani. *Optimal transport, old and new.* Volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2008. 32, 52

[67] B. Wirth. Phase field models for two-dimensional branched transportation problems. arXiv:1805.05141 [math.OC]. 104

[68] Q. Xia. Optimal paths related to transport problems. *Communications in contemporary mathematics,* 5(2):251–279, 2003. 33, 34, 37, 48

[69] Q. Xia. Numerical simulation of optimal transport paths. In *2010 Second International Conference on Computer Modeling and Simulation*, 2008. 48, 52

[70] Q. Xia. Motivations, ideas and applications of ramified optimal transportation. *ESAIM: Mathematical Modelling and Numerical Analysis,* 49(6):1791–1832, 2015. 48

[71] G. Xue, T. P. Lillys, and D. E. Dougherty. Computing the minimum cost pipe network interconnecting one sink and many sources. *SIAM Journal on Optimization,* 10(1):22–42, 1999. 52