Adaptive Numerical Methods for Optimal and Branched Transport Problems

Dissertation
for the award of the degree
"Doctor rerum naturalium" (Dr.rer.nat.)
of the Georg-August-Universität Göttingen

within the doctoral program "Programme for Computer Science"
of the Georg-August University School of Science (GAUSS)

submitted by
Olga Minevich
from St.Petersburg, Russia

Göttingen, 2023

Thesis Committee

Prof. Dr. Bernhard Schmitzer
Research Group Optimal Transport, Institute of Computer Science,
Georg-August-Universität Göttingen

Prof. Dr. Alexander Ecker
Research Group Neural Data Science, Institute of Computer Science,
Georg-August-Universität Göttingen

Prof. Dr. Benedikt Wirth
Working Group Mathematical Optimization, Institute for Analysis and Numerics,
Westfälische Wilhelms-Universität Münster

Members of the Examination Board

Reviewer
Prof. Dr. Bernhard Schmitzer
Research Group Optimal Transport, Institute of Computer Science,
Georg-August-Universität Göttingen

Second Reviewer
Prof. Dr. Benedikt Wirth
Working Group Mathematical Optimization, Institute for Analysis and Numerics,
Westfälische Wilhelms-Universität Münster

Further members of the Examination Board

Prof. Dr. Alexander Ecker
Research Group Neural Data Science, Institute of Computer Science,
Georg-August-Universität Göttingen

Prof. Dr. Florin Manea
Research Group Theoretical Computer Science, Institute of Computer Science,
Georg-August-Universität Göttingen

Prof. Dr. Constantin Pape
Research Group Computational Cell Analytics, Institute of Computer Science,
Georg-August-Universität Göttingen

Prof. Dr. Fabian Sinz
Research Group Machine Learning, Institute of Computer Science,
Georg-August-Universität Göttingen

Date of the oral examination: 15.11.2023

Abstract

Optimal transport is an area of mathematical research that has been gaining popularity in recent years in various application fields such as economics, statistics or machine learning. Two important factors behind its increasing popularity are its modelling flexibility and the ever-expanding range of available dedicated computational tools.

Unbalanced optimal transport is a generalization that allows for the comparison of measures with different mass, which is more appropriate in some applications. In this thesis, we consider the barycenter problem (i.e. finding a weighted average) between several input measures with respect to the unbalanced Hellinger–Kantorovich metric. In particular, we focus on the case with an uncountable number of Dirac input measures. We study existence, uniqueness and stability of the solutions, and demonstrate the intricate behavior of the barycenters with respect to the length scale parameter using analytical and numerical tools.

Another important variant is *branched* transport, where the transport cost encourages the formation of branched transportation networks. We focus in this thesis on its convex relaxation in terms of multimaterial transport. In particular, we study the multimaterial problem in a setting when only a single topology of the solution is admissible and describe the simple structure of the dual solution in this case. We then formulate a problem with 3 sources and 1 sink where two candidate solutions of different topologies give the same transportation cost, study its properties and characterize the solution set.

Acknowledgements

I would like to thank my supervisor, Prof. Bernhard Schmitzer, for his guidance and feedback, and generally for the invaluable experience I gained in the course of my PhD studies.

I would also like to express my gratitude to the members of my examination board for their valuable time.

Special thanks of course to my colleagues, especially, although not exclusively, those who shared with me a lot of espressos during the early 11 o'clock meetings.

I would like to thank my main collaborators in the past few years, Dr. Mauro Bonafini, Dr. Julius Lohmann, (again) Prof. Bernhard Schmitzer and Prof. Benedikt Wirth for their expertise and input.

I am also thankful to DFG SPP 1962 for the opportunity to perform, present and discuss my research within the framework of the Priority Program.

My deepest gratitude without any unnecessary explanations goes to my family.

"I move uphill without a stop,
I waste no time.
And in the world there is no top
You cannot climb."

— *Vladimir Vysotsky*

Contents

1 Introduction

Optimal transport is a scientific domain which studies the problem of finding the most cost-efficient way to transport one input distribution to another. As many processes in various spheres of life and science are controlled by a *shortest path* principle or a *minimum energy* principle, optimal transport finds applications in different areas. While historically optimal transport was developed for engineering [97], and logistics and economics [116, 75, 79], nowadays it is also efficiently used in computer graphics [94], computer vision [114], computational biology [109], medical imaging [70], statistics [78], machine learning [47], fluid mechanics [6] and many other fields.

Optimal transport offers a natural way to define a distance in the space of distributions, allowing to compare and interpolate between input distributions (see [4, Lecture 8] on metric properties of optimal transport and Wasserstein distance). With its broad range of generalizations and adaptations, optimal transport offers a powerful analytical tool (e.g. in the studies of partial differential equations [60]) as well as a rich variety of computational methods [104]. However, it should be noted that optimal transport problems (especially in higher dimensions) are computationally expensive (as compared to the alternative measure comparison options, such as L^p or Mahalanobis distance), which leads to a constant interest in the development and adaptation of computationally efficient numerical methods [49, 67].

A major drawback of the standard optimal transport is that it only works for measures of the same total mass, which is not suitable in many applications, for example when the initial data is subject to substantial noise and renormalization can lead to undesired effects. One of the settings proposed to lift his constraint is *unbalanced transport*, which generalizes the optimal transport problem to input measures with unequal total mass. Unbalanced optimal transport has been studied for example in [10, 40, 42, 86] and found its applications in many fields, such as machine learning [8], medical imaging [62] or bioinformatics [73]. We focus in this work in particular on the Hellinger–Kantorovich distance (see [86] for an in-depth study and [36] for a compact summary of some important properties and comparison with the Wasserstein distance).

An interesting problem arising in different application areas is the *barycenter problem* – the problem of finding a (weighted) average of several input distributions with respect to a chosen metric. The barycenter problem in balanced optimal transport setting (in particular, in Wasserstein-2 distance) was proposed in [1] and studied in more detail in [103, 50, 21]. The barycenter problem has also been extended to the unbalanced setting, in particular, for the Hellinger–Kantorovich distance the barycenter problem has been studied in [65, 43]. One of the settings proposed in [65] is the barycenter problem between several Dirac measures: While the Wasserstein barycenter of any number of input Dirac measures is known to be a single Dirac measure [1], the Hellinger–Kantorovich barycenter exhibited non-trivial structure depending on the length scale parameter, even including some diffused solutions, which presents an interesting challenge for further research.

While standard optimal transport does not take the interaction between masses into consideration, in many applications combined transfer of resources is more advantageous than the individual one, leading to the so-called *branching behavior*. Generally, a lot of natural and man-made objects appear as branching structures, such as for example vascular systems in plants [95] and animals [123], tree branches [105], lightning strikes [74] or computer networks [118], gas pipelines [20], and electric power lines [106].

There are different optimal transport models incorporating the described property: In branched transport [122, 90, 32], the optimal transport cost is selected to be concave, to encourage the masses to travel in bulk. In urban planning [31], the cost of transportation is defined separately, for transport on a network and (a generally higher cost) for transport outside of the network, thus bringing the network ramification. In multimaterial transport [92], transport of mass of different types is performed, and the cost of transporting different materials together is reduced when compared to separate transport. These three problems, although seemingly describing the optimal transport from different perspectives, turn out to have a lot in common, namely, under some assumptions, branched transport and urban planning problems can be shown to be equivalent [88, 89], and the multimaterial transport problem can be shown to be their convex relaxation [92, 87]. Multimaterial transport has also been gaining some popularity on its own in recent years, including for example the mailing problem [37] or the power line communication technology modelling [91].

Multimaterial transport problem (in the general setting) is an infinite dimensional convex optimization problem. Therefore, the numerical schemes have to rely on discretization [22, 93], which in the case of the multimaterial problem has to be considered with care, as the chosen mesh has to be a good representation of the edges of the optimal network. The development of efficient numerical methods is therefore only possible if the behavior of solutions of the multimaterial transport problem is examined more closely, at least in simplified settings.

The thesis is organised as follows.

Chapter 2: Background provides some background for this study, and is used to lay out the notation, recall some definitions from different areas of mathematics, in particular convex analysis, optimal transport theory, and numerical mathematics, and show the context of the research presented in this thesis.

Chapter 3: Hellinger–Kantorovich Barycenter discusses the properties of barycenters in the Hellinger–Kantorovich metric. Motivated by the non-trivial behavior discussed in previous studies of the problem, we investigate the barycenters in the particular cases of a continuum of general measures and of Dirac measures using both analytical and numerical tools. The results presented in this chapter have been published in [23].

Chapter 4: Branched and Multimaterial Transport focuses on the branched and multimaterial transport problems. As these problems are known to be difficult to solve, we focus on some specific settings of multimaterial transport, namely, we consider a case when only one network configuration is possible from the givens, and a low-dimensional case when two specific configurations are allowed simultaneously. We also present two numerical schemes for the multimaterial problem and perform some experiments to support our findings and gain more insights into the intricate behavior of the solutions. The results, presented in this chapter, are currently being prepared for publication as [24].

2 Background

This chapter presents the necessary background for the thesis. Here we introduce notation, recall definitions and provide important theoretical results, which will be used in further chapters.

The first section focuses on optimal transport in balanced and unbalanced settings and provides a brief discussion of displacement interpolation. The next section is devoted to the elements of convex analysis and convex duality theory. The last section of this chapter gives a small overview of common numerical approaches for optimal transport.

Throughout the thesis, for a compact metric space (X, d_X) we denote by $\mathrm{C}(X)$ the space of continuous real valued functions equipped with the sup-norm and by $\mathrm{C}^1(X)$ the set of continuously differentiable functions whenever applicable.

We denote by $\mathcal{M}(X)$ the space of Radon measures equipped with the total variation norm, the subsets of non-negative and probability measures are denoted by $\mathcal{M}_+(X)$ and $\mathcal{P}(X)$ respectively. We consider on $\mathcal{M}(X)$ the weak* topology induced via duality with $\mathrm{C}(X)$. For $\mu \in \mathcal{M}_+(X)$ one has $\|\mu\| = \mu(X)$.

2.1 Optimal transport

2.1.1 Balanced optimal transport

Monge formulation

The Monge formulation of optimal transport problem relies on the notion of push-forward measures:

Definition 2.1.1. Let X, Y be compact metric spaces, $\mu \in \mathcal{M}(X)$ a signed Radon measure, $T : X \to Y$ a measurable function. The measure $T_{\#}\mu \in \mathcal{M}(Y)$ is called a *push-forward* of μ along T if

$$\int_Y \sigma(y)\mathrm{d}(T_{\#}\mu)(y) = \int_X \sigma(T(x))\mathrm{d}\mu(x) \quad \forall \sigma \in \mathrm{C}(Y).$$

In modern terms, the formulation of the optimal transport problem first proposed by Gaspard Monge [97] is as follows: Given two probability measures $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ and a Borel cost function $c : X \times Y \to [0, \infty]$, find a *transport map* T such that

$$\inf_T \left\{ \int_X c(x, T(x))\mathrm{d}\mu(x) \;\middle|\; T : X \to Y \,, T_{\#}\mu = \nu \right\}. \tag{2.1.1}$$

Because of its nonlinearity, problem (2.1.1) and its properties have remained unstudied for a long time. In 1885 the French Academy of Science announced a prize for the research of properties of the original problem given by Monge (with the cost function $c(x, y) = |x - y|$) under the assumption that the solutions exist. The prize was awarded to Paul Appell (see [5] and [108]

for historical context). The questions of existence of solutions were only later formulated by Alexander Vershik in 1970 [119], and then studied further in the works of Vladimir Sudakov [115], Luigi Ambrosio [3], Lawrence Evance and Wilfried Gangbo [60], Neil Trudinger and Xu-Jia Wang [117], Luis Caffarelli, Mikhail Fekdman and Robert McCann [35] and others. Some of the results were obtained using the relaxed formulation of the optimal transport problem, given by Kantorovich, which we recall next.

Kantorovich formulation

More than 150 years later, the soviet mathematician Leonid Kantorovich introduced a different formulation of the optimal transport problem [75]. This formulation, usually referred to by his name, relies on the notion of *transport plans* or *couplings*:

Definition 2.1.2. Given two probability measures $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$, the *set of transport plans* between the two is defined as

$$\Gamma(\mu,\nu) = \{\gamma \in \mathcal{P}(X \times Y) \;\mid\; \mathrm{P}_1\gamma = \mu,\ \mathrm{P}_2\gamma = \nu\}\,, \tag{2.1.2}$$

where operators P_1 and P_2 map their operands to first and second marginal respectively, that is $\mathrm{P}_1\gamma = [(x,y) \mapsto x]_{\#}\gamma$ and $\mathrm{P}_2\gamma = [(x,y) \mapsto y]_{\#}\gamma$.

Kantorovich's formulation of optimal transport is to find a transport plan that solves the following problem:

$$\inf\left\{\int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y) \;\middle|\; \gamma \in \Gamma(\mu,\nu)\right\}. \tag{2.1.3}$$

One can see that this formulation of optimal transport problem is in fact a (potentially infinite dimensional) linear program, i.e. an optimization problem with linear objective function under linear constraints, which makes it far easier to study its properties. It should also be noted that when an optimal solution of the Monge problem T exists, an optimal solution of the Kantorovich problem can be constructed as $\gamma = (Id, T)_{\#}\mu$, where Id is the identity function. Therefore, the Kantorovich formulation is sometimes referred to as the *relaxation* of the Monge problem.

The transport problem was also considered (in finite-dimensional setting) independently by other authors, for instance in 1930 by Aleksei Tolstoy [116], in 1941 by Frank Hitchcock [72], and later in the 1940s by Tjalling Koopmans [79] and George Dantzig [51, 52]. After learning about each other's works, Koopmans and Kantorovich have communicated on the topic in the later 1950s [120], and in 1975 they were awarded the the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel for "their contributions to the theory of optimum allocation of resources" [99].

Wasserstein distance

An important property of optimal transport is that one can define a distance between probability measures:

Definition 2.1.3. Assume $X = Y$, and let $c(x,y) = d(x,y)^p$ for some $p \in [1,\infty)$, where d is a distance on X. Then the *p-Wasserstein distance* on X is

$$W_p(\mu,\nu) = \left\{\inf \int_{X,X} d(x,y)^p \mathrm{d}\gamma(x,y) \;\middle|\; \gamma \in \Gamma(\mu,\nu)\right\}^{\frac{1}{p}}. \tag{2.1.4}$$

The p-Wasserstein distance can be shown to satisfy the distance axioms (see e.g. [104, Proposition 2.3]). An important property of the Wasserstein distance is its relation to the weak convergence:

Theorem 2.1.4 (Basic properties of W_p [121, Theorems 6.9, 6.18])**.** *Let (X, d) be a compact metric space. The Wasserstein distance W metrizes the weak* topology over $\mathcal{P}(X)$. The metric space $(\mathcal{P}(X), W)$ is separable and complete.*

2.1.2 Unbalanced optimal transport

In the previous section, we were considering the transport between two probability measures $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$. However in practice the marginals $\mu \in \mathcal{M}_+(X), \nu \in \mathcal{M}_+(Y)$ sometimes have different total mass, meaning that their renormalization could lead to some undesired effects. Therefore, it has been proposed to use the *unbalanced* optimal transport – an unconstrained optimization problem, where the marginal deviation is penalized with some divergence function $\mathcal{D}$ (see [104, Section 10.2]):

$$\inf_{\gamma \in \mathcal{M}_+(X \times Y)} \int_{X \times Y} c(x, y) \mathrm{d}\gamma(x, y) + \varepsilon_1 \mathcal{D}(\mathrm{P}_1 \gamma | \mu) + \varepsilon_2 \mathcal{D}(\mathrm{P}_2 \gamma | \nu).$$

One of the common choices is the Kullback–Leibler divergence $\mathcal{D}(\cdot|\cdot) = \mathrm{KL}(\cdot|\cdot)$.

Definition 2.1.5. For $\mu \in \mathcal{M}(X)$, $\nu \in \mathcal{M}_+(X)$ the *Kullback–Leibler divergence* (or relative entropy) of μ w.r.t. ν is given by

$$\mathrm{KL}(\mu|\nu) = \begin{cases} \int_X \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu, & \text{if } \mu \ll \nu,\ \mu \geq 0, \\ +\infty, & \text{else,} \end{cases}$$

where $\varphi \colon \mathbb{R} \to \mathbb{R} \cup \infty$ is defined by

$$\varphi(s) = \begin{cases} s \log(s) - s + 1, & \text{if } s > 0, \\ 1, & \text{if } s = 0, \\ +\infty, & \text{else,} \end{cases}$$

$\mu \ll \nu$ means that μ is absolutely continuous with respect to ν, i.e. for every measurable subset $A \subset X$, $\nu(A) = 0$ implies $\mu(A) = 0$, and $\mathrm{d}\mu/\mathrm{d}\nu$ denotes the Radon–Nikodym derivative of μ w.r.t. ν.

In particular, with $\mathcal{D} = \mathrm{KL}$, for $c(x, y) = \|x - y\|^2$ the unbalanced problem leads to the squared Gaussian–Hellinger distance [86] and for

$$c(x, y) = \begin{cases} -2 \log \cos(|x - y|/\kappa), & \text{if } |x - y| < \kappa\pi/2, \\ +\infty, & \text{otherwise} \end{cases}$$

it gives the squared scaled Hellinger–Kantorovich distance (also known as Wasserstein–Fisher–Rao distance) [41, 86], with $\kappa > 0$ a length scale parameter (see a more detailed description in Chapter 3).

2.1.3 Displacement interpolation

Returning to the Monge formulation (2.1.1) in $\mathbb{R}^n$ and assuming an optimal map T with $T_{\#}\mu = \nu$ exists, let us select a parameter $t \in [0, 1]$ and consider the map

$$T_t = (1 - t)Id + tT.$$

The interpolation

$$\rho_t = (T_t)_{\#}\mu = \big((1 - t)Id + tT\big)_{\#}\mu$$

is called *displacement interpolation* or *McCann's interpolation* and can be interpreted as moving along a geodesic with respect to the Wasserstein distance W_p [66].

In the case $p = 2$, displacement interpolation can also be obtained as a solution of

$$\rho_t = \operatorname*{argmin}_{\rho} \big\{(1 - t)W_2(\rho, \mu)^2 + tW_2(\rho, \nu)^2\big\},$$

see for example [104, Section 7.1]. This perspective allows generalizing the problem to the barycenter problem. Let $(\mu_k)_k$ with $\mu_k \in \mathcal{P}(X)$ be the input probability measures, and let $(\lambda_k)_k$: $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$ be the corresponding weights. The Wasserstein barycenter $\rho \in \mathcal{P}(X)$ can then be defined as follows [1]:

$$\rho \in \operatorname*{argmin}_{\tau} \sum_k \lambda_k W_2(\tau, \mu_k)^2.$$

The barycenter problem has also been generalized to the unbalanced optimal transport setting, see for example [43, 65] and Chapter 5 of this work.

The notion of displacement interpolation is also closely related to the so-called *dynamic* (or *geodesic*) formulations of the optimal transport problem.

Jean-David Benamou and Yann Brenier in the seminal work [11] presented the following result, which is now often referred to as *Benamou–Brenier formula*. When $X = \mathbb{R}^n$ and cost $c(x, y) = |x - y|^2$, the squared Wasserstein-2 distance between $\mu, \nu \in \mathcal{P}(X)$ (see (2.1.4)) is equivalent to the problem of finding an optimal curve ρ_t and associated velocity field v_t minimizing the following functional

$$\inf \left\{ \int_0^1 \int_{R^n} |v_t(x)|^2 \mathrm{d}\rho_t(x)\mathrm{d}t \;\middle|\; (\rho_t, v_t) : \partial_t \rho_t + \operatorname{div}_x(v_t \rho_t) = 0, \rho_0 = \mu, \rho_1 = \nu \right\}, \tag{2.1.5}$$

i.e. the Wasserstein-2 distance can be expressed in terms of solutions of a continuity equation (here we only present the dynamic problem formally, see the original work [11] and [4, Chapter 9] for the details). The authors also proposed a momentum change of variables motivated by the fluid mechanics interpretation which allowed them to obtain a convex objective functional with affine constraints.

The Benamou–Brenier formula has later been generalized for other spaces and cost functions, including some results for the unbalanced transport (see for example [10, 63, 13, 42, 111]).

As the dynamic model of Benamou and Brenier is formulated in terms of curves and velocity fields, it belongs to the class of *Eulerian* models: Eulerian flow field models focus on space and characteristics which are associated with it, while the alternative Lagrangian flow field models focus on individual particles and their properties. There are also Lagrangian dynamic formulations of optimal transport, which are specifically common in the applications allowing interaction between mass particles (see for example [15, 33]).

2.2 Elements of convex analysis and duality theory

This section contains some definitions and known results from convex analysis and (convex) duality theory which will be used further in the thesis. Unless specified explicitly, the definitions are given following [9].

Definition 2.2.1. Let X a locally convex space and $S \subset X$ a convex set, let $f : S \to \mathbb{R}$. A functional in the dual space $v \in X^*$ is called the *subgradient* of f at $x_0 \in S$ if

$$f(x) - f(x_0) \geq \langle v, x - x_0 \rangle \quad \forall x \in S. \tag{2.2.1}$$

The set of all subgradients of function f at x_0 is called the *subdifferential* of f at x_0 and is denoted $\partial f(x_0)$.

The subdifferential is by definition a convex and closed set, although it might be empty.

We also introduce here the proximal operator as an important tool for analysis and numerical methods.

Definition 2.2.2. Let X be a Hilbert space and $f : X \to \mathbb{R} \cup \infty$ proper lower semicontinuous convex function. The *proximal operator* is defined as

$$\text{prox}_f(y) = \arg\min_{x \in X} \left\{ f(x) + \frac{1}{2} \|x - y\|_X^2 \right\}. \tag{2.2.2}$$

Proximal operator is closely related to the subdifferential:

$$p = \text{prox}_f(x) \Leftrightarrow x - p \in \partial f(p). \tag{2.2.3}$$

Definition 2.2.3. Let X be a normed vector space and $f : X \to \mathbb{R} \cup \infty$ be a function operating on that space. The *convex conjugate* or *Fenchel–Legendre conjugate* of function f is the function $f^* : X^* \to \mathbb{R} \cup \infty$, defined on the dual space X^* as

$$f^*(x^*) = \sup_{x \in X} \{\langle x, x^* \rangle - f(x)\} \quad \forall x^* \in X^*.$$

The following useful formula can be shown for the convex conjugate computations (see e.g. [28, Section 3.3]). For some $a, b, d, e \in \mathbb{R}$ and $c \in \mathbb{R}_{++}$ let $g(x) = a + bx + cf(dx + e)$. Then for the convex conjugate one finds

$$g^*(x^*) = -a - e\frac{x^* - b}{d} + cf^*\left(\frac{x^* - b}{cd}\right). \tag{2.2.4}$$

The following important result relates the convex conjugates and the subgradient.

Proposition 2.2.4 (Fenchel–Young inequality [107, Section 12])**.** *For any function $f : X \to \mathbb{R} \cup \infty$ and its convex conjugate $f^* : X^* \to \mathbb{R} \cup \infty$, for every $x \in X$ and $x^* \in X^*$ it holds that*

$$\langle x, x^* \rangle \leq f(x) + f^*(x^*).$$

The equality holds if and only if $x^* \in \partial f(x)$.

It can easily be seen from the definition that the convex conjugate of a function is always lower semicontinuous. In the view of that, the following result is important:

Theorem 2.2.5 (Fenchel–Moreau [4, Theorem 3.4]). *For a proper function f and its biconjugate $f^{**} := (f^*)^*$, it holds that $f = f^{**}$ if and only if f is a lower semicontinuous convex function.*

We next state an important result on duality for optimization problems, which is used often throughout the thesis.

Theorem 2.2.6 (Fenchel–Rockafellar [107, Theorem 31.1]). *Let X, Y be normed vector spaces, $G : X \to \mathbb{R} \cup \infty$ and $F : Y \to \mathbb{R} \cup \infty$. Let $A : X \to Y$ be a linear bounded transformation from X to Y.*

Let $p \in \mathbb{R} \cup \infty$ be the primal value

$$p = \inf_{x \in X} F(Ax) + G(x), \tag{2.2.5}$$

and $d \in \mathbb{R} \cup -\infty$ be the dual value

$$d = \sup_{y^* \in Y^*} -F^*(-y^*) - G^*(A^* y^*), \tag{2.2.6}$$

where Y^* is the dual space of Y, F^* and G^* are the convex conjugates of functions F and G, and A^* is the adjoint operator of A.

Then these values satisfy the weak duality, i.e. $p \geq d$.

If additionally functions F and G are proper convex, lower semicontinuous and either of the following conditions is satisfied

$$\exists \bar{x} \in X : (G(\bar{x}) < +\infty) \text{ and } (F(A\bar{x}) < +\infty) \text{ and } \left[(G \text{ is continuous at } \bar{x}) \text{ or } (F \text{ is continuous at } A\bar{x})\right], \tag{2.2.7}$$

$$\exists \bar{y}^* \in Y^* : (F^*(\bar{y}^*) < +\infty) \text{ and } (G^*(-A^*\bar{y}^*) < +\infty) \text{ and } \left[(F^* \text{ is continuous at } \bar{y}^*) \text{ or } (G^* \text{ is continuous at } -A^*\bar{y}^*)\right], \tag{2.2.8}$$

then the strong duality holds, i.e. $p = d$. When (2.2.7) is satisfied, the supremum of the dual problem is attained whenever it is finite. When (2.2.8) is satisfied, the infimum of the primal problem is attained whenever it is finite.

Next we state and briefly discuss the dual optimal transport problem, for which the following definition will be useful.

Definition 2.2.7. For a set $S \subset \mathbb{R}^n$ we denote by $\imath_S$ its *indicator function*:

$$\imath_S(s) = \begin{cases} 0, & \text{if } s \in S, \\ +\infty, & \text{else.} \end{cases}$$

Function $\imath_S$ is a convex if and only if S is a convex set.

We note that this function is sometimes called *characteristic function* of the set, however we are following the convention of [107].

Example 2.2.8 (Dual formulation of optimal transport)**.** The dual problem for the balanced optimal transport problem (2.1.3) is

$$\sup\left\{ \int_X \varphi(x)\mathrm{d}\mu(x) + \int_Y \psi(y)\mathrm{d}\nu(y) \;\middle|\; (\varphi,\psi) \in \mathrm{C}(X)\times\mathrm{C}(Y) : \varphi(x)+\psi(y) \le c(x,y) \;\; \forall x,y \right\}. \quad (2.2.9)$$

This problem is sometimes called *Kantorovich–Rubinstein formula*, and the dual variables φ and ψ are called *dual potentials* or *Kantorovich potentials*.

Let us briefly show the derivation using the Fenchel–Rockafellar theorem in the case when the cost function $c(x,y) = \|x-y\|^2$ (see [3, Lecture 3] for more general settings). We choose

$$\begin{aligned} G &: \mathcal{P}(X\times Y) \to \mathbb{R}\cup\infty, & G(\gamma) &= \begin{cases} \int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y), & \text{if } \gamma \ge 0, \\ +\infty, & \text{otherwise,} \end{cases} \\ F &: \mathcal{P}(X)\times\mathcal{P}(Y) \to \{0,\infty\}, & F(u,v) &= \imath_{\{u=\mu, v=\nu\}}, \\ A &: \mathcal{P}(X\times Y) \to \mathcal{P}(X)\times\mathcal{P}(Y), & A\gamma &= (\mathrm{P}_1\gamma, \mathrm{P}_2\gamma). \end{aligned}$$

Recalling here that the convex conjugate of a linear functional $f(x) = \langle a, x\rangle$ is $f^*(x^*) = \imath_{x^*=a}(x^*)$ and the conjugate of an indicator function of the form $f(x) = \imath_{x=a}(x)$ is $f^*(x*) = \langle a, x^*\rangle$ (see for example [28, Chapter 3]), we can then write down the following expressions for the convex conjugates G^* and F^*:

$$\begin{aligned} G^* &: \mathrm{C}(X\times Y) \to \{0,\infty\}, & G^*(u) &= \begin{cases} 0, & \text{if } u(x,y) \le c(x,y) \;\; \forall x\in X, y\in Y, \\ +\infty, & \text{otherwise,} \end{cases} \\ F^* &: \mathrm{C}(X)\times\mathrm{C}(Y) \to \mathbb{R}, & F^*(\varphi,\psi) &= \int_X \varphi(x)\mathrm{d}\mu(x) + \int_Y \psi(y)\mathrm{d}\nu(y). \end{aligned}$$

From the definition of the adjoint operator $\langle Ax, x^*\rangle = \langle x, A^*x^*\rangle \;\forall x\in X, x^*\in X^*$, we can compute also

$$A^* : \mathrm{C}(X)\times\mathrm{C}(Y) \to \mathrm{C}(X\times Y), \quad A^*(\varphi,\psi)(x,y) = \varphi(x)+\psi(y).$$

We also note that in the chosen setting both F and G are proper convex and lower semicontinuous, operator A is a linear bounded map. Let $\varphi(x) = x$ and $\psi(y) = y$ for $x\in X$ and $y\in Y$. Function F^* is continuous and finite at (x,y) and $G^*(-A^*(x,y)) = 0$ is finite, so condition (2.2.8) is satisfied and therefore strong duality holds.

Theorem 2.2.9 (Primal–dual optimality conditions [107, Theorem 31.3])**.** *For the optimization problems as in Theorem 2.2.6, a primal-dual pair $(x,y^*)\in X\times Y^*$ is optimal if and only if*

$$Ax \in \partial F^*(-y^*) \;\Leftrightarrow\; -y^* \in \partial F(Ax) \;\Leftrightarrow\; F(Ax) + F^*(-y^*) = -\langle A^*y^*, x\rangle, \quad (2.2.10)$$

$$A^*y^* \in \partial G(x) \;\Leftrightarrow\; x \in \partial G^*(A^*y^*) \;\Leftrightarrow\; G(x) + G^*(A^*y^*) = \langle A^*y^*, x\rangle. \quad (2.2.11)$$

Example 2.2.10 (Primal-dual optimality conditions for optimal transport)**.** For the balanced optimal transport, a primal feasible γ (2.1.3) and a dual feasible pair (φ, ψ) (2.2.9) are optimal if and only if

$$\operatorname{spt}(\gamma) \subset \{(x, y) \in X \times Y : \varphi(x) + \psi(y) = c(x, y)\}. \tag{2.2.12}$$

For the primal candidate γ, dual candidates (φ, ψ) and F, G and A as in the Example 2.2.8 condition (2.2.10) gives the marginal constraints:

$$\mathrm{P}_1\gamma = \mu, \qquad\qquad \mathrm{P}_2\gamma = \nu.$$

The second condition (2.2.11) can be written as

$$\langle \varphi + \psi - c, \gamma \rangle = \begin{cases} 0, & \text{if } \gamma(x, y) \geq 0, \;\; \varphi(x) + \psi(y) - c(x, y) \leq 0 \;\; \forall x \in X, y \in Y, \\ +\infty, & \text{otherwise.} \end{cases}$$

By studying the case when the obtained expression is finite, we can find out the following slackness conditions: Whenever $\gamma(x, y) > 0$ and $\varphi(x) + \psi(y) - c(x, y) \leq 0$, the right side of the expression is equal to 0, and so should the left side, meaning that $\varphi(x) + \psi(y) - c(x, y) = 0$; when $\gamma(x, y) = 0$ and $\varphi(x) + \psi(y) - c(x, y) \leq 0$, both sides are equal to 0 even if the second condition is satisfied as a strict inequality. Therefore, we conclude that for any $x \in X, y \in Y$, if $\gamma(x, y) > 0$, then $\varphi(x) + \psi(y) = c(x, y)$, which can also be written as (2.2.12).

2.3 Numerical approaches for optimal transport

In this section, we briefly discuss some numerical methods and techniques for optimal transport. It is not meant as a full taxonomy, but rather just shows the diversity and variety of the existing methods and puts this research into context.

Linear programming

As mentioned earlier in this chapter, optimal transport can be seen as a special case of linear programming. Therefore, the solvers developed for linear programs can also be applied to discrete optimal transport problems.

The first broad class of methods for linear programming is the methods which follow the boundary of the feasible set of the problem until a vertex of the optimal solution is reached. The methods of this class are the simplex method developed by Dantzig in the 1940s (the original paper is classified, see instead the paper on generalized simplex method by Dantzig et al. [55] and the historic overview [53]) and its variations and generalizations, such as network simplex method [52, 48], dual simplex method [83], or primal-dual simplex method [54].

Another large group of methods commonly used for linear programming is interior point methods, also known as barrier functions methods. In the methods of this class, the linear program is converted into an unconstrained minimization problem by adding a barrier term which penalizes leaving the feasible set. Different versions of interior point methods for linear programming were proposed among others by Dikin [56], Khachyan [77] and Karmakar [76].

It can be noted that there are also methods for solving linear programs that do not belong to either of the two classes above (e.g. a method proposed by Seidel [112] or Meggido's algorithms [58]), however, the methods from the discussed classes remain prevailing.

Linear assignment problem

An important special case of optimal transport problem is the *linear assignment problem*:

$$\min \left\{ \langle C, x \rangle \;\middle|\; x \in \{0,1\}^{n \times n} : \mathrm{P}_1 x = 1_n, \mathrm{P}_2 x = 1_n \right\}, \tag{2.3.1}$$

where $C \in \mathbb{R}^{n \times n}$ is the cost matrix, $\langle \cdot, \cdot \rangle$ denotes Frobenius inner product, 1_n is the vector with all entries 1, and the projection operators are $\mathrm{P}_1 x = \sum_{j=1}^{n} x_{ij}$ and $\mathrm{P}_2 x = \sum_{i=1}^{n} x_{ij}$.

One of the first efficient numerical methods for the linear assignment problem called *Hungarian method* was proposed by Kuhn [81]. It is a direct method, which first selects a dual feasible solution and then improves it on every step until the optimal assignment is found or the problem's infeasibility is established.

Another commonly used linear assignment solver is the *auction algorithm* due to Bertsekas [16]. It is an iterative algorithm, which solves a relaxed assignment problem and finds partial assignments on each iteration. The method is especially efficient when implemented together with an ε-scaling technique – applying the algorithm several times with a decreasing relaxation parameter ε [17]. The auction algorithm has later been adapted to the general discrete transportation problem [19].

Some of the general linear programming algorithms have also been adapted to the specifics of the linear assignment problem, for example, the dual simplex [69]. See also [57] for a broad overview of other algorithms for linear assignment problems.

Although the methods of linear programming are well-developed and well-studied, they generally lack efficiency when applied to large-scale optimal transport problems, which becomes particularly daunting when one (or both) of the input measures are *not* represented as a sum of single Dirac measures and require discretization.

Based on partial differential equations

As already stated in Section 2.1.3, in some cases, the Wasserstein distance between two given measures can be expressed through solutions of a continuity equation. The original paper by Benamou and Brenier proposes a variant of an augmented Lagrangian method on a finite difference staggered grid for numerical solution of the dedicated optimization problem [11]. Numerical schemes involving similar grids combined with various proximal splitting methods have been proposed and studied in [102]. A solver based on nested finite volume discretization and interior point method was proposed in [98].

Another wide class of methods is based on solving the Monge–Ampère equation. The Monge–Ampère equation is a non-linear elliptic partial differential equation, which we introduce as follows after [14]. Assuming $X, Y \subset \mathbb{R}^n$ compact and turning to the Monge formulation of optimal transport (2.1.1) for quadratic cost $c(x, y) = \|x - y\|^2$, one can write the mass preservation condition for a map $T : X \to Y$ as

$$\nu(T) \det(\nabla T) = \mu, \tag{2.3.2}$$

where ∇u is the gradient of the function u. One can then represent the unique map T_* minimizing (2.1.1) as a gradient of a convex function $u : X \to \mathbb{R}$, i.e. $T_* = \nabla u$. Formally substituting it into (2.3.2), one obtains the Monge–Ampère equation:

$$\det(D^2 u(x)) = \frac{\mu(x)}{\nu(\nabla u(x))}, \quad \text{for } x \in X,$$

with the conditions that u is convex and that the gradient map takes X to Y: $\nabla u(X) = Y$ (here D^2u is the Hessian of the function u).

The authors of [14] propose relaxation of the boundary conditions and offer a review of several dedicated domain discretization approaches and solution methods.

Some other insightful examples of partial differential equations in connection with optimal transport together with a discussion of strategies for their solution can be found in [60].

Semi-discrete transport

In the problems of semi-discrete transport, the transport is conducted between a discrete and a continuous measure. It should be noted that semi-discrete transport has its own applications, however it can also be used as a discretization of a continuous transport problem [96]. In 2D and 3D the problem can be constructively viewed from a geometric perspective: For costs $c(x, y) = \|x - y\|$ and $c(x, y) = \|(\| x - y)^2$, the space of the continuous input measure can be discretized using Laguerre cells (or their generalizations), where each cell should be mapped into a separate component of the discrete measure [104, Chapter 5]. Numerical schemes for more general settings based on solving the Monge–Ampère equation have been proposed and discussed in [100, 7, 96]. Broad reviews of the numerical approaches can be found in [85] and [26].

Entropic optimal transport

Here we primarily follow the description of the entropic optimal transport and Sinkhorn algorithm from [104, Chapter 4].

Definition 2.3.1. *Entropy-regularized optimal transport problem* is

$$\inf \left\{ \int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y) + \varepsilon \,\mathrm{KL}(\gamma \mid \mu \otimes \nu) \;\middle|\; \gamma \in \Gamma(\mu,\nu) \right\}, \tag{2.3.3}$$

where $\varepsilon > 0$ is a regularization parameter, $\mu \otimes \nu \in \mathcal{M}_+(X \times Y)$ is the product measure, i.e.

$$\int_{X\times Y} f(x,y)\mathrm{d}(\mu\otimes\nu)(x,y) = \int_{X\times Y} f(x,y)\mathrm{d}\mu(x)\mathrm{d}\nu(y),$$

and

$$\mathrm{KL}(\gamma \mid \pi) = \int_{X,Y} \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\pi}(x,y)\right)\mathrm{d}\gamma(x,y) + \int_{X,Y} (\mathrm{d}\pi(x,y) - \mathrm{d}\gamma(x,y))$$

is a generalization of the discrete Kullback–Leibler divergence (see Definition 2.1.5) with a convention $\mathrm{KL}(\gamma \mid \pi) = +\infty$ if γ does not have density w.r.t. π.

With $\varepsilon > 0$, problem (2.3.3) is strictly convex. It can be shown that the unique solution of (2.3.3) converges with $\varepsilon \to 0$ to the maximum entropy solution of the unregularized problem (2.1.3) [84].

The dual problem to (2.3.3) is given by

$$\sup \Big\{ \int_X \varphi(x)\mathrm{d}\mu(x) + \int_Y \psi(y)\mathrm{d}\nu(y) - \varepsilon \int_{X\times Y} e^{\frac{-c(x,y)+\varphi(x)+\psi(y)}{\varepsilon}}\mathrm{d}\mu(x)\mathrm{d}\nu(y) \;\Big|\; (\varphi,\psi) \in \mathrm{C}(X)\times\mathrm{C}(Y)\Big\}.$$

Let the input marginals μ, ν be discrete so that $\mu = \sum_{i=1}^m \bar{\mu}_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n \bar{\nu}_j \delta_{y_j}$ with $(x_i) \subset X, (y_j) \subset Y$, $\bar{\mu} \in \mathbb{R}_+^m$, $\bar{\nu} \in \mathbb{R}_+^n$. Let matrix $C \in \mathbb{R}^{m\times n}$ store the values of the cost function c at the support of the discrete measures: $C_{ij} = c(x_i, y_j)$. Let $K \in \mathbb{R}^{m\times n}$ be defined as

$$K_{ij} = e^{-\frac{C_{ij}}{\varepsilon}}.$$

Sinkhorn's algorithm consists in performing subsequent updates

$$u^{(k+1)} = \frac{\bar{\mu}}{Kv^{(k)}}, \qquad v^{(k+1)} = \frac{\bar{\nu}}{Ku^{(k+1)}} \tag{2.3.4}$$

for $k = 0, 1, 2, \ldots$ with arbitrary $v^{(0)} \in \mathbb{R}^n_+$; the division is performed component-wise. Vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ are called *(diagonal) scaling factors*. The regularized transport plan $\gamma \in \mathbb{R}^{m \times n}$ can then be retrieved as

$$\gamma_{ij} = u_i K_{ij} v_j, \quad i = 1, \ldots, m, \; j = 1, \ldots, n.$$

The dual potentials $\varphi \in \mathbb{R}^m, \psi \in \mathbb{R}^n$ can be retrieved as

$$\varphi = \varepsilon \log(u), \qquad \psi = \varepsilon \log(v).$$

The algorithm is named after Richard Sinkhorn who first published convergence results for iterations of the type (2.3.4) in [113], although methods based on similar iterations have been considered before (see a short overview in [104, Remark 4.5]).

It is important to note that when the regularization parameter $\varepsilon \to 0$, the regularized optimal transport problem converges to the unregularized one, but the numerical stability of the Sinkhorn iterations deteriorates, as the values of matrix K become too small and eventually comparable to the computational error. This problem can be overcome by performing the computations in the log domain, which however can lead to increased computational complexity. We refer to [110, Section 3.1] for an approach offering a balance between stability and speed.

Apart from numerical instability, when $\varepsilon \to 0$, the algorithm also exhibits a slower convergence rate. To remedy this, it has been proposed to use an ε-scaling scheme [80, 110], similar to the scheme used in the auction algorithm.

Another potential computational issue is the sparsity of the solution. In unregularized optimal transport problems, the optimal transport plan is usually concentrated on a small subset of the product space $X \times Y$, and while the introduction of the regularization term diffuses the solution, with the decrease of the value of ε the sparsity should be retrieved again. Therefore, the numerical solver should be flexible to allow for different levels of resolution depending on the current approximation. A dedicated kernel truncation procedure based on local duality gap estimations together with a hierarchical multi-scale scheme can be found in [110]. An alternative technique for sparsity preservation (based on Lagrangian problem description) has been proposed in [61, Chapter 4].

It should be also noted that the Sinkhorn algorithm can be generalized and adapted to other optimal transport problems, including unbalanced settings and barycenter problem [12, 41].

3 Hellinger–Kantorovich Barycenter

This chapter is devoted to the study of the barycenter problem in the Hellinger–Kantorovich metric. The results presented in this chapter were published in [23].

Wasserstein barycenter problem introduced in [1] (see the statement in Section 2.1.3) can be extended to the unbalanced setting by considering the barycenter problem with an unbalanced optimal transport metric. For scaled Hellinger–Kantorovich distance (see Section 2.1.2), the barycenter problem has been studied in [65, 43]. In particular, to better understand the role of the scaling parameter κ, the authors of [65] studied the case where the input measures are all *Dirac measures*. Unexpectedly, even on some simple analytical and preliminary numerical examples, they observed a non-trivial structure depending on the length scale parameter. The behavior seemed reminiscent of hierarchical clustering methods where the number of clusters is chosen automatically, depending on κ. Between transitions of different cluster numbers sometimes a diffuse intermediate solution was observed. In these cases, the solution was shown to be non-unique and a discrete solution was always shown to exist as well.

The goal of the research presented in this chapter is to study the Hellinger–Kantorovich barycenter problem and the properties of its solutions in more detail, with a focus on the case of Dirac marginals. This chapter is organized as follows.

Section 3.1 recalls some background on the Hellinger–Kantorovich distance.

Throughout Section 3.2 we study the HK barycenter problem between a continuum of general (non-Dirac) input measures. We provide existence and stability under changes in the distribution of input measures and scaling parameter κ, including the limits $\kappa \to 0$ (Hellinger limit) or ∞ (Wasserstein limit). A dual problem is derived that will become instrumental in the analysis of the Dirac case.

Section 3.3 is dedicated to the Dirac case. We give simplified expressions for the primal and dual objectives and show existence and uniqueness of dual solutions, primal-dual optimality conditions, and dual stability with respect to the variations in the distribution of the input Dirac measures and the length scale parameter κ. The solution for the $\kappa = 0$ limit is given explicitly and the asymptotic behavior as $\kappa \to 0$ is described in terms of total mass and local mass density of the minimizer. Finally, we turn to the question of the sparsity of the minimizers. We give an alternative proof to that of [65] that discrete minimizers exist when the distribution of the input measures consists of a finite number of Diracs. But conversely, we also give analytical examples for which no discrete minimizers exist for a continuum of marginal measures.

Section 3.4 discusses numerical approximation and examples. We propose a non-convex Lagrangian discretization, reminiscent of methods for quantization problems. It provides high spatial accuracy in the case of sparse solutions. Unlike the quantization problem, missing points can be detected by sampling the dual potential. We illustrate that the evolution of the barycenter is stable with respect to κ, but far from a simple successive merging of clusters. Instead, a

wide variety of transition behaviors is documented. The convergence of the barycenter as the sequence of sampled empirical distributions of the input data converges to a "true" distribution is visualized. We observe that for some values of κ the HK barycenter seems to be unique and can be approximated well numerically, whereas for other values (usually the "transition regimes") this proves to be quite challenging since either it is non-unique or the basin around the minimizer is extremely shallow, as evidenced by very degenerate primal-dual slackness conditions. Non-uniqueness of minimizers and a vast set of near-optimizers are common phenomena in non-convex measure quantization, see e.g. [30, Section 4.1].

In conclusion, the HK barycenter between Dirac measures does not provide a novel straightforward method for hierarchical point clustering, since the evolution of the minimizer with respect to the length scale parameter does not correspond to a simple successive merging of clusters, and sometimes even only diffuse solutions exist. But it does provide an interpolation between the input data and a single Dirac measure, parametrized by a single length scale parameter, that can be interpreted as gradual coarse graining. It is provably stable with respect to the input data and scale changes and comes with a corresponding sequence of dual problems with unique solutions, that provide additional interpretation via the primal-dual optimality relations and information for numerical approximation. A summarizing discussion is given in Section 3.5.

Author's contribution

The author has made minor contributions to Section 3.2.

The author has made major contributions to Sections 3.3 and 3.4. In particular, in Section 3.3 the author contributed to formulating, validating or disproving the conjectures about the properties of the barycenter of a continuum of Dirac measures. In Section 3.4 the author was responsible for implementing, testing and improving the numerical schemes, as well as for conducting numerical experiments and interpreting the results.

3.1 Hellinger–Kantorovich distance

Here we briefly recall the properties of Hellinger–Kantorovich distance required in this chapter. First we give an explicit expression for it already mentioned in Section 2.1.2 and then give an alternative formulation.

Throughout this chapter we let $\Omega \subset \mathbb{R}^n$ be a compact and convex set with non-empty interior.

For $\mu, \nu \in \mathcal{M}_+(\Omega)$ the scaled *Hellinger–Kantorovich distance* HK_κ is given by [86]

$$\mathrm{HK}_\kappa^2(\mu,\nu) := \inf\left\{ \int_{\Omega^2} \hat{c}_\kappa(x,y)\,\mathrm{d}\gamma(x,y) + \mathrm{KL}(\mathrm{P}_1\gamma|\mu) + \mathrm{KL}(\mathrm{P}_2\gamma|\nu) \middle| \gamma \in \mathcal{M}_+(\Omega^2) \right\}, \tag{3.1.1}$$

where

$$\hat{c}_\kappa(x,y) := \begin{cases} -2\log\cos(|x-y|/\kappa) & \text{for } |x-y| < \kappa\pi/2, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.1.2}$$

This is an optimal transport problem where the marginal constraints are relaxed and deviations from the marginals μ and ν are admissible and penalized by the Kullback–Leibler divergence, allowing for changes of mass. Parameter $\kappa > 0$ is a length scale parameter that balances the trade-off between transport and mass change. From the definition of $\hat{c}_\kappa$ we infer that mass is never transported further than $\kappa\pi/2$, in particular κ effectively re-scales the Euclidean distance on Ω, as elaborated in the following Remark.

Remark 3.1.1. For $\kappa \in (0,\infty)$ let $S : \Omega \to \Omega/\kappa$, $x \mapsto x/\kappa$. Then for $\mu,\nu \in \mathcal{M}_+(\Omega)$ one obtains that $\mathrm{HK}^2_\kappa(\mu,\nu) = \mathrm{HK}^2_1(S_\#\mu, S_\#\nu)$ where the latter distance is computed on $\mathcal{M}_+(\Omega/\kappa)$. This follows quickly from the fact that S is a homeomorphism between Ω and Ω/κ, implying for instance $\mathrm{KL}(\mathrm{P}_1\gamma|\mu) = \mathrm{KL}(\mathrm{P}_1 S_\#\gamma|S_\#\mu)$.

We also recall an alternative formulation for HK_κ, as given by [42]. Let $\mathrm{Cos}\colon \mathbb{R}\to\mathbb{R}$ denote the truncated cosine function defined as

$$\mathrm{Cos}(s) := \cos\left(\min\{|s|,\ \pi/2\}\right) \quad \text{for } s\in\mathbb{R}$$

and consider the following cost function $c_\kappa\colon \Omega\times\mathbb{R}\times\Omega\times\mathbb{R}\to\mathbb{R}\cup\{\infty\}$,

$$c_\kappa(x_1,m_1,x_2,m_2) := \begin{cases} m_1+m_2-2\sqrt{m_1m_2}\,\mathrm{Cos}(|x_1-x_2|/\kappa) & \text{if } m_1,m_2\geq 0,\\ +\infty & \text{otherwise.}\end{cases}$$

Then, the scaled Hellinger–Kantorovich distance HK_κ can be written as

$$\mathrm{HK}^2_\kappa(\mu,\nu) = \inf\left\{ \int_{\Omega^2} c_\kappa\left(x, \tfrac{\mathrm{d}\gamma_1}{\mathrm{d}\gamma}(x,y), y, \tfrac{\mathrm{d}\gamma_2}{\mathrm{d}\gamma}(x,y)\right)\mathrm{d}\gamma(x,y) \;\middle|\; \gamma_1,\gamma_2,\gamma\in\mathcal{M}_+(\Omega^2),\ \gamma_1,\gamma_2\ll\gamma \text{ and } \mathrm{P}_1\gamma_1=\mu, \mathrm{P}_2\gamma_2=\nu\right\}. \tag{3.1.3}$$

Note that γ in (3.1.3) is just an auxiliary variable and the integral does not depend on the choice of γ by positive 1-homogeneity of c_κ in its second and fourth argument.

A dual formulation for (3.1.1) and (3.1.3) is given by [42]

$$\mathrm{HK}^2_\kappa(\mu,\nu) = \sup_{(\psi,\phi)\in Q_\kappa}\left[\int_\Omega \psi(x)\,\mathrm{d}\mu(x) + \int_\Omega \phi(y)\,\mathrm{d}\nu(y)\right], \tag{3.1.4}$$

where the set Q_κ is defined by

$$Q_\kappa := \left\{ (\psi,\phi)\in \mathrm{C}(\Omega)\times\mathrm{C}(\Omega)\ \text{ s.t. } \begin{array}{l}\psi(x),\phi(y)\in(-\infty,1],\\ (1-\psi(x))(1-\phi(y))\geq \mathrm{Cos}^2(|x-y|/\kappa)\end{array} \quad \text{for all } x,y\in\Omega\right\}. \tag{3.1.5}$$

Theorem 3.1.2 (Basic properties of HK_κ)**.** *For any $\kappa\in(0,\infty)$, the Hellinger–Kantorovich distance HK_κ metrizes the weak* topology over $\mathcal{M}_+(\Omega)$. The metric space $(\mathcal{M}_+(\Omega),\mathrm{HK}_k)$ is separable and complete. Furthermore, it is a proper metric space, i.e. every bounded set is relatively compact (see [86, Section 7.5]).*

As we seek to study the evolution of the Hellinger–Kantorovich barycenter over varying length scales we now recall the corresponding result. For $\mu,\nu\in\mathcal{M}_+(\Omega)$ the *Hellinger distance* is defined as

$$\mathrm{Hell}^2(\mu,\nu) := \int_\Omega \left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\tau}} - \sqrt{\frac{\mathrm{d}\nu}{\mathrm{d}\tau}}\right)^2 \mathrm{d}\tau, \tag{3.1.6}$$

where $\tau\in\mathcal{M}_+(\Omega)$ is an arbitrary measure such that $\mu,\nu\ll\tau$. Again, since the function $(s,t)\mapsto(\sqrt{s}-\sqrt{t})^2$ is positively 1-homogeneous, the definition of $\mathrm{Hell}(\mu,\nu)$ does not depend on the choice of the (admissible) τ. As observed in [86], the Hellinger–Kantorovich distance converges towards the Hellinger and the Wasserstein distance as one sends $\kappa\to 0$ or $\kappa\to\infty$ respectively.

Theorem 3.1.3 (Scaling limits [86, Theorems 7.22, 7.24]). *For $\mu, \nu \in \mathcal{M}_+(\Omega)$, one finds that the function $(0,\infty) \ni \kappa \mapsto \mathrm{HK}^2_\kappa(\mu,\nu)$ is non-increasing and*

$$\lim_{\kappa\to 0} \mathrm{HK}^2_\kappa(\mu,\nu) \nearrow \mathrm{Hell}^2(\mu,\nu). \tag{3.1.7}$$

On the other hand, the function $(0,\infty) \ni \kappa \mapsto \kappa^2 \cdot \mathrm{HK}^2_\kappa(\mu,\nu)$ is non-decreasing and

$$\lim_{\kappa\to\infty} \kappa^2\, \mathrm{HK}^2_\kappa(\mu,\nu) \nearrow W^2(\mu,\nu). \tag{3.1.8}$$

The function $(0,\infty) \ni \kappa \mapsto \mathrm{HK}^2_\kappa(\mu,\nu)$ is continuous.

Continuity of HK^2_κ with respect to κ follows directly from the fact that the function is non-increasing while $\kappa \mapsto \kappa^2 \cdot \mathrm{HK}^2_\kappa(\mu,\nu)$ is non-decreasing. Note that the case $\|\mu\| \neq \|\nu\|$ is explicitly allowed as $\kappa \to \infty$, in which case the limiting value is $+\infty$. These scaling limits can be guessed from (3.1.1): As $\kappa \to 0$, the function $\hat{c}_\kappa$ goes to infinity everywhere except on the diagonal, restricting asymptotically feasible γ to the diagonal. One can then quickly verify that minimizing (3.1.1) only over diagonal γ yields (3.1.6). Conversely, looking at $\kappa^2 \cdot \mathrm{HK}^2_\kappa(\mu,\nu)$, one can guess from $\lim_{\kappa\to\infty} \kappa^2 \cdot \hat{c}_\kappa(x,y) = \|x-y\|^2$ that the integral $\int_{\Omega^2} \hat{c}_\kappa \mathrm{d}\gamma$ converges to the standard Wasserstein transport cost, while the term $\kappa^2 \cdot \mathrm{KL}(\mathrm{P}_1\gamma|\mu)$ increasingly penalizes deviations between $\mathrm{P}_1\gamma$ and μ, and likewise for the second marginal term, thus asymptotically enforcing $\gamma \in \Gamma(\mu,\nu)$.

The following bounds can be shown to hold.

Proposition 3.1.4 (Mass-rescaling for HK_κ). *For $\kappa \in (0,\infty)$ and $\mu, \nu \in \mathcal{M}_+(\Omega)$, one finds*

$$\mathrm{HK}^2_\kappa(\mu,\nu) = \sqrt{\|\mu\|\|\nu\|}\, \mathrm{HK}^2_\kappa\left(\frac{\mu}{\|\mu\|}, \frac{\nu}{\|\nu\|}\right) + (\sqrt{\|\mu\|} - \sqrt{\|\nu\|})^2 \tag{3.1.9}$$

with the convention $\mu/\|\mu\| = 0$ in the case of $\mu = 0$ (and likewise for ν). Additionally,

$$\mathrm{HK}^2_\kappa(\mu,\nu) \leq \|\mu\| + \|\nu\|. \tag{3.1.10}$$

Proof. The equality in (3.1.9) follows from [82, Theorem 3.3]. By Theorem 3.1.3, we have $\mathrm{HK}^2_\kappa(\mu,\nu) \leq \mathrm{Hell}^2(\mu,\nu)$, and (3.1.10) follows directly. □

3.2 Hellinger–Kantorovich barycenter of a continuum of measures

3.2.1 Problem setup

The barycenter between a finite collection of measures with respect to the Hellinger–Kantorovich metric has been studied in [43, 65]. In this section we generalize these results to infinitely many input measures, including the uncountable case of a continuum of input measures.

For a constant $M \in (0,\infty)$, which we will assume to be fixed throughout this chapter, we define

$$\mathfrak{C} := \{\mu \in \mathcal{M}_+(\Omega) \mid \|\mu\| \leq M\}.$$

Since $\mathfrak{C}$ is weak* closed, by Theorem 3.1.2 the metric space $(\mathfrak{C}, \mathrm{HK}_\kappa)$ is compact for all $\kappa \in (0,\infty)$. We will describe the collection of input measures (and their weights) of which to compute the barycenter as a probability measure $\Lambda \in \mathcal{P}(\mathfrak{C})$ where $\mathfrak{C}$ is equipped with the Borel σ-algebra induced by HK_κ (which is the same for any $\kappa \in (0,\infty)$). Since $(\mathfrak{C}, \mathrm{HK}_\kappa)$ is compact, weak* convergence on $\mathcal{P}(\mathfrak{C})$ is metrized by the Wasserstein distance over $\mathcal{P}(\mathfrak{C})$ (see Theorem 2.1.4).

For $\Lambda \in \mathcal{P}(\mathfrak{C})$ and for $\kappa \in (0,\infty)$, the primal problem we are interested in is

$$\inf \left\{ J_{\Lambda,\kappa}(\nu) := \int_{\mathfrak{C}} \mathrm{HK}_\kappa^2(\mu,\nu)\,\mathrm{d}\Lambda(\mu) \,\middle|\, \nu \in \mathcal{M}_+(\Omega) \right\}. \tag{$\mathcal{P}_{\Lambda,\kappa}$}$$

The finite case of computing the barycenter between input measures $\mu_1,\dots,\mu_n \in \mathfrak{C}$ with weights $\lambda_1,\dots,\lambda_n$ where $\lambda_i > 0$ and $\sum_{i=1}^n \lambda_i = 1$ is recovered by setting $\Lambda := \sum_{i=1}^n \lambda_i\,\delta_{\mu_i}$.

3.2.2 Existence and stability of minimizers

Proposition 3.2.1. *Let $\Lambda \in \mathcal{P}(\mathfrak{C})$ and $\kappa \in (0,\infty)$. Then, ($\mathcal{P}_{\Lambda,\kappa}$) admits a minimizer $\nu \in \mathfrak{C}$.*

Proof. We first observe that, by means of the upper bound in (3.1.10), we have

$$(\mathcal{P}_{\Lambda,\kappa}) \;\le J_{\Lambda,\kappa}(0) = \int_{\mathfrak{C}} \mathrm{HK}_\kappa^2(\mu,0)\,\mathrm{d}\Lambda(\mu) \le \int_{\mathfrak{C}} \|\mu\|\,\mathrm{d}\Lambda(\mu) \le M.$$

Let $(\nu_n)_n \subset \mathcal{M}_+(\Omega)$ be a minimizing sequence for ($\mathcal{P}_{\Lambda,\kappa}$). For each $n > 0$, we can assume without loss of generality that $\nu_n \in \mathfrak{C}$. Indeed, suppose this is not the case, i.e. $\|\nu_n\| > M$. Then, for all $\mu \in \mathfrak{C}$, by means of (3.1.9), we have

$$\begin{aligned}
\mathrm{HK}_\kappa^2\left(\mu, \frac{M}{\|\nu_n\|}\nu_n\right) &= \sqrt{\|\mu\| M}\;\mathrm{HK}_\kappa^2\left(\frac{\mu}{\|\mu\|}, \frac{\nu_n}{\|\nu_n\|}\right) + (\sqrt{\|\mu\|} - \sqrt{M})^2 \\
&< \sqrt{\|\mu\|\|\nu_n\|}\;\mathrm{HK}_\kappa^2\left(\frac{\mu}{\|\mu\|}, \frac{\nu_n}{\|\nu_n\|}\right) + (\sqrt{\|\mu\|} - \sqrt{\|\nu_n\|})^2 = \mathrm{HK}_\kappa^2(\mu,\nu_n),
\end{aligned}$$

so that $J_{\Lambda,\kappa}(M/\|\nu_n\|\cdot\nu_n) < J_{\Lambda,\kappa}(\nu_n)$. Hence, upon possibly replacing ν_n with $M/\|\nu_n\|\cdot\nu_n$, the sequence $(\nu_n)_n$ is entirely contained in $\mathfrak{C}$. By compactness of $(\mathfrak{C},\mathrm{HK}_\kappa)$, there exists a cluster point $\nu \in \mathfrak{C}$ such that, up to a subsequence, $\nu_n \rightharpoonup^* \nu$ as $n \to \infty$, or equivalently $\mathrm{HK}_\kappa(\nu_n,\nu) \to 0$ as $n \to \infty$. By the upper bound in (3.1.10) and by triangle inequality we also have

$$\begin{aligned}
|\,\mathrm{HK}_\kappa^2(\mu,\nu_n) - \mathrm{HK}_\kappa^2(\mu,\nu)| &= |\,\mathrm{HK}_\kappa(\mu,\nu_n) + \mathrm{HK}_\kappa(\mu,\nu)|\,|\,\mathrm{HK}_\kappa(\mu,\nu_n) - \mathrm{HK}_\kappa(\mu,\nu)| \\
&\le 2\sqrt{2M}\cdot \mathrm{HK}_\kappa(\nu_n,\nu) \to 0 \quad \text{as } n\to\infty \text{ for all } \mu \in \mathfrak{C}.
\end{aligned} \tag{3.2.1}$$

By means of Fatou's lemma and recalling that $(\nu_n)_n$ is a minimizing sequence, we conclude

$$J_{\Lambda,\kappa}(\nu) = \int_{\mathfrak{C}} \mathrm{HK}_\kappa^2(\mu,\nu)\,\mathrm{d}\Lambda(\mu) = \int_{\mathfrak{C}} \lim_{n\to\infty} \mathrm{HK}_\kappa^2(\mu,\nu_n)\,\mathrm{d}\Lambda(\mu) \le \liminf_{n\to\infty} \underbrace{\int_{\mathfrak{C}} \mathrm{HK}_\kappa^2(\mu,\nu_n)\,\mathrm{d}\Lambda(\mu)}_{J_{\Lambda,\kappa}(\nu_n)} = (\mathcal{P}_{\Lambda,\kappa}),$$

which provides minimality of ν for ($\mathcal{P}_{\Lambda,\kappa}$). □

Proposition 3.2.2 (Stability)**.** *Fix $\kappa \in (0,\infty)$. Let $(\Lambda_n)_{n\in\mathbb{N}}$ be a weak* convergent sequence in $\mathcal{P}(\mathfrak{C})$ with limit $\Lambda \in \mathcal{P}(\mathfrak{C})$ and let $(\nu_n)_{n\in\mathbb{N}}$ be a weak* convergent sequence in $\mathcal{M}_+(\Omega)$ with limit $\nu \in \mathcal{M}_+(\Omega)$. Then,*

$$J_{\Lambda,\kappa}(\nu) = \lim_{n\to\infty} J_{\Lambda_n,\kappa}(\nu_n). \tag{3.2.2}$$

Proof. As in (3.2.1), the sequence of functions $(\mu \mapsto \mathrm{HK}_\kappa^2(\mu,\nu_n))_n$ converges uniformly to $(\mu \mapsto \mathrm{HK}_\kappa^2(\mu,\nu))$ in $\mathrm{C}(\mathfrak{C})$. This, together with weak* convergence of Λ_n to Λ, leveraging duality between $\mathrm{C}(\mathfrak{C})$ and $\mathcal{M}_+(\mathfrak{C})$, leads to

$$\lim_{n\to\infty} J_{\Lambda_n,\kappa}(\nu_n) = \lim_{n\to\infty} \int_{\mathfrak{C}} \mathrm{HK}_\kappa^2(\mu,\nu_n)\,\mathrm{d}\Lambda_n(\mu) = \int_{\mathfrak{C}} \mathrm{HK}_\kappa^2(\mu,\nu)\,\mathrm{d}\Lambda(\mu) = J_{\Lambda,\kappa}(\nu),$$

which provides (3.2.2). □

Corollary 3.2.3 (Convergence of minimizers)**.** *Fix $\kappa \in (0,\infty)$. Let $(\Lambda_n)_{n\in\mathbb{N}}$ be a weak* convergent sequence in $\mathcal{P}(\mathfrak{C})$ with limit $\Lambda \in \mathcal{P}(\mathfrak{C})$ and, for each n, let ν_n be a minimizer of $(\mathcal{P}_{\Lambda_n,\kappa})$. Then, the sequence $(\nu_n)_{n\in\mathbb{N}}$ is weak* pre-compact and each cluster point $\nu \in \mathfrak{C}$ is a minimizer of $(\mathcal{P}_{\Lambda,\kappa})$.*

Proof. By Proposition 3.2.1, the sequence of minimizers $(\nu_n)_{n\in\mathbb{N}}$ lies entirely in $\mathfrak{C}$, hence by compactness of $(\mathfrak{C}, \mathrm{HK}_\kappa)$ it is weak* pre-compact. Fix now any weak* cluster point ν of $(\nu_n)_n$ and a corresponding subsequence $(\nu_{n'})_{n'}$ such that $\nu_{n'} \rightharpoonup^* \nu$ as $n' \to \infty$. Fix any $\tilde{\nu} \in \mathcal{M}_+(\Omega)$. Minimality of each ν_n for $(\mathcal{P}_{\Lambda_n,\kappa})$ and a double application of Proposition 3.2.2 provide

$$J_{\Lambda,\kappa}(\nu) = \lim_{n'\to\infty} J_{\Lambda_{n'},\kappa}(\nu_{n'}) \le \lim_{n'\to\infty} J_{\Lambda_{n'},\kappa}(\tilde{\nu}) = J_{\Lambda,\kappa}(\tilde{\nu}),$$

which proves minimality of ν for $(\mathcal{P}_{\Lambda,\kappa})$. □

3.2.3 Scaling limits for the metric

Now let us look at the limit problems as we send $\kappa \to 0$ and $\kappa \to \infty$ respectively. Based on Theorem 3.1.3 we expect to recover the pure Hellinger and pure Wasserstein barycenter problems (after suitable re-scaling). The expected limit functionals are therefore:

$$J_{\Lambda,0}(\nu) := \int_{\mathfrak{C}} \mathrm{Hell}^2(\mu,\nu)\,\mathrm{d}\Lambda(\mu), \qquad J_{\Lambda,\infty}(\nu) := \int_{\mathfrak{C}} W^2(\mu,\nu)\,\mathrm{d}\Lambda(\mu). \tag{3.2.3}$$

In particular, we obtain as a by-product the existence of minimizers for such limiting barycenter problems.

Proposition 3.2.4. *Let $\Lambda \in \mathcal{P}(\mathfrak{C})$, let $(\kappa_n)_n$ be a sequence in $(0,\infty)$ with $\lim_n \kappa_n = \kappa_\infty \in [0,\infty)\cup\{\infty\}$. For each n, let $\nu_n \in \mathfrak{C}$ be a minimizer of J_{Λ,κ_n}. Then, the sequence $(\nu_n)_n$ is weak* pre-compact and each cluster point $\nu_\infty \in \mathfrak{C}$ is a minimizer of $J_{\Lambda,\kappa_\infty}$. Furthermore,*

$$J_{\Lambda,\kappa_\infty}(\nu_\infty) = \lim_{n\to\infty} J_{\Lambda,\kappa_n}(\nu_n) \quad \text{if} \quad \kappa_\infty \in [0,\infty) \tag{3.2.4}$$

and

$$J_{\Lambda,\infty}(\nu_\infty) = \lim_{n\to\infty} \kappa_n^2 J_{\Lambda,\kappa_n}(\nu_n) \quad \text{if} \quad \kappa_\infty = \infty. \tag{3.2.5}$$

Proof. By Proposition 3.2.1 the sequence $(\nu_n)_n$ lies in $\mathfrak{C}$, thus has uniformly bounded mass and thus is weak* pre-compact. Let us assume for now that the sequence $(\kappa_n)_n$ is monotone and that the corresponding sequence $(\nu_n)_n$ converges weak* to $\nu_\infty \in \mathfrak{C}$.

Step 1.1 ($\kappa_n \nearrow \kappa_\infty$)
Assume the sequence $(\kappa_n)_n$ is non-decreasing and converging to $\kappa_\infty \in (0,\infty)$, and assume the corresponding sequence $(\nu_n)_n$ converges weak* to $\nu_\infty \in \mathfrak{C}$. By Theorem 3.1.3 the function $(0,\infty) \ni \kappa \mapsto \kappa^2 \cdot \mathrm{HK}^2_\kappa(\mu,\nu)$ is non-decreasing for all $\mu,\nu \in \mathfrak{C}(\Omega)$ and $\lim_{n\to\infty} \kappa_n^2\, \mathrm{HK}^2_{\kappa_n}(\mu,\nu) \nearrow \kappa_\infty^2\, \mathrm{HK}^2_{\kappa_\infty}(\mu,\nu)$. Therefore, for each $\nu \in \mathfrak{C}$, the function $(0,\infty) \ni \kappa \mapsto \kappa^2\, J_{\Lambda,\kappa}(\nu)$ is non-decreasing, so that

$$\kappa_m^2\, J_{\Lambda,\kappa_m}(\nu) \le \kappa_n^2\, J_{\Lambda,\kappa_n}(\nu) \quad \text{for all } n > m > 0, \text{ and all } \nu \in \mathcal{M}_+(\Omega), \tag{3.2.6}$$

and by monotone convergence

$$\kappa_n^2\, J_{\Lambda,\kappa_n}(\nu) \nearrow \kappa_\infty^2 J_{\Lambda,\infty}(\nu) \text{ as } n \to \infty, \text{ for all } \nu \in \mathcal{M}_+(\Omega). \tag{3.2.7}$$

Let us fix $m \in \mathbb{N}$. Thanks to (3.2.6), applied with $\nu = \nu_n$, we find

$$\kappa_m^2 \, J_{\Lambda,\kappa_m}(\nu_n) \leq \kappa_n^2 \, J_{\Lambda,\kappa_n}(\nu_n) \quad \text{for all } n > m,$$

so that, passing to the limit as $n \to \infty$, we obtain

$$\kappa_m^2 \, J_{\Lambda,\kappa_m}(\nu_\infty) \overset{(3.2.2)}{=} \kappa_m^2 \lim_{n\to\infty} J_{\Lambda,\kappa_m}(\nu_n) \leq \liminf_{n\to\infty} \kappa_n^2 \, J_{\Lambda,\kappa_n}(\nu_n).$$

Passing now to the limit as $m \to \infty$ we conclude

$$\kappa_\infty^2 J_{\Lambda,\kappa_\infty}(\nu_\infty) \overset{(3.2.7)}{=} \lim_{m\to\infty} \kappa_m^2 \, J_{\Lambda,\kappa_m}(\nu_\infty) \leq \liminf_{n\to\infty} \kappa_n^2 \, J_{\Lambda,\kappa_n}(\nu_n). \tag{3.2.8}$$

On the other hand, using (3.2.7) with $\nu = \nu_\infty$ and leveraging minimality of each ν_n, one has

$$\kappa_\infty^2 J_{\Lambda,\kappa_\infty}(\nu_\infty) \geq \kappa_n^2 \, J_{\Lambda,\kappa_n}(\nu_\infty) \geq \kappa_n^2 \, J_{\Lambda,\kappa_n}(\nu_n) \quad \text{for every } n,$$

and a passage to the limit as $n \to \infty$ directly provides

$$\kappa_\infty^2 J_{\Lambda,\kappa_\infty}(\nu_\infty) \geq \limsup_{n\to\infty} \kappa_n^2 \, J_{\Lambda,\kappa_n}(\nu_n). \tag{3.2.9}$$

Combining (3.2.9) and (3.2.8) provides (3.2.4) for the particular class of sequences under consideration. Assume now ν_∞ were not optimal for $J_{\Lambda,\kappa_\infty}$, i.e. there exists some ν'_∞ with a strictly better score. By minimality of each ν_n, we have $J_{\Lambda,\kappa_n}(\nu_n) \leq J_{\Lambda,\kappa_n}(\nu'_\infty)$ so that, passing to the limit one has

$$\lim_{n\to\infty} J_{\Lambda,\kappa_n}(\nu_n) \leq \lim_{n\to\infty} J_{\Lambda,\kappa_n}(\nu'_\infty) \overset{(3.2.7)}{=} J_{\Lambda,\kappa_\infty}(\nu'_\infty) < J_{\Lambda,\kappa_\infty}(\nu_\infty) \overset{(3.2.4)}{=} \lim_{n\to\infty} J_{\Lambda,\kappa_n}(\nu_n),$$

hence the sought for contradiction. The same argument applies if $\kappa_n \nearrow \infty$, taking into account that $\lim_{n\to\infty} \kappa_n^2 \, \mathrm{HK}^2_{\kappa_n}(\mu,\nu) \nearrow W^2(\mu,\nu)$ for all $\mu,\nu \in \mathfrak{C}(\Omega)$.

Step 1.2 ($\kappa_n \searrow \kappa_\infty$)
If we assume instead that the sequence $(\kappa_n)_n$ is non-increasing and $\kappa_n \searrow \kappa_\infty \in [0,\infty)$, a completely symmetric argument as in Step 1.1 can be applied after dropping the scaling factors κ_n^2. Indeed, by Theorem 3.1.3 the function $(0,\infty) \ni \kappa \mapsto \mathrm{HK}^2_\kappa(\mu,\nu)$ is non-increasing for all $\mu,\nu \in \mathfrak{C}(\Omega)$ and so $\lim_{n\to\infty} \mathrm{HK}^2_{\kappa_n}(\mu,\nu) \nearrow \mathrm{HK}^2_{\kappa_\infty}(\mu,\nu)$ (with limit $\mathrm{Hell}^2(\mu,\nu)$ if $\kappa_\infty = 0$). Thus, the same monotonicity arguments apply.

Step 2. Assume now $\nu_\infty \in \mathfrak{C}$ is any cluster point of $(\nu_n)_n$. Hence, there exists a subsequence $(\nu_{n'})_{n'}$ such that $\nu_{n'} \rightharpoonup^* \nu_\infty$ as $n' \to \infty$. We can extract an additional subsequence such that $(\kappa_{n''})_{n''}$ is either non-increasing or non-decreasing. Step 1 then provides minimality of ν_∞ for $(\mathcal{P}_{\Lambda,\kappa_\infty})$.

We are left to prove that the sequence of energies $(J_{\Lambda,\kappa_n}(\nu_n))_n$ converges as a whole. Consider any subsequence $(J_{\Lambda,\kappa_{n'}}(\nu_{n'}))_{n'}$. By pre-compactness of the corresponding sequence $(\nu_{n'})_{n'}$, we can identify a further subsequence such that $(\nu_{n''})_{n''}$ converges weak* to some cluster point $\nu_\infty \in \mathfrak{C}$. and in turn extract an additional subsequence such that $(\kappa_{n'''})_{n'''}$ is either non-increasing or non-decreasing. By Step 1 we have

$$\lim_{n'''\to\infty} J_{\Lambda,n'''}(\nu_{n'''}) = (\mathcal{P}_{\Lambda,\kappa_\infty}).$$

Hence, every subsequence of $(J_{\Lambda,\kappa_n}(\nu_n))_n$ admits a converging subsequence to the same limit $(\mathcal{P}_{\Lambda,\kappa_\infty})$. This provides convergence of the full sequence to $(\mathcal{P}_{\Lambda,\kappa_\infty})$ and proves (3.2.4) for the whole sequence of minimizing energies. □

For $\kappa \in (0, \infty]$ it is known that barycenters are not necessarily unique (see, e.g., [65, Section 6]), hence there may be multiple corresponding cluster points ν_∞ in the above result. We now show that for $\kappa = 0$ uniqueness holds in general.

Corollary 3.2.5 (Convergence of minimizers for $\kappa \to 0$)**.** *Let $\Lambda \in \mathcal{P}(\mathfrak{C})$ and for each $\kappa > 0$ let $\nu_\kappa \in \mathfrak{C}$ be a minimizer of $J_{\Lambda,\kappa}$. Then, there exists some $\nu_0 \in \mathfrak{C}$ such that $\nu_k \rightharpoonup^* \nu_0$ as $\kappa \to 0$. In particular, ν_0 is the unique minimizer of $J_{\Lambda,0}$.*

Proof. By Proposition 3.2.4, there exists a minimizer $\nu_0 \in \mathcal{M}_+(\Omega)$ of $J_{\Lambda,0}$. Such a minimizer is indeed unique: assume this were not the case, so that there exists a second minimizer $\nu_0' \in \mathcal{M}_+(\Omega)$. Fix any $\tau \in \mathcal{M}_+(\Omega)$ such that $\nu_0, \nu_0' \ll \tau$ and define $v_0 = \mathrm{d}\nu_0/\mathrm{d}\tau$ and $v_0' = \mathrm{d}\nu_0'/\mathrm{d}\tau$. Let $\bar{v} = \left(\frac{1}{2}\sqrt{v_0} + \frac{1}{2}\sqrt{v_0'}\right)^2$ and define $\bar{\nu} = \bar{v}\tau$ (note that this definition does not depend on the choice of τ by positive 1-homogeneity). For any $\mu \in \mathfrak{C}$, fix any $\tau_\mu \in \mathcal{M}_+(\Omega)$ such that $\tau, \mu \ll \tau_\mu$ and, by strict convexity of $x \mapsto x^2$, compute

$$\begin{aligned}
\mathrm{Hell}^2(\bar{\nu}, \mu) &= \int_\Omega \left(\sqrt{\frac{\mathrm{d}\bar{\nu}}{\mathrm{d}\tau_\mu}} - \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\tau_\mu}} \right)^2 \mathrm{d}\tau_\mu \\
&= \int_\Omega \left(\frac{1}{2}\left(\sqrt{v_0 \frac{\mathrm{d}\tau}{\mathrm{d}\tau_\mu}} - \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\tau_\mu}} \right) + \frac{1}{2}\left(\sqrt{v_0' \frac{\mathrm{d}\tau}{\mathrm{d}\tau_\mu}} - \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\tau_\mu}} \right) \right)^2 \mathrm{d}\tau_\mu \\
&< \frac{1}{2}\int_\Omega \left(\sqrt{v_0 \frac{\mathrm{d}\tau}{\mathrm{d}\tau_\mu}} - \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\tau_\mu}} \right)^2 \mathrm{d}\tau_\mu + \frac{1}{2}\int_\Omega \left(\sqrt{v_0' \frac{\mathrm{d}\tau}{\mathrm{d}\tau_\mu}} - \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\tau_\mu}} \right)^2 \mathrm{d}\tau_\mu \\
&= \frac{1}{2}(\mathrm{Hell}^2(\nu_0, \mu) + \mathrm{Hell}^2(\nu_0', \mu)).
\end{aligned}$$

An integration in μ eventually provides $J_{\Lambda,0}(\bar{\nu}) < \frac{1}{2}(J_{\Lambda,0}(\nu_0) + J_{\Lambda,0}(\nu_0'))$, which contradicts minimality of ν_0 and ν_0' simultaneously and provides uniqueness of the minimizer of $J_{\Lambda,0}$.

Let now $(\kappa_n)_n$ be any sequence converging to 0 and $(\kappa_{n'})_{n'}$ be any subsequence. By Proposition 3.2.4 there exists an additional subsequence $(\nu_{n''})_{n''}$ such that $\nu_{n''}$ converges weak* to a minimizer of $J_{\Lambda,0}$, hence it converges to ν_0 by uniqueness. Since every subsequence of $(\kappa_n)_n$ admits a subsequence converging to the same limit ν_0, we conclude that the whole sequence $(\nu_n)_n$ converges to ν_0. In turn, since any sequence $(\kappa_n)_n$ converging to 0 admits the same limit ν_0, the continuous limit as $\kappa \to 0$ follows. □

Remark 3.2.6 (Joint stability under changes of κ and Λ)**.** In the case when $\kappa_\infty \in (0, \infty)$ the behavior of the HK_κ-barycenter w.r.t. variations in κ can be reduced to the study of variations in Λ via Remark 3.1.1, by working in some finitely re-scaled Ω/κ, for a sufficiently small but finite κ instead, and by relocating the mass of Λ onto the re-scaled measures μ. By applying the results from Section 3.2.2 one then finds that one can consider joint limits in Λ and κ, and that the order in which the limits are taken does not matter.

The situation is more intricate when $\kappa_\infty \in \{0, \infty\}$. In the latter case, it can be problematic when Λ is not exclusively supported on measures of equal mass. In the former case one may obtain different cluster points of minimizers ν, depending on the order or relative speed in which Λ and κ approach their limits. An example is given in Remark 3.3.12 further below. Combining the above results we find that in both cases one obtains the limit minimizer for Λ_∞ and κ_∞ by first going to the limit in Λ and then in κ.

3.2.4 Duality

We now show that a dual problem for $(\mathcal{P}_{\Lambda,\kappa})$ can be formulated as

$$\sup\left\{\int_{\mathfrak{C}}\int_{\Omega}\Psi(\mu,x)\,\mathrm{d}\mu(x)\,\mathrm{d}\Lambda(\mu)\;\middle|\;\Psi,\Phi\in\mathrm{C}(\mathfrak{C}\times\Omega),(\Psi(\mu,\cdot),\Phi(\mu,\cdot))\in Q_\kappa\text{ for all }\mu\in\mathfrak{C},\right.$$
$$\left.\text{and }\int_{\mathfrak{C}}\Phi(\mu,y)\,\mathrm{d}\Lambda(\mu)\geq 0\text{ for all }y\in\Omega\right\}.\quad(\mathcal{D}_{\Lambda,\kappa})$$

We will study this duality in more detail in Section 3.3 (including dual existence and primal-dual optimality conditions) for the specific case when Λ is concentrated on the set of Dirac measures. For the general case we content ourselves with equality of optimal values. In [65] duality of $(\mathcal{P}_{\Lambda,\kappa})$ was established by combining all pairwise optimization problems (3.1.3) for $\mathrm{HK}^2_\kappa(\mu_i,\nu)$ in the discrete version of $(\mathcal{P}_{\Lambda,\kappa})$ (with a finite collection of input measures μ_i) and then dualizing them jointly. Here we generalize this combination to the case of uncountably many input measures.

Proposition 3.2.7. *Let $\Lambda\in\mathcal{P}(\mathfrak{C})$ and $\kappa\in(0,\infty)$. Then, $(\mathcal{D}_{\Lambda,\kappa})$ is a dual problem to $(\mathcal{P}_{\Lambda,\kappa})$, more precisely*

$$(\mathcal{D}_{\Lambda,\kappa})=(\mathcal{P}_{\Lambda,\kappa}).\tag{3.2.10}$$

Proof. For given $\Lambda\in\mathcal{P}(\mathfrak{C})$, define the measure $\Lambda\cdot\mu\in\mathcal{M}_+(\mathfrak{C}\times\Omega)$ as

$$\int_{\mathfrak{C}\times\Omega}\phi\,\mathrm{d}(\Lambda\cdot\mu):=\int_{\mathfrak{C}}\int_{\Omega}\phi(\mu,x)\,\mathrm{d}\mu(x)\,\mathrm{d}\Lambda(\mu)\quad\text{for }\phi\in\mathrm{C}(\mathfrak{C}\times\Omega).$$

We start with the primal problem and estimate

$$(\mathcal{P}_{\Lambda,\kappa})=\inf_{\nu\in\mathcal{M}_+(\Omega)}\int_{\mathfrak{C}}\mathrm{HK}^2_\kappa(\mu,\nu)\,\mathrm{d}\Lambda(\mu)\tag{3.2.11}$$

$$\overset{(3.1.3)}{=}\inf_{\nu\in\mathcal{M}_+(\Omega)}\int_{\mathfrak{C}}\left[\inf_{\substack{\gamma_1,\gamma_2,\gamma\in\mathcal{M}_+(\Omega^2),\gamma_i\ll\gamma\\ [(x,y)\mapsto x]_\#(\gamma_1)=\mu\\ [(x,y)\mapsto y]_\#(\gamma_2)=\nu}}\int_{\Omega\times\Omega}c_\kappa\left(x,\tfrac{\mathrm{d}\gamma_1}{\mathrm{d}\gamma}(x,y),y,\tfrac{\mathrm{d}\gamma_2}{\mathrm{d}\gamma}(x,y)\right)\mathrm{d}\gamma(x,y)\right]\mathrm{d}\Lambda(\mu)\tag{3.2.12}$$

$$\leq\inf_{\nu\in\mathcal{M}_+(\Omega)}\inf_{\substack{\Gamma_1,\Gamma_2,\Gamma\in\mathcal{M}_+(\mathfrak{C}\times\Omega^2),\Gamma_i\ll\Gamma\\ [(\mu,x,y)\mapsto(\mu,x)]_\#(\Gamma_1)=\Lambda\cdot\mu\\ [(\mu,x,y)\mapsto(\mu,y)]_\#(\Gamma_2)=\Lambda\otimes\nu}}\int_{\mathfrak{C}\times\Omega\times\Omega}c_\kappa\left(x,\tfrac{\mathrm{d}\Gamma_1}{\mathrm{d}\Gamma}(\mu,x,y),y,\tfrac{\mathrm{d}\Gamma_2}{\mathrm{d}\Gamma}(\mu,x,y)\right)\mathrm{d}\Gamma(\mu,x,y)\tag{3.2.13}$$

$$=\inf_{\substack{\Gamma_1,\Gamma_2,\Gamma\in\mathcal{M}_+(\mathfrak{C}\times\Omega^2),\Gamma_i\ll\Gamma\\ [(\mu,x,y)\mapsto(\mu,x)]_\#(\Gamma_1)=\Lambda\cdot\mu\\ \exists\nu\in\mathcal{M}_+(\Omega)\text{ s.t. }[(\mu,x,y)\mapsto(\mu,y)]_\#(\Gamma_2)=\Lambda\otimes\nu}}\int_{\mathfrak{C}\times\Omega\times\Omega}c_\kappa\left(x,\tfrac{\mathrm{d}\Gamma_0}{\mathrm{d}\Gamma}(\mu,x,y),y,\tfrac{\mathrm{d}\Gamma_1}{\mathrm{d}\Gamma}(\mu,x,y)\right)\mathrm{d}\Gamma(\mu,x,y).\tag{3.2.14}$$

The inequality from (3.2.12) to (3.2.13) follows since every admissible candidate in the latter induces a family of admissible candidates for the former. Indeed, let Γ_1,Γ_2,Γ be admissible in (3.2.13). By the constraints it follows that $[(\mu,x,y)\mapsto\mu]_\#(\Gamma_i)\ll\Lambda$. As in (3.1.3), since c_κ is positively 1-homogeneous in its second and fourth argument, the value of the integral does not depend on the choice of Γ, as long as $\Gamma_i\ll\Gamma$. Therefore, w.l.o.g. we may choose Γ such that

$[(\mu, x, y) \mapsto \mu]_\#(\Gamma) \ll \Lambda$. Let now $(\gamma_{1,\mu})_{\mu\in\mathfrak{C}}, (\gamma_{2,\mu})_{\mu\in\mathfrak{C}}$ and $(\gamma_\mu)_{\mu\in\mathfrak{C}}$ be the disintegrations of Γ_1, Γ_2 and Γ with respect to Λ. These three families of measures are then admissible in (3.2.12) and yield the same score.

Now set $X := \mathcal{M}(\mathfrak{C} \times \Omega^2)$, $Y := \mathcal{M}(\mathfrak{C} \times \Omega)$, and define

$$
\begin{aligned}
G\colon X \times X \to \mathbb{R} \cup \{\infty\}, \qquad & (\Gamma_1, \Gamma_2) \mapsto \int_{\mathfrak{C}\times\Omega^2} c_\kappa \left(x, \tfrac{\mathrm{d}\Gamma_1}{\mathrm{d}\Gamma}, y, \tfrac{\mathrm{d}\Gamma_2}{\mathrm{d}\Gamma}\right) \mathrm{d}\Gamma,\\
F_1\colon Y \to \mathbb{R} \cup \{\infty\}, \qquad & \tau \mapsto \begin{cases} 0 & \text{if } \tau = \Lambda \cdot \mu,\\ +\infty & \text{else,}\end{cases}\\
F_2\colon Y \to \mathbb{R} \cup \{\infty\}, \qquad & \tau \mapsto \begin{cases} 0 & \text{if } \tau = \Lambda \otimes \nu \text{ for some } \nu \in \mathcal{M}_+(\Omega),\\ +\infty & \text{else,}\end{cases}
\end{aligned}
$$

where in the definition of G the measure Γ is any positive measure such that $\Gamma_1 \ll \Gamma$ and $\Gamma_2 \ll \Gamma$ (note that $G(\Gamma_1, \Gamma_2)$ is finite only if both Γ_1 and Γ_2 are non-negative). Let us also define the two linear (projection) operators $Q_1, Q_2\colon X \to Y$ as

$$
Q_1\Gamma := [(\mu, x, y) \mapsto (\mu, x)]_\#(\Gamma) \quad \text{and} \quad Q_2\Gamma := [(\mu, x, y) \mapsto (\mu, y)]_\#(\Gamma).
$$

Hence, we can rewrite (3.2.14) as

$$
\inf_{\Gamma_1,\Gamma_2\in X} G(\Gamma_1, \Gamma_2) + F_1(Q_1\Gamma_1) + F_2(Q_2\Gamma_2). \tag{$\mathcal{P}$}
$$

By Fenchel–Rockafellar duality (Theorem 2.2.6) one finds $(\mathcal{P}) \leq (\mathcal{D})$, where $(\mathcal{D})$ is

$$
\sup_{\Psi,\Phi\in \mathrm{C}(\mathfrak{C}\times\Omega)} -G^*(Q_1^*\Psi, Q_2^*\Phi) - F_1^*(-\Psi) - F_2^*(-\Phi). \tag{$\mathcal{D}$}
$$

Note that we do not insist on a vanishing duality gap here. Direct computation quickly yields

$$
\begin{aligned}
F_1^* : \mathrm{C}(\mathfrak{C} \times \Omega) \to \mathbb{R} \cup \{\infty\}, \qquad & \Psi \mapsto \int_{\mathfrak{C}\times\Omega} \Psi \,\mathrm{d}(\Lambda \cdot \mu),\\
F_2^* : \mathrm{C}(\mathfrak{C} \times \Omega) \to \mathbb{R} \cup \{\infty\}, \qquad & \Phi \mapsto \begin{cases} 0 & \text{if } \int_{\mathfrak{C}} \Phi(\mu, y) \,\mathrm{d}\Lambda(\mu) \leq 0 \ \ \forall\, y \in \Omega,\\ +\infty & \text{else,}\end{cases}
\end{aligned}
$$

and using [42, Lemma 2.9]

$$
G^*(Q_1^*\cdot, Q_2^*\cdot) : \mathrm{C}(\mathfrak{C} \times \Omega)^2 \to \mathbb{R} \cup \{\infty\}, \quad (\Psi, \Phi) \mapsto \begin{cases} 0 & \text{if } (\Psi(\mu, \cdot), \Phi(\mu, \cdot)) \in Q_\kappa \text{ for all } \mu \in \mathfrak{C},\\ +\infty & \text{else.}\end{cases}
$$

With this, $(\mathcal{D})$ becomes

$$
\sup \left\{ \int_{\mathfrak{C}} \int_{\Omega} \Psi(\mu, x) \,\mathrm{d}\mu(x) \,\mathrm{d}\Lambda(\mu) \,\middle|\, \Psi, \Phi \in \mathrm{C}(\mathfrak{C} \times \Omega), (\Psi(\mu, \cdot), \Phi(\mu, \cdot)) \in Q_\kappa \text{ for all } \mu \in \mathfrak{C}, \right.
$$
$$
\left. \text{and } \int_{\mathfrak{C}} \Phi(\mu, y) \,\mathrm{d}\Lambda(\mu) \geq 0 \text{ for all } y \in \Omega \right\},
$$

which is exactly $(\mathcal{D}_{\Lambda,\kappa})$. Let us now fix a minimizer $\nu_\kappa \in \mathcal{M}_+(\Omega)$ of $J_{\Lambda,\kappa}$ and continue from above

$$(\mathcal{P}_{\Lambda,\kappa}) \leq (3.2.14) = (\mathcal{P}) \leq (\mathcal{D}) = (\mathcal{D}_{\Lambda,\kappa}) \tag{3.2.15}$$

$$= \sup_{\substack{\Psi,\Phi\in C(\mathfrak{C}\times\Omega)\\ (\Psi(\mu,\cdot),\Phi(\mu,\cdot))\in Q_\kappa\ \forall \mu\in\mathfrak{C}\\ \int_{\mathfrak{C}}\Phi(\mu,y)\,\mathrm{d}\Lambda(\mu)\geq 0\ \forall y\in\Omega}} \int_{\mathfrak{C}}\int_\Omega \Psi(\mu,x)\,\mathrm{d}\mu(x)\,\mathrm{d}\Lambda(\mu) \tag{3.2.16}$$

$$\leq \sup_{\substack{\Psi,\Phi\in C(\mathfrak{C}\times\Omega)\\ (\Psi(\mu,\cdot),\Phi(\mu,\cdot))\in Q_\kappa\ \forall \mu\in\mathfrak{C}\\ \int_{\mathfrak{C}}\Phi(\mu,y)\,\mathrm{d}\Lambda(\mu)\geq 0\ \forall y\in\Omega}} \int_{\mathfrak{C}}\int_\Omega \Psi(\mu,x)\,\mathrm{d}\mu(x)\,\mathrm{d}\Lambda(\mu) + \int_\Omega\int_{\mathfrak{C}} \Phi(\mu,y)\,\mathrm{d}\Lambda(\mu)\,\mathrm{d}\nu_\kappa(y) \tag{3.2.17}$$

$$\leq \sup_{\substack{\Psi,\Phi\in C(\mathfrak{C}\times\Omega)\\ (\Psi(\mu,\cdot),\Phi(\mu,\cdot))\in Q_\kappa\ \forall \mu\in\mathfrak{C}}} \int_{\mathfrak{C}}\left[\int_\Omega \Psi(\mu,x)\,\mathrm{d}\mu(x) + \int_\Omega \Phi(\mu,y)\,\mathrm{d}\nu_\kappa(y)\right]\mathrm{d}\Lambda(\mu) \tag{3.2.18}$$

$$\leq \int_{\mathfrak{C}}\left[\sup_{\substack{\psi,\phi\in C(\Omega)\\ (\psi,\phi)\in Q_\kappa}} \int_\Omega \psi(x)\,\mathrm{d}\mu(x) + \int_\Omega \phi(y)\,\mathrm{d}\nu_\kappa(y)\right]\mathrm{d}\Lambda(\mu) = \int_{\mathfrak{C}} \mathrm{HK}^2_\kappa(\mu,\nu_\kappa)\,\mathrm{d}\Lambda(\mu) = (\mathcal{P}_{\Lambda,\kappa}). \tag{3.2.19}$$

The chain of inequalities (3.2.11)-(3.2.19) is then actually a chain of equalities. Hence, $(\mathcal{D}_{\Lambda,\kappa})$ is a dual problem to $(\mathcal{P}_{\Lambda,\kappa})$ and the optimal values coincide. □

Remark 3.2.8 (Formal Wasserstein limit of $(\mathcal{D}_{\Lambda,\kappa})$)**.** Considering Theorem 3.1.3 and Proposition 3.2.4 one might expect to recover a dual problem for the Wasserstein-2 distance by considering $\kappa^2 \cdot (\mathcal{D}_{\Lambda,\kappa})$ and then sending $\kappa \to \infty$. The problem $\kappa^2 \cdot (\mathcal{D}_{\Lambda,\kappa})$ can be written as

$$\sup\left\{ \int_{\mathfrak{C}}\int_\Omega \Psi(\mu,x)\,\mathrm{d}\mu(x)\,\mathrm{d}\Lambda(\mu) \,\middle|\, \Psi,\Phi\in C(\mathfrak{C}\times\Omega), (\Psi(\mu,\cdot)/\kappa^2,\Phi(\mu,\cdot)/\kappa^2)\in Q_\kappa \text{ for all } \mu\in\mathfrak{C}, \right.$$
$$\left. \text{and } \int_{\mathfrak{C}}\Phi(\mu,y)\,\mathrm{d}\Lambda(\mu)\geq 0 \text{ for all } y\in\Omega\right\}.$$

At a purely intuitive level one can then consider the limit of the condition $(\Psi(\mu,\cdot)/\kappa^2,\Phi(\mu,\cdot)/\kappa^2)\in Q_\kappa$ for some $\mu\in\mathfrak{C}$ as $\kappa\to\infty$:

$$1-\Psi(\mu,x)/\kappa^2-\Phi(\mu,y)/\kappa^2+o(1/\kappa^2) = (1-\Psi(\mu,x)/\kappa^2)(1-\Phi(\mu,y)/\kappa^2)$$
$$\geq \mathrm{Cos}^2(|x-y|/\kappa) = 1-|x-y|^2/\kappa^2+o(1/\kappa^2).$$

That is, we expect to obtain the limit condition $\Psi(\mu,x)+\Phi(\mu,y)\leq|x-y|^2$, which would turn $(\mathcal{D}_{\Lambda,\kappa})$ into a version of the well-known dual problem for the Wasserstein barycenter. To the best of our knowledge this dual has so far not yet been stated in the literature for the case of a continuum of input measures.

3.3 Hellinger–Kantorovich barycenter of a continuum of Dirac measures

3.3.1 Problem setup and basic properties

Throughout Section 3.3 we study the particular case when Λ is concentrated on the set of unit Dirac measures, i.e. Λ-almost every μ is of the form δ_x for some $x \in \Omega$. In this case Λ can be represented by a measure $\rho \in \mathcal{P}(\Omega)$ which gives the distribution of the locations $x \in \Omega$. More precisely, for any $\rho \in \mathcal{P}(\Omega)$ we define the measure $\Lambda_\rho := T_\# \rho$ where $T : \Omega \to \mathcal{P}(\Omega)$, $x \mapsto \delta_x$, or equivalently

$$\int_{\mathfrak{C}} \phi(\mu) \,\mathrm{d}\Lambda_\rho(\mu) = \int_\Omega \phi(\delta_x) \,\mathrm{d}\rho(x) \quad \text{for all } \phi \in \mathrm{C}(\mathfrak{C}).$$

In this particular case the primal problem $(\mathcal{P}_{\Lambda_\rho,\kappa})$ simplifies to

$$\inf \left\{ J_{\rho,\kappa}(\nu) := \int_\Omega \mathrm{HK}^2_\kappa(\delta_x, \nu) \,\mathrm{d}\rho(x) \middle| \nu \in \mathcal{M}_+(\Omega) \right\}. \tag{$\mathcal{P}_{\rho,\kappa}$}$$

For the Wasserstein case (i.e. $\kappa = \infty$) this problem is trivial, the unique minimizer being given by $\nu = \delta_{\overline{x}}$ where $\overline{x} := \int_\Omega x \,\mathrm{d}\rho(x)$ is the center of mass of ρ ($\overline{x} \in \Omega$ by convexity of Ω). For ρ being a finite superposition of Dirac measures, i.e. $\rho = \sum_{i=1}^n m_i \,\delta_{x_i}$, and $\kappa \in (0, \infty)$ the problem was studied in [65]. It was shown that for κ sufficiently large the minimizer ν is again a single Dirac measure (consistent with the scaling limit of Proposition 3.2.4). However, for smaller κ, the minimizer ν may contain multiple Diracs or even be diffuse.

Therefore, we now study ($\mathcal{P}_{\rho,\kappa}$) in some more depth. First, we will further simplify the expression of ($\mathcal{P}_{\rho,\kappa}$) by making the expression $\mathrm{HK}^2_\kappa(\delta_x, \nu)$ more explicit. Then, in Section 3.3.2 we revisit the dual problem, derive dual existence and primal-dual optimality conditions. In Section 3.3.3 we present some results on whether barycenters ν are discrete or diffuse and we study the asymptotic behavior of the barycenter as $\kappa \to 0$ in Section 3.3.4.

Proposition 3.3.1. *Let $\kappa \in (0, \infty)$, $m > 0$, $\bar{x} \in \Omega$, $\nu \in \mathcal{M}_+(\Omega)$. One finds*

$$\begin{aligned} \mathrm{HK}^2_\kappa(m\delta_{\bar{x}}, \nu) &= \sup_{\xi < 1} \left[m\xi + \|\nu\| - \frac{1}{1-\xi} \int_\Omega \mathrm{Cos}^2(|\bar{x} - y|/\kappa) \,\mathrm{d}\nu(y) \right] \\ &= m + \|\nu\| - 2\sqrt{m} \sqrt{\int_\Omega \mathrm{Cos}^2(|\bar{x} - y|/\kappa) \,\mathrm{d}\nu(y)}. \end{aligned}$$

Proof. Let $\mu = m\,\delta_{\bar{x}}$. We recall from (3.1.4) that

$$\mathrm{HK}^2_\kappa(\mu, \nu) = \sup_{(\psi,\phi) \in Q_\kappa} \int_\Omega \psi(x) \,\mathrm{d}\mu(x) + \int_\Omega \phi(y) \,\mathrm{d}\nu(y) = \sup_{(\psi,\phi) \in Q_\kappa} m\psi(\bar{x}) + \int_\Omega \phi(y) \,\mathrm{d}\nu(y),$$

where

$$Q_\kappa = \left\{ (\psi, \phi) \in \mathrm{C}(\Omega)^2 \;\text{ s.t. } \begin{array}{l} \psi(x), \phi(y) \in (-\infty, 1] \\ (1 - \psi(x))(1 - \phi(y)) \geq \mathrm{Cos}(|x - y|/\kappa)^2 \end{array} \quad \forall x, y \in \Omega \right\}.$$

Note that only the value of ψ at $\bar{x}$ enters the energy. For any $\psi_{\bar{x}} \in \mathbb{R}$ and $n \in \mathbb{N}$ set $\psi_n(x) := \psi_{\bar{x}} - n \cdot \|x - \bar{x}\|$. For each ψ_n the remaining supremum over ϕ is then attained by

$$\phi_n(y) := \inf_{x \in \Omega} \; 1 - \frac{\mathrm{Cos}^2(|x - y|/\kappa)}{1 - \psi_n(x)},$$

which is indeed a continuous function in y. As $n \to \infty$ one has $\phi_n \nearrow \phi$ pointwise for

$$\phi(y) := 1 - \frac{\operatorname{Cos}^2(|\bar{x} - y|/\kappa)}{1 - \psi_{\bar{x}}}$$

i.e. only the constraint for $x = \bar{x}$ in Q_κ remains. Therefore, by monotone convergence, the problem reduces to

$$\begin{aligned}\operatorname{HK}^2_\kappa(m\delta_{\bar{x}}, \nu) &= \sup_{\psi_{\bar{x}} < 1} \left[m\psi_{\bar{x}} + \|\nu\| - \frac{1}{1 - \psi_{\bar{x}}} \int_\Omega \operatorname{Cos}^2(|\bar{x} - y|/\kappa) \, \mathrm{d}\nu(y) \right] \\ &= m + \|\nu\| - 2\sqrt{m} \sqrt{\int_\Omega \operatorname{Cos}^2(|\bar{x} - y|/\kappa) \, \mathrm{d}\nu(y)}. \end{aligned}$$

□

Corollary 3.3.2. *A primal minimizer for* $\operatorname{HK}^2_\kappa(m\delta_{\bar{x}}, \nu)$ *in* (3.1.1) *is given by*

$$\gamma = \delta_{\bar{x}} \otimes \sigma \qquad \text{with} \qquad \sigma = \nu \ \operatorname{Cos}^2(|\bar{x} - \cdot|/\kappa) \ \sqrt{\frac{m}{\|\nu \ \operatorname{Cos}^2(|\bar{x} - \cdot|/\kappa)\|}} \, . \tag{3.3.1}$$

Proof. This follows directly by plugging expression (3.3.1) into (3.1.1) and comparing the objective with Proposition 3.3.1. □

If we were to interpret an HK-barycenter ν as a "generalized clustering" of some input data ρ, then for each $\bar{x} \in \Omega$, the corresponding measure σ (with $m = 1$) could be interpreted as the association strength of the point at $\bar{x}$ with each of the points in the clustering. It would be a common occurrence that a point is associated with multiple 'clusters' at the same time.

Next, Proposition 3.3.1 also yields a simpler form of the primal objective which we will subsequently study in more detail.

Corollary 3.3.3. *Let* $\kappa \in (0, \infty)$ *and* $\rho \in \mathcal{P}(\Omega)$. *Then* $(\mathcal{P}_{\rho,\kappa})$ *admits a minimizer* $\nu \in \mathcal{M}_+(\Omega)$ *and the objective function* $J_{\rho,\kappa}$ *in* $(\mathcal{P}_{\rho,\kappa})$ *takes the form*

$$J_{\rho,\kappa}(\nu) = 1 + \|\nu\| - 2 \int_\Omega \sqrt{\int_\Omega \operatorname{Cos}^2(|x - y|/\kappa) \, \mathrm{d}\nu(y)} \, \mathrm{d}\rho(x). \tag{3.3.2}$$

Proof. Existence of a minimizer follows by Proposition 3.2.1 applied for $\Lambda = \Lambda_\rho$, the simplified objective in (3.3.2) follows by applying Proposition 3.3.1 for the integrand $\operatorname{HK}^2_\kappa(\delta_x, \nu)$ inside $J_{\rho,\kappa}$ in $(\mathcal{P}_{\rho,\kappa})$. □

3.3.2 Duality

Next, we prove that when $\Lambda = \Lambda_\rho$, the dual $(\mathcal{D}_{\Lambda,\kappa})$ takes the specific form

$$\begin{aligned}\sup \Bigg\{ \int_\Omega \psi(x) \, \mathrm{d}\rho(x) \;\Bigg|\; & \psi \in \mathrm{C}(\Omega), \, \psi < 1 \\ & \text{and } F_{\rho,\kappa}(\psi)(y) := \int_\Omega \frac{\operatorname{Cos}^2(|x - y|/\kappa)}{1 - \psi(x)} \, \mathrm{d}\rho(x) \leq 1 \text{ for all } y \in \Omega \Bigg\}. \end{aligned} \tag{$\mathcal{D}_{\rho,\kappa}$}$$

In the following, we will refer to $F_{\rho,\kappa}$ as the *constraint function*.

Proposition 3.3.4. *Let $\rho \in \mathcal{P}(\Omega)$ and $\kappa \in (0,\infty)$. Then,*

(i) $(\mathcal{D}_{\rho,\kappa})$ *is a dual problem to* $(\mathcal{P}_{\rho,\kappa})$ *and* $(\mathcal{D}_{\rho,\kappa}) = (\mathcal{P}_{\rho,\kappa})$,

(ii) $(\mathcal{D}_{\rho,\kappa})$ *admits a maximizer $\psi \in \mathrm{C}(\Omega)$ which is unique on the support of ρ and for any primal optimizer ν we have*

$$\psi(x) = 1 - \sqrt{\int_\Omega \mathrm{Cos}^2(|x-y|/\kappa)\,\mathrm{d}\nu(y)} \qquad \text{for } \rho\text{-a.e. } x \in \Omega, \tag{3.3.3a}$$

$$F_{\rho,\kappa}(\psi)(y) = 1 \qquad \text{for } \nu\text{-a.e. } y \in \Omega, \tag{3.3.3b}$$

(iii) an admissible couple $(\nu,\psi) \in \mathcal{M}_+(\Omega) \times \mathrm{C}(\Omega)$ is optimal if and only if (3.3.3) holds.

Proof. Step 1. Primal optimality conditions. Let $\nu \in \mathcal{M}_+(\Omega)$ be an optimizer of $(\mathcal{P}_{\rho,\kappa})$ and consider any non-negative measure $\tilde\nu \in \mathcal{M}_+(\Omega)$. By optimality of ν, one has

$$\frac{\mathrm{d}}{\mathrm{d}t} J_{\rho,\kappa}(\nu + t\tilde\nu)|_{t=0^+} \geq 0.$$

Using (3.3.2), we compute

$$0 \leq \frac{\mathrm{d}}{\mathrm{d}t} J_{\rho,\kappa}(\nu + t\tilde\nu)|_{t=0^+} = \frac{\mathrm{d}}{\mathrm{d}t}\left(1 + \|\nu\| + t\|\tilde\nu\| - 2\int_\Omega \sqrt{\int_\Omega \mathrm{Cos}^2(|x-y|/\kappa)\,\mathrm{d}(\nu+t\tilde\nu)(y)}\,\mathrm{d}\rho(x)\right)\Bigg|_{t=0^+}$$

$$= \int_\Omega \left[1 - \int_\Omega \frac{\mathrm{Cos}^2(|x-y|/\kappa)}{\sqrt{\int_\Omega \mathrm{Cos}^2(|x-z|/\kappa)\,\mathrm{d}\nu(z)}}\,\mathrm{d}\rho(x)\right] \mathrm{d}\tilde\nu(y).$$

Since $\tilde\nu$ is an arbitrary non-negative measure, we conclude that, for any optimal ν,

$$\int_\Omega \frac{\mathrm{Cos}^2(|x-y|/\kappa)}{\sqrt{\int_\Omega \mathrm{Cos}^2(|x-z|/\kappa)\,\mathrm{d}\nu(z)}}\,\mathrm{d}\rho(x) \leq 1 \quad \text{for all } y \in \Omega. \tag{3.3.4}$$

Now consider as variation $\tilde\nu = -\nu$ such that $\nu + t\tilde\nu \geq 0$ for $t \in [0,1]$. As above, we find

$$0 \leq \frac{\mathrm{d}}{\mathrm{d}t} J_{\rho,\kappa}(\nu + t\tilde\nu)|_{t=0^+} = -\int_\Omega \left[1 - \int_\Omega \frac{\mathrm{Cos}^2(|x-y|/\kappa)}{\sqrt{\int_\Omega \mathrm{Cos}^2(|x-z|/\kappa)\,\mathrm{d}\nu(z)}}\,\mathrm{d}\rho(x)\right] \mathrm{d}\nu(y).$$

Since from above we know that the expression in squared brackets must be non-negative, we now deduce that

$$\int_\Omega \frac{\mathrm{Cos}^2(|x-y|/\kappa)}{\sqrt{\int_\Omega \mathrm{Cos}^2(|x-z|/\kappa)\,\mathrm{d}\nu(z)}}\,\mathrm{d}\rho(x) = 1 \quad \text{for } \nu\text{-a.e. } y \in \Omega. \tag{3.3.5}$$

Step 2. Duality. By Proposition 3.2.7 as applied to Λ_ρ, we have $(\mathcal{P}_{\rho,\kappa}) = (\mathcal{P}_{\Lambda_\rho,\kappa}) = (\mathcal{D}_{\Lambda_\rho,\kappa})$. The latter problem, taking into account the definition of Λ_ρ, reduces to

$$\sup\left\{ \int_\Omega \Psi(\delta_x, x)\,\mathrm{d}\rho(x) \,\middle|\, \Psi, \Phi \in \mathrm{C}(\mathfrak{C}\times\Omega), (\Psi(\mu,\cdot), \Phi(\mu,\cdot)) \in Q_\kappa \text{ for all } \mu \in \mathfrak{C}, \right.$$
$$\left. \text{and } \int_\Omega \Phi(\delta_x, y)\,\mathrm{d}\rho(x) \geq 0 \text{ for all } y \in \Omega \right\} \quad (\mathcal{D}_{\Lambda_\rho,\kappa})$$

We now show that $(\mathcal{D}_{\Lambda_\rho,\kappa}) \leq (\mathcal{D}_{\rho,\kappa})$. Let $\Psi, \Phi \in \mathrm{C}(\mathfrak{C} \times \Omega)$ be any two admissible functions for $(\mathcal{D}_{\Lambda_\rho,\kappa})$. Define $\psi\colon \Omega \to \mathbb{R}$ as $\psi(x) := \Psi(\delta_x, x)$. Then $\psi \in \mathrm{C}(\Omega)$. Since $(\Psi(\mu,\cdot), \Phi(\mu,\cdot)) \in Q_\kappa$ for all $\mu \in \mathfrak{C}$, one has $\psi(x) = \Psi(\delta_x, x) < 1$ for all $x \in \Omega$ and in particular

$$(1 - \Psi(\delta_x, x))(1 - \Phi(\delta_x, y)) \geq \mathrm{Cos}(|x-y|/\kappa) \quad \text{for all } x, y \in \Omega,$$

so that

$$\frac{\mathrm{Cos}(|x-y|/\kappa)}{1-\psi(x)} \leq 1 - \Phi(\delta_x, y) \text{ for all } x, y \in \Omega.$$

Integrating against ρ and using that $\|\rho\| = 1$, we get

$$F_{\rho,\kappa}(\psi)(y) = \int_\Omega \frac{\mathrm{Cos}(|x-y|/\kappa)}{1-\psi(x)}\,\mathrm{d}\rho(x) \leq 1 - \int_\Omega \Phi(\delta_x, y)\,\mathrm{d}\rho(x) \leq 1 \quad \text{for all } y \in \Omega.$$

Hence ψ is admissible for $(\mathcal{D}_{\rho,\kappa})$ and $(\mathcal{D}_{\Lambda_\rho,\kappa}) \leq (\mathcal{D}_{\rho,\kappa})$. Let now $\psi \in \mathrm{C}(\Omega)$, $\psi < 1$, be admissible for $(\mathcal{D}_{\rho,\kappa})$ and let $\nu \in \mathcal{M}_+(\Omega)$ be any optimizer of $(\mathcal{P}_{\rho,\kappa})$. Then, since $F_{\rho,\kappa}(\psi) \leq 1$, one obtains

$$\begin{aligned}
\int_\Omega \psi(x)\,\mathrm{d}\rho(x) &\leq \int_\Omega \psi(x)\,\mathrm{d}\rho(x) + \int_\Omega \left[1 - \int_\Omega \frac{\mathrm{Cos}^2(|x-y|/\kappa)}{1-\psi(x)}\,\mathrm{d}\rho(x)\right]\mathrm{d}\nu(y) \\
&= \int_\Omega \left[\psi(x) + \|\nu\| - \frac{1}{1-\psi(x)}\int_\Omega \mathrm{Cos}^2(|x-y|/\kappa)\,\mathrm{d}\nu(y)\right]\mathrm{d}\rho(x) \\
&\leq \int_\Omega \sup_{\xi<1}\left[\xi + \|\nu\| - \frac{1}{1-\xi}\int_\Omega \mathrm{Cos}^2(|x-y|/\kappa)\,\mathrm{d}\nu(y)\right]\mathrm{d}\rho(x) \\
&= \int_\Omega \mathrm{HK}^2_\kappa(\delta_x, \nu)\,\mathrm{d}\rho(x) = (\mathcal{P}_{\rho,\kappa}),
\end{aligned} \tag{3.3.6}$$

where the first equality in the last line follows by Proposition 3.3.1. All in all, we showed $(\mathcal{P}_{\rho,\kappa}) = (\mathcal{P}_{\Lambda_\rho,\kappa}) = (\mathcal{D}_{\Lambda_\rho,\kappa}) \leq (\mathcal{D}_{\rho,\kappa}) \leq (\mathcal{P}_{\rho,\kappa})$, hence $(\mathcal{D}_{\rho,\kappa})$ is a dual to $(\mathcal{P}_{\rho,\kappa})$ and $(\mathcal{P}_{\rho,\kappa}) = (\mathcal{D}_{\rho,\kappa})$.

Step 3. Existence and characterization of the dual optimizer. Fix a primal optimizer ν_κ and define $\psi_\kappa \in \mathrm{C}(\Omega)$ as

$$(1 - \psi_\kappa(x))^2 = \int_\Omega \mathrm{Cos}^2(|x-y|/\kappa)\,\mathrm{d}\nu_\kappa(y). \tag{3.3.7}$$

By the optimality condition (3.3.4) we have $F_{\rho,\kappa}(\psi_\kappa) \leq 1$, so that ψ_κ is dual admissible. Furthermore, thanks to (3.3.5), we have $F_{\rho,\kappa}(\psi_\kappa) = 1$ ν_κ-almost everywhere. Thus, for $\nu = \nu_\kappa$ and $\psi = \psi_\kappa$, the chain of inequalities (3.3.6) becomes a chain of equalities and ψ_κ provides an optimizer for $(\mathcal{D}_{\rho,\kappa})$. In particular, from (3.3.6) we also deduce that

- any optimal ψ has to satisfy

  $$(1 - \psi(x))^2 = \int_\Omega \mathrm{Cos}^2(|x-y|/\kappa)\,\mathrm{d}\nu(y) \quad \text{for } \rho\text{-a.e. } x \in \Omega, \text{ for any primal optimizer } \nu,$$

  in order to have an equality between the second and third line (and recall that feasible ψ must be < 1),

- for any optimal ψ we have

  $$F_{\rho,\kappa}(\psi)(y) = 1 \text{ for } \nu\text{-a.e. } y \in \Omega, \text{ for any primal optimizer } \nu,$$

  in order to have an equality in the first line.

Hence, *(ii)* follows. Assume now $(\nu, \psi) \in \mathcal{M}_+(\Omega) \times \mathrm{C}(\Omega)$ is an admissible couple such that (3.3.3) holds. The same derivation as in (3.3.6) provides

$$\int_\Omega \psi(x)\,\mathrm{d}\rho(x) = \int_\Omega \mathrm{HK}^2_\kappa(\delta_x, \nu)\,\mathrm{d}\rho(x),$$

which holds if and only if ν is a primal minimizer and ψ is a dual maximizer, thus completing the proof. □

Corollary 3.3.5. *The constraint functions $F_{\rho,\kappa}(\psi)$ for any dual optimizer ψ are identical.*

Proof. This follows from the fact that all dual maximizers agree ρ-a.e., see (3.3.3a), and the constraint function only evaluates ψ on this set. □

Similar to the primal case (see Sections 3.2.2 and 3.2.3) dual maximizers and the constraint function are stable under small perturbations in ρ and κ.

Proposition 3.3.6 (Dual stability)**.** *Let $(\kappa_n)_n$ and $(\rho_n)_n$ be convergent (weak* in the latter case) sequences in $(0,\infty)$ and $\mathcal{P}(\Omega)$, with limits $\kappa_\infty \in (0,\infty)$ and $\rho_\infty \in \mathcal{P}(\Omega)$, respectively. Let $(\nu_n)_n$ be a corresponding sequence of primal minimizers and set*

$$\psi_n(x) := 1 - \sqrt{\int_\Omega \mathrm{Cos}^2(|x-y|/\kappa_n)\,\mathrm{d}\nu_n(y)}, \quad \psi_\infty(x) := 1 - \sqrt{\int_\Omega \mathrm{Cos}^2(|x-y|/\kappa_\infty)\,\mathrm{d}\nu_\infty(y)},$$

where ν_∞ is some cluster point of $(\nu_n)_n$. Then ψ_∞ is a dual maximizer of the limit problem for κ_∞ and ρ_∞, the sequence $(\psi_n)_n$ converges uniformly to ψ_∞ on the support of ρ_∞, and the sequence of constraint functions $(F_{\rho_n,\kappa_n}(\psi_n))_n$ converges uniformly to $F_{\rho_\infty,\kappa_\infty}(\psi_\infty)$ on Ω.

While the primal minimizer might not always be unique, the set where $F_{\rho,\kappa}(\psi) = 1$ for a dual maximizer ψ is unique and stable under small perturbations in ρ and κ. Since one must have $F_{\rho,\kappa}(\psi)(y) = 1$ for ν-almost all y, the set where $F_{\rho,\kappa}(\psi)$ is (close to) 1 therefore provides an alternative and unique interpretation of clustering.

Proof. By Corollary 3.2.3 the sequence $(\nu_n)_n$ is weak* pre-compact and any cluster point is a minimizer of the primal limit problem, see Remark 3.2.6 for the incorporation of a sequence of changing $(\kappa_n)_n$, with limit in $(0,\infty)$. By Proposition 3.3.4 a dual maximizer for the limit problem is then given through (3.3.3a), which gives dual optimality of ψ_∞.

Since the family of functions $(y \mapsto \mathrm{Cos}^2(|x-y|/\kappa_n))_{x\in\Omega,n}$ is uniformly equicontinuous, so are the $(\psi_n)_n$ and one has that the subsequence of functions $(\psi_{n_k})_k$ is uniformly convergent for every weak* convergent subsequence $(\nu_{n_k})_k$ of $(\nu_n)_n$ and each limit must be a dual maximizer. Since the limit dual maximizer is unique on the support of ρ_∞, all cluster points of $(\psi_n)_n$ must agree on this set (and no other cluster points can exist, e.g. since all cluster points $(\nu_n)_n$ are primal minimizers).

Finally, let us consider the sequence of constraint functions. For brevity set $F_n := F_{\rho_n,\kappa_n}(\psi_n)$. By assumption the sequence $(\kappa_n)_n$ is bounded away from zero, so there exists a finite set $Y \subset \Omega$ such that $\sum_{y\in Y} \mathrm{Cos}(|x-y|^2/\kappa_n) \geq 1$ for all $x \in \Omega$ and n. Then, with $F_n(y) \leq 1$ for all $y \in \Omega$, $n \in \mathbb{N}$ it follows that

$$\int_\Omega \frac{1}{1-\psi_n(x)}\mathrm{d}\rho_n(x) \leq \sum_{y\in Y} \int_\Omega \frac{\mathrm{Cos}(|x-y|^2/\kappa_n)}{1-\psi_n(x)}\mathrm{d}\rho_n(x) = \sum_{y\in Y} F_n(y) \leq |Y|.$$

Sequence $((1-\psi_n)^{-1}\cdot\rho_n)_n$ is therefore a sequence of bounded non-negative measures on Ω and thus weak* pre-compact. Again, by equicontinuity of the $(y\mapsto \mathrm{Cos}^2(|x-y|/\kappa_n))_{x,n}$ follows the pre-compactness of the sequence $(F_n)_n$ for the uniform convergence. Let now $(n_k)_k$ be a subsequence such that $((1-\psi_{n_k})^{-1}\cdot\rho_{n_k})_k$ converges weak*, let F_∞ be the limit of $(F_{n_k})_k$. We find that $F_{n_k}(y)$ converges pointwise to $F_{\rho_\infty,\kappa_\infty}(\psi_\infty)(y)$ for all y and thus we must have that $F_\infty = F_{\rho_\infty,\kappa_\infty}(\psi_\infty)$. Since the latter only depends on the value of ψ_∞ on the support of ρ_∞ and ψ_∞ is unique on this support, this means that all cluster points F_∞ must be identical. □

3.3.3 Discrete and diffuse barycenters

In [65] it was observed that the HK barycenter between a finite number of Dirac measures was sometimes discrete and sometimes diffuse. In the latter case it was shown that the solution is non-unique and that a discrete solution also exists [65, Proposition 6.2]. In this Section we give an alternative proof for this result. Then we turn to the question of existence of discrete solutions for a diffuse ρ and provide a negative answer: sometimes no discrete minimizers exist. For illustration we also briefly discuss an example on the torus.

Proposition 3.3.7 (Discrete barycenters for finite number of Dirac input measures)**.** *Let $\rho := \sum_{i=1}^n m_i\,\delta_{x_i}$ for $n\in\mathbb{N}$, $m=(m_1,\dots,m_n)\in\mathbb{R}_+^n$, $\sum_{i=1}^n m_i = 1$ and $x_1,\dots,x_n\in\Omega$. Then, $(\mathcal{P}_{\rho,\kappa})$ has a minimizer ν of the form*

$$\nu = \sum_{i=1}^{k} \tilde{m}_i\,\delta_{\tilde{x}_i}$$

for a positive integer $k\le n$, non-negative mass coefficients $\tilde{m}_1,\dots,\tilde{m}_k$ and positions $\tilde{x}_1,\dots,\tilde{x}_k\in\Omega$.

Proof. Let $\nu\in\mathcal{M}_+(\Omega)$ be a minimizer of $(\mathcal{P}_{\rho,\kappa})$ and let $\psi\in\mathrm{C}(\Omega)$ be the optimal dual defined via (3.3.3). Since ψ is uniquely determined on $\operatorname{spt}\rho$, we can reduce our focus to a vector $\psi=(\psi_1,\dots,\psi_n)\in\mathbb{R}^n$ with entries defined as

$$\psi_i = \int_\Omega \mathrm{Cos}^2(|x_i-y|/\kappa)\,\mathrm{d}\nu(y) \quad \text{for all } i=1,\dots,n.$$

Consider now a discrete approximating sequence $\{\nu^s\}_{s\in\mathbb{N}}$ for ν, $\operatorname{spt}\nu^s\subset\operatorname{spt}\nu$, with

$$\nu^s = \sum_{j=1}^{s}\bar{m}_j^s\delta_{x_j^s} \quad \text{for } \bar{m}^s=(\bar{m}_1^s,\dots,\bar{m}_s^s)\in\mathbb{R}_+^s,\ x_1^s,\dots,x_s^s\in\Omega,$$

such that $\nu^s \rightharpoonup^* \nu$ as $s\to\infty$. Each measure ν^s defines a vector $\psi^s\in\mathbb{R}^n$ by setting

$$\psi_i^s := \int_\Omega \mathrm{Cos}^2(|x_i-y|/\kappa)\,\mathrm{d}\nu^s(y) = \sum_{j=1}^{s}\mathrm{Cos}^2(|x_i-x_j^s|/\kappa)\,\bar{m}_j^s \quad \text{for all } i=1,\dots,n.$$

This can be written as $\psi^s = A^s\bar{m}^s$ for a matrix $A^s\in\mathbb{R}^{n\times s}$ with entries $A_{ij}^s := \mathrm{Cos}^2(|x_i-x_j^s|/\kappa)$. Clearly we have $\operatorname{rank}(A^s)\le n$, and so we can find a vector $\bar{m}^{s,n}\in\mathbb{R}_+^s$ with at most n strictly positive entries such that $\psi^s = A^s\cdot\bar{m}^{s,n}$. In turn, this defines a discrete non-negative measure $\nu^{s,n}$ supported on at most n points such that

$$\psi_i^s = \int_\Omega \mathrm{Cos}^2(|x_i-y|/\kappa)\,\mathrm{d}\nu^{s,n}(y) \quad \text{for all } i=1,\dots,n.$$

By compactness, there exists a cluster point $\nu^n \in \mathcal{M}_+(\Omega)$ such that, up to selection of a subsequence, $\nu^{s,n} \rightharpoonup^* \nu^n$ and ν^n is supported on at most n points because each measure $\nu^{s,n}$ is. Hence, using that $\nu^s \rightharpoonup^* \nu$ and $\nu^{s,n} \rightharpoonup^* \nu^n$ as $s \to \infty$ and that $A^s \bar{m}^s = A^s \bar{m}^{s,n}$, we obtain

$$\begin{aligned}\psi_i &= \int_\Omega \mathrm{Cos}^2(|x_i - y|/\kappa)\,\mathrm{d}\nu(y) = \lim_{s\to\infty} \int_\Omega \mathrm{Cos}^2(|x_i - y|/\kappa)\,\mathrm{d}\nu^s(y) \\ &= \lim_{s\to\infty} \int_\Omega \mathrm{Cos}^2(|x_i - y|/\kappa)\,\mathrm{d}\nu^{s,n}(y) = \int_\Omega \mathrm{Cos}^2(|x_i - y|/\kappa)\,\mathrm{d}\nu^n(y).\end{aligned}$$

Since $F_{\rho,\kappa}(\psi)(y) = 1$ for every $y \in \operatorname{spt}\nu$, we also have $F_{\rho,\kappa}(\psi)(x_j^s) = 1$ for every $j = 1, \dots, s$, and any $s > 0$. In particular, when passing to the limit as $s \to \infty$, one observes that $F_{\rho,\kappa}(\psi)(y) = 1$ for every $y \in \operatorname{spt}\nu^n$. By Proposition 3.3.4, point (iii), ν^n is primal optimal and the result follows. □

In this manuscript Ω is a subset of $\mathbb{R}^d$ and equipped with the Euclidean distance. The Hellinger–Kantorovich distance can be defined for non-negative measures over more general metric spaces [86] and in particular it can be shown that the barycenter problem between Dirac measures can be extended to the d-torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. One merely has to replace any occurrence of Ω by $\mathbb{T}^d$ and the Euclidean distance $|x - y|$ by the geodesic distance dst on $\mathbb{T}^d$. We now show that on the torus it may happen that no discrete minimizer ν exists.

Proposition 3.3.8 (Diffuse barycenters on the torus)**.** *Let $d \in \mathbb{N}$, $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$ be the d-dimensional unit torus (with circumference 1 along each dimension), equipped with its geodesic distance* dst *and let $\rho \in \mathcal{P}(\mathbb{T}^d)$ be the uniform probability measure on $\mathbb{T}^d$. Then for $d > 1$ there is no discrete barycenter ν. For $d = 1$ there is no discrete barycenter when $\kappa \cdot \pi$ is irrational or $\kappa \cdot \pi > 1$, otherwise discrete barycenters exist.*

Proof. Adapting Proposition 3.3.4 we obtain existence of a dual maximizer $\psi \in \mathrm{C}(\mathbb{T}^d)$, which is unique ρ-a.e., i.e. it is unique since ρ has full support. It is characterized by

$$(1 - \psi(x))^2 = \int_{\mathbb{T}^d} \mathrm{Cos}^2(\mathrm{dst}(x, y)/\kappa)\,\mathrm{d}\nu(y) \tag{3.3.8}$$

and the condition $\psi(x) < 1$, where ν is an arbitrary primal minimizer. By symmetry and convexity of the problem, ψ must be translation invariant, i.e. it must be constant, and therefore we have $\psi \in C^\infty(\mathbb{T}^d)$.

Assume now $d > 1$. Note that $g_y : x \mapsto \mathrm{Cos}^2(\mathrm{dst}(x, y)/\kappa)$ for some fixed $y \in \mathbb{T}^d$ is merely C^1 when $\kappa\pi/2 \le 1/2$ and even only C^0 otherwise, due to its behavior on the sphere of radius $\kappa\pi/2$ around y or on the cut locus of $\mathrm{dst}(\cdot, y)$. Also, it is not possible to "cancel" these irregularities by carefully combining a countable number of g_y for different y with positive weights that have a finite sum. Therefore, ψ cannot be constructed from a discrete ν via (3.3.8).

Now let $d = 1$. For $\kappa\pi > 1$ the function g_y is merely C^0 and a finite number of g_y cannot be combined into a smoother function and thus, as above, no discrete barycenter can exist. Let now $\kappa\pi \le 1$ and in addition $\kappa\pi \in \mathbb{Q}$. Then for any $y \in \mathbb{T}^1$ the set

$$S_y := \{y + k \cdot \kappa\pi \,|\, k \in \mathbb{Z}\}$$

with the obvious interpretation of addition on the torus and identification of points that differ by an integer, is finite. Then by setting $\nu := m \sum_{y' \in S_y} \delta_{y'}$ with a suitable $m > 0$ (depending on κ), one will find that it is possible to construct a constant ψ via (3.3.8) and hence this ν is a discrete

primal minimizer. For $\kappa\pi \notin \mathbb{Q}$ this construction fails since S_y will not be finite and therefore no countable number of g_y for different y with positive weights that have a finite sum yields a C^∞-function. □

From this example we draw the following intuition for $\mathbb{R}^d$: When ρ has a large (compared to κ) region of constant density, the question whether the barycenter ν can be discrete or diffuse it not decided in the bulk of the region but at its boundary. Therefore, by careful design of the boundary region it might be possible to construct ρ for which no discrete barycenter exists. We confirm this in the next proposition.

Proposition 3.3.9 (Diffuse barycenters in $\mathbb{R}^d$)**.** *Let $d \in \mathbb{N}$ and $\kappa \in (0, \infty)$. There exist a compact, closed, convex set $\Omega \subset \mathbb{R}^d$ and a measure $\rho \in \mathcal{M}_+(\Omega)$ such that $(\mathcal{P}_{\rho,\kappa})$ has no discrete optimizer.*

Proof. Let $L \in (0, \infty)$ and let $\Omega := \bar{B}(0, L + \kappa\pi/2) \subset \mathbb{R}^d$ be the closed ball centered at the origin with radius $L + \kappa\pi/2$. Define the function $\sigma \in \mathrm{C}(\Omega)$ as

$$\sigma^2(x) := \int_{B(0,L)} \mathrm{Cos}^2(|x - y|/\kappa)\,\mathrm{d}y \quad \text{for all } x \in \Omega.$$

Denote $C_d := \|\mathrm{Cos}^2(|\cdot|)\|_{L^1(\mathbb{R}^d)}$ and $a := 1/\|\sigma\|_{L^1(\Omega)}$, and consider the probability measure $\rho := a \cdot \sigma \cdot \mathcal{L}^d$. Let $(\nu, \psi) \in \mathcal{M}_+(\Omega) \times \mathrm{C}(\Omega)$ be defined as

$$\nu := C_d^2 \kappa^{2d} a^2 \cdot \mathcal{L}^d \llcorner B(0, L) \quad \text{and} \quad \psi := 1 - C_d \kappa^d \cdot a \cdot \sigma.$$

The pair $(\nu, \psi) \in \mathcal{M}_+(\Omega) \times \mathrm{C}(\Omega)$ is an optimal primal-dual pair for $(\mathcal{P}_{\rho,\kappa})$ and $(\mathcal{D}_{\rho,\kappa})$. Indeed, one readily checks that

$$F_{\rho,\kappa}(\psi)(y) = \int_\Omega \frac{\mathrm{Cos}^2(|x - y|/\kappa)}{1 - \psi(x)} \mathrm{d}\rho(x) = \frac{1}{C_d \kappa^d} \int_\Omega \mathrm{Cos}^2(|x - y|/\kappa)\,\mathrm{d}x \leq 1 \quad \text{for all } y \in \Omega$$

and $F_{\rho,\kappa}(\psi)(y) = 1$ for all $y \in \bar{B}(0, L) = \operatorname{spt} \nu$. Further, by construction,

$$(1 - \psi(x))^2 = C_d^2 \kappa^{2d} a^2 \sigma^2(x) = C_d^2 \kappa^{2d} a^2 \int_{B(0,L)} \mathrm{Cos}^2(|x - y|/\kappa)\,\mathrm{d}y = \int_\Omega \mathrm{Cos}^2(|x - y|/\kappa)\,\mathrm{d}\nu(x).$$

Hence, optimality of ν and ψ follows from Proposition 3.3.4, point (iii). Since ψ is unique on $\operatorname{spt} \rho$, the function $F_{\rho,\kappa}(\psi)$ is unique and identical for all dual maximizers, therefore the set $\{y \in \Omega | F_{\rho,\kappa}(\psi)(y) = 1\} = \bar{B}(0, L)$ is unique, and finally we find that any primal minimizer must be concentrated on $\bar{B}(0, L)$.

Denote by $\mathrm{Cos}^2_\kappa : \mathbb{R}^d \to [0, 1]$ the function $x \mapsto \mathrm{Cos}^2(|x|/\kappa)$. Extend the measure ν from Ω to $\mathbb{R}^d$ by zero. Then by the above, one obtains for the convolution

$$(\mathrm{Cos}^2_\kappa * \nu)(y) := \int_{\mathbb{R}^d} \mathrm{Cos}^2_\kappa(y - x)\,\mathrm{d}\nu(x) = \begin{cases} \sigma^2(y) & \text{for } y \in \Omega, \\ 0 & \text{else,} \end{cases}$$

that is, it is known on all of $\mathbb{R}^d$. Now the proof strategy is to show that since the convolution of ν with a compact kernel is fully known, ν must indeed be uniquely determined and be equal to the above, and hence no other primal minimizer exists, in particular none that is discrete.

Denote by $\mathcal{F}$ the Fourier transform on $\mathbb{R}^d$, acting on suitable functions f and measures μ as

$$(\mathcal{F}f)(k) := \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} f(x)\,\exp(ikx)\,\mathrm{d}x, \qquad (\mathcal{F}\mu)(k) := \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\exp(ikx)\,\mathrm{d}\mu(x)$$

whenever these integrals are well-defined. Since $(\mathrm{Cos}_\kappa^2 * \nu) \in L^2(\mathbb{R}^d)$, $\mathcal{F}(\mathrm{Cos}_\kappa^2 * \nu)$ is well-defined. The convolution theorem now corresponds to the observation that for almost every k one has

$$\begin{aligned}(\mathcal{F}\,\mathrm{Cos}_\kappa^2 * \nu)(k) &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}(\mathrm{Cos}_\kappa^2 * \nu)(x)\exp(ikx)\mathrm{d}x\\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mathrm{Cos}_\kappa^2(x-y)\mathrm{d}\nu(y)\exp(ikx)\mathrm{d}x = \int_{\mathbb{R}^d}(\mathcal{F}\,\mathrm{Cos}_\kappa^2)(k)\exp(iky)\mathrm{d}\nu(y)\\ &= (2\pi)^d(\mathcal{F}\,\mathrm{Cos}_\kappa^2)(k)\,(\mathcal{F}\nu)(k),\end{aligned}$$

where we swapped the order of integration by Fubini's theorem.

Since Cos_κ^2 has compact support, $\mathcal{F}\,\mathrm{Cos}_\kappa^2(k) \neq 0$ k-almost everywhere and thus we obtain that

$$(\mathcal{F}\nu)(k) = \frac{(\mathcal{F}\,\mathrm{Cos}_\kappa^2 * \nu)(k)}{(2\pi)^d(\mathcal{F}\,\mathrm{Cos}_\kappa^2)(k)}$$

for almost all k, i.e. the Fourier transform of all primal minimizers must agree almost everywhere.

We now show that for a finite measure μ on $\mathbb{R}^d$ with compact support one finds $[\mathcal{F}\mu(k) = 0\ k\text{-a.e.}] \Rightarrow [\mu = 0]$ and thus by linearity of $\mathcal{F}$ this implies that knowing $(\mathcal{F}\nu)$ k-almost everywhere uniquely determines ν. Let g be a continuous convolution kernel with compact support and total mass 1, and for $\varepsilon > 0$ let $g_\varepsilon(x) := \varepsilon^{-d}\cdot g(x/\varepsilon)$ be the re-scaled version. Then clearly $g_\varepsilon * \mu \rightharpoonup^* \mu$ as $\varepsilon \to 0$. Let $\varphi \in \mathrm{C}_c(\mathbb{R}^d)$ be continuous with compact support. Then one finds

$$\begin{aligned}\int_{\mathbb{R}^d}\varphi\,\mathrm{d}\mu &= \lim_{\varepsilon\searrow 0}\int_{\mathbb{R}^d}\varphi(x)\,(g_\varepsilon * \mu)(x)\,\mathrm{d}x = \lim_{\varepsilon\searrow 0}(2\pi)^d\int_{\mathbb{R}^d}\overline{(\mathcal{F}\varphi)(k)}\,(\mathcal{F}(g_\varepsilon * \mu))(k)\,\mathrm{d}k\\ &= \lim_{\varepsilon\searrow 0}(2\pi)^{2d}\int_{\mathbb{R}^d}\overline{(\mathcal{F}\varphi)(k)}\,(\mathcal{F}g_\varepsilon)(k)(\mathcal{F}\mu)(k)\,\mathrm{d}k = 0,\end{aligned}$$

where we first used unitarity (up to normalization) of the Fourier transform on $L^2(\mathbb{R}^d, \mathbb{C})$ and that $\varphi, (g_\varepsilon * \mu) \in L^2(\mathbb{R}^d, \mathbb{C})$ for all $\varepsilon > 0$, then again the convolution theorem as above, and finally the assumption $\mathcal{F}\mu(k) = 0$ for almost all k. Since this holds for all $\varphi \in \mathrm{C}_c(\mathbb{R}^d)$, we must have that $\mu = 0$.

In conclusion, the minimizer ν constructed above must be unique and therefore no discrete minimizers can exist.

□

Note that the above argument with the convolution only works since $\operatorname{spt}\rho \supset B(0, \kappa\pi/2) + \operatorname{spt}\nu$ (where the plus denotes the Minkowski sum). In other cases, the convolution $\mathrm{Cos}_\kappa^2 * \nu$ may not be fully known and consequently ν may be non-unique.

3.3.4 Asymptotic behavior for $\kappa \to 0$

Now we look more closely at the limiting behavior of the functional as $\kappa \to 0$ (the case $\kappa \to \infty$ is given by the Wasserstein limit and well-understood). We start by specifying the unique minimizer ν_κ for $\kappa = 0$.

Proposition 3.3.10. *Let $\rho \in \mathcal{P}(\Omega)$ and consider the decomposition*

$$\rho = \rho_c + \sum_{i=1}^{\infty} m_i \delta_{x_i}, \quad x_i \in \Omega, m_i \geq 0 \text{ for all } i \geq 1, \tag{3.3.9}$$

with $\rho_c \in \mathcal{M}_+(\Omega)$ atomless. Then,

$$\nu = \sum_{i=1}^{\infty} m_i^2 \delta_{x_i} \quad \text{is the unique optimizier of} \quad \inf \left\{ \int_\Omega \mathrm{Hell}^2(\delta_x, \nu) \, \mathrm{d}\rho(x) \middle| \nu \in \mathcal{M}_+(\Omega) \right\}.$$

Proof. Taking into account (3.3.9), for any $\nu \in \mathcal{M}_+(\Omega)$, we can write

$$\int_\Omega \mathrm{Hell}^2(\delta_x, \nu) \, \mathrm{d}\rho(x) = \sum_{i=1}^{\infty} m_i \, \mathrm{Hell}^2(\delta_{x_i}, \nu) + \int_\Omega \mathrm{Hell}^2(\delta_x, \nu) \, \mathrm{d}\rho_c(x).$$

For the first term, with (3.1.6) observe that for any $x \in \Omega$,

$$\mathrm{Hell}^2(\delta_x, \nu) = (1 - \sqrt{\nu(\{x\})})^2 + \nu(\Omega \setminus \{x\}) = 1 - 2\sqrt{\nu(\{x\})} + \|\nu\|.$$

For the second term, let $m = \sum_i m_i \in [0, 1]$ be the total mass of the atomic part of ρ. Observe now that, since ρ_c is atomless, we have $\nu(\{x\}) = 0$ for ρ_c-a.e. $x \in \Omega$, so that the second term in the sum above simplifies into

$$\int_\Omega \mathrm{Hell}^2(\delta_x, \nu) \, \mathrm{d}\rho_c(x) = \|\rho_c\|(1 + \|\nu\|) = (1 - m)\,(1 + \|\nu\|).$$

Therefore, one obtains

$$\int_\Omega \mathrm{Hell}^2(\delta_x, \nu) \, \mathrm{d}\rho(x) = 1 + \|\nu\| - 2\sum_{i=1}^{\infty} m_i \sqrt{\nu(\{x_i\})}.$$

Hence, any optimal ν must be supported on $\{x_i\}_{i=1}^{\infty}$, and the Hellinger barycenter problem for ρ reduces to

$$\inf \left\{ 1 + \sum_{i=1}^{\infty} n_i - 2\sum_{i=1}^{\infty} m_i \sqrt{n_i} \middle| \nu = \sum_{j=1}^{\infty} n_j \delta_{x_j}, n_j \geq 0 \right\}.$$

The result follows by first order optimality conditions for each n_i. □

For some $\rho \in \mathcal{P}(\Omega)$ and $\kappa \in [0, \infty)$ let now ν_κ be a primal optimizer of $(\mathcal{P}_{\rho,\kappa})$. Then, by Corollary 3.2.5, as $\kappa \to 0$, ν_k converges to the unique minimizer ν_0 of the Hellinger barycenter problem for ρ which is specified by Proposition 3.3.10. We find that the only contributions to ν_0 arise from the atoms of ρ, all other contributions must tend to 0 as $\kappa \to 0$. The following Lemma provides a rough estimate on the corresponding rate. It is related to the concentration of ρ.

Lemma 3.3.11. *Let $\rho \in \mathcal{P}(\Omega)$. For $\kappa \in (0, \infty)$, let $\nu_\kappa \in \mathcal{M}_+(\Omega)$ be a minimizer of $(\mathcal{P}_{\rho,\kappa})$. Denote*

$$C_{\rho,\kappa} := \sup_{y \in \Omega} \left[\rho(B(y, \kappa \cdot \pi/2)) \right].$$

Then,

$$\|\nu_k\| \leq 4 C_{\rho,\kappa}. \tag{3.3.10}$$

Proof. Via reverse Jensen's inequality, we have

$$\int_\Omega \sqrt{\int_\Omega \mathrm{Cos}^2(|x-y|/\kappa)\,\mathrm{d}\nu_\kappa(y)}\,\mathrm{d}\rho(x) \le \sqrt{\int_\Omega\int_\Omega \mathrm{Cos}^2(|x-y|/\kappa)\,\mathrm{d}\nu_\kappa(y)\mathrm{d}\rho(x)} \le \sqrt{C_{\rho,\kappa}\|\nu_\kappa\|}.$$

Taking into account that the zero measure provides an upper bound for the optimal value, the inequality above provides

$$1 = J_{\rho,\kappa}(0) \ge J_{\rho,\kappa}(\nu_\kappa) \ge 1 + \|\nu_k\| - 2\sqrt{C_{\rho,\kappa}}\sqrt{\|\nu_\kappa\|},$$

so that $\|\nu_k\| \le 4C_{\rho,\kappa}$, which establishes (3.3.10). □

If ρ contains atoms, then $\lim_{\kappa\searrow 0} C_{\rho,\kappa} > 0$ and $\|\nu_0\| > 0$, in agreement with Proposition 3.3.10. If ρ is atomless, then $\lim_{\kappa\searrow 0} C_{\rho,\kappa} = 0$ (by outer regularity of Radon measures). The rate will depend on ρ, and will be slower, for instance, when ρ is concentrated on a lower-dimensional submanifold. When $\lim_{\kappa\searrow 0} C_{\rho,\kappa} = 0$ we also obtain $\psi_\kappa \to 1$ uniformly for dual maximizers via (3.3.3a).

Remark 3.3.12 (Different limits as $\kappa \to \infty$)**.** We briefly resume the discussion of Remark 3.2.6. Let $(\kappa_n)_n$ be a positive sequence, converging to $\kappa_\infty = 0$, let x be in the interior of Ω, and let ρ_n be convolutions of δ_x with some compact mollifier, with width going to zero as $n \to \infty$, such that $\rho_n \overset{*}{\to} \rho_\infty := \delta_x$. Then the minimizer of J_{ρ_n,κ_∞} will be $\nu = 0$ for all $n < \infty$, whereas it will be $\nu = \delta_x$ for J_{ρ_∞,κ_n} for all n up to $n = \infty$.

Finally, by assuming that ρ has an L^2-density with respect to the Lebesgue measure, we will now provide a more precise statement on the asymptotic behavior of ν_κ and ψ_κ.

Proposition 3.3.13. *Let $\rho \in \mathcal{P}(\Omega)$ and assume $\rho \ll \mathcal{L}^d \llcorner \Omega$ with $\mathrm{d}\rho/\mathrm{d}\mathcal{L}^d \in L^2(\Omega)$. Denote by $C_d := \|\mathrm{Cos}^2(|\cdot|)\|_{L^1(\mathbb{R}^d)}$. For any $\kappa \in (0,\infty)$, let $(\nu_\kappa, \psi_\kappa) \in \mathcal{M}_+(\Omega) \times \mathrm{C}(\Omega)$ be an optimal pair for $(\mathcal{P}_{\rho,\kappa})$ and $(\mathcal{D}_{\rho,\kappa})$. Then, $\nu_\kappa \rightharpoonup^* 0$ as $\kappa \to 0$ with*

$$\|\nu_k\| \le 4C_d \left\|\tfrac{\mathrm{d}\rho}{\mathrm{d}\mathcal{L}^d}\right\|^2_{L^2(\Omega)} \cdot \kappa^d. \tag{3.3.11}$$

In particular, $\nu_\kappa/(C_d\kappa^d) \rightharpoonup^* \left(\frac{\mathrm{d}\rho}{\mathrm{d}\mathcal{L}^d}\right)^2 \cdot \mathcal{L}^d$ and $\|1-\psi_\kappa\|_\infty = O(\kappa^{d/2})$ as $\kappa \to 0$.

Proof. Step 0. Preliminaries on mollifiers. For each $\kappa > 0$, consider the continuous function $\eta_\kappa \colon \mathbb{R}^d \to \mathbb{R}$ defined as

$$\eta_\kappa(x) := \frac{\mathrm{Cos}^2(|x|/\kappa)}{C_d\kappa^d}.$$

The collection $(\eta_\kappa)_{\kappa>0}$ provides a family of compactly supported continuous mollifiers, which are also positive and radially symmetric. For any measure $\nu \in \mathcal{M}(\mathbb{R}^d)$, we define as usual

$$(\eta_\kappa * \nu)(x) := \int_{\mathbb{R}^d} \eta_\kappa(x-y)\,\mathrm{d}\nu(y) \quad \text{for } x \in \mathbb{R}^d$$

and extend such definition to functions $f \in L^p_{\mathrm{loc}}(\mathbb{R}^d)$, $p \ge 1$, setting $\eta_\kappa * f := \eta_\kappa * (f\mathcal{L}^d)$. We recall from [59, Appendix C, Theorem 6] the following classical results:

- Let $f \in L^p(\mathbb{R}^d)$, $p \ge 1$. Then, $\eta_\kappa * f \in \mathrm{C}(\mathbb{R}^d)$ and $\eta_\kappa * f \to f$ in $L^p(\mathbb{R}^d)$ as $\kappa \to 0$.

- Let $f \in \mathrm{C}(\mathbb{R}^d)$. Then, $\eta_\kappa * f \to f$ uniformly on compact subsets of $\mathbb{R}^d$ as $\kappa \to 0$.

From now on, for simplicity, we denote by ρ also the Lebesgue density of ρ, so that $\rho \in L^2(\Omega)$. We extend ρ to $L^2(\mathbb{R}^d)$ by assigning the value 0 outside of Ω.

Step 1. Mass bound and dual convergence. Let $\nu_\kappa \in \mathcal{M}_+(\Omega)$ be a minimizer of $J_{\rho,\kappa}$. Lemma 3.3.11 combined with the absolute continuity of ρ with respect to $\mathcal{L}^d$ provides $\|\nu_k\| \leq \mathcal{L}^d(B(0,\pi/2))\|\rho\|^2_{L^2(\Omega)} \cdot \kappa^{d/2}$, which shows a decay rate which is slower than the stated rate. Hence, the estimate of Lemma 3.3.11 has to be refined. By minimality of ν_κ we have $J_{\rho,\kappa}(\nu_\kappa) \leq J_{\rho,\kappa}(0) = 1$ and so, by the Cauchy–Schwartz inequality in $L^2(\Omega)$, we obtain

$$\begin{aligned} 1 = J_{\rho,\kappa}(0) \geq J_{\rho,\kappa}(\nu_\kappa) &= 1 + \|\nu_k\| - 2\langle \sqrt{C_d\kappa^d(\eta_\kappa * \nu_\kappa)}, \rho\rangle_{L^2(\Omega)} \\ &\geq 1 + \|\nu_k\| - 2\left\| \sqrt{C_d\kappa^d(\eta_\kappa * \nu_\kappa)} \right\|_{L^2(\Omega)} \|\rho\|_{L^2(\Omega)} \geq 1 + \|\nu_k\| - 2\sqrt{C_d\kappa^d}\sqrt{\|\nu_\kappa\|}\|\rho\|_{L^2(\Omega)}, \end{aligned}$$

where we used $\|\eta_\kappa * \nu_\kappa\|_{L^1(\Omega)} = \|\nu_\kappa\|$, so that $\|\nu_k\| \leq 4C_d\|\rho\|^2_{L^2(\Omega)} \cdot \kappa^d$, which establishes (3.3.11). In particular, thanks to (3.3.3), we also have

$$(1 - \psi_\kappa(x))^2 = \int_\Omega \mathrm{Cos}^2(|x-y|/\kappa)\, \mathrm{d}\nu_\kappa(y) \leq \|\nu_\kappa\|$$

and by recalling that $\psi_\kappa \leq 1$, see ($\mathcal{D}_{\rho,\kappa}$), and with (3.3.11) this establishes $\|1 - \psi_\kappa\|_\infty = O(\kappa^{d/2})$ as $\kappa \to 0$.

Step 2. Energy bound. Consider $\hat{\nu}_\kappa = C_d\kappa^d\rho^2\mathcal{L}^d \in \mathcal{M}_+(\Omega)$. One finds

$$J_{\rho,\kappa}(\hat{\nu}_\kappa) = 1 + C_d\kappa^d \left(\|\rho\|^2_{L^2(\Omega)} - 2\langle \sqrt{\eta_\kappa * \rho^2}, \rho\rangle_{L^2(\Omega)} \right). \tag{3.3.12}$$

By strong convergence of the mollified functions, we have $\eta_\kappa * \rho^2 \to \rho^2$ in $L^1(\mathbb{R}^d)$ as $\kappa \to 0$. Using that $(a-b)^2 \leq |a^2 - b^2|$ for any $a, b \geq 0$, we eventually obtain that

$$\sqrt{\eta_\kappa * \rho^2} \to \rho \text{ in } L^2(\mathbb{R}^d) \text{ as } \kappa \to 0, \quad \text{which implies} \quad \langle \sqrt{\eta_\kappa * \rho^2}, \rho\rangle_{L^2(\Omega)} = \|\rho\|^2_{L^2(\Omega)} + o(1).$$

Thus, substituting this expansion into (3.3.12), one gets

$$\min_{\nu \in \mathcal{M}_+(\Omega)} J_{\rho,\kappa}(\nu) \leq J_{\rho,\kappa}(\hat{\nu}_\kappa) = 1 - C_d\kappa^d\|\rho\|^2_{L^2(\Omega)} + o(\kappa^d). \tag{3.3.13}$$

Step 3. Convergence of the rescaled minimizers. Let $\nu_\kappa \in \mathcal{M}_+(\Omega)$ be a minimizer of $J_{\rho,\kappa}$. Using (3.3.13) we estimate

$$\begin{aligned} 0 = \frac{J_{\rho,\kappa}(\nu_\kappa) - \min_\nu J_{\rho,\kappa}(\nu)}{C_d\kappa^d} &\geq \frac{\|\nu_\kappa\| - 2\langle \sqrt{C_d\kappa^d(\eta_\kappa * \nu_\kappa)}, \rho\rangle_{L^2(\Omega)} + C_d\kappa^d\|\rho\|^2_{L^2(\Omega)} + o(\kappa^d)}{C_d\kappa^d} \\ &= \frac{\left\|\sqrt{\eta_\kappa * \nu_\kappa}\right\|^2_{L^2(\mathbb{R}^d)}}{C_d\kappa^d} - 2\left\langle \sqrt{\frac{\eta_\kappa * \nu_\kappa}{C_d\kappa^d}}, \rho \right\rangle_{L^2(\mathbb{R}^d)} + \|\rho\|^2_{L^2(\mathbb{R}^d)} + o(1) \\ &= \left\| \sqrt{\frac{\eta_\kappa * \nu_\kappa}{C_d\kappa^d}} - \rho \right\|^2_{L^2(\mathbb{R}^d)} + o(1), \end{aligned}$$

where we used again $\|\nu_\kappa\| = \|\eta_\kappa * \nu_\kappa\|_{L^1(\Omega)}$ from the first to the second line. Therefore, $\sqrt{(\eta_\kappa * \nu_\kappa)/(C_d\kappa^d)} \to \rho$ in $L^2(\mathbb{R}^d)$ as $\kappa \to 0$. In particular,

$$\frac{\eta_\kappa * \nu_k}{C_d\kappa^d} \to \rho^2 \text{ in } L^1(\mathbb{R}^d) \text{ as } \kappa \to 0.$$

Fix any $\phi \in \mathrm{C}(\Omega)$ and consider a continuous bounded extension $\tilde{\phi} \in \mathrm{C}(\mathbb{R}^d)$ such that $\tilde{\phi}\llcorner\Omega = \phi$. We compute

$$\begin{aligned}\int_\Omega \phi \,\mathrm{d}\left(\frac{\nu_k}{C_d\kappa^d}\right) &= \int_{\mathbb{R}^d} \tilde{\phi}\,\mathrm{d}\left(\frac{\eta_\kappa * \nu_k}{C_d\kappa^d}\mathcal{L}^d\right) + \int_{\mathbb{R}^d} \tilde{\phi}\,\mathrm{d}\left(\frac{\nu_k}{C_d\kappa^d} - \frac{\eta_\kappa * \nu_k}{C_d\kappa^d}\mathcal{L}^d\right)\\ &= \int_{\mathbb{R}^d} \tilde{\phi}(x) \cdot \left(\frac{(\eta_\kappa * \nu_k)(x)}{C_d\kappa^d}\right) \mathrm{d}x + \int_{\mathbb{R}^d} (\tilde{\phi}(x) - (\eta_\kappa * \tilde{\phi})(x))\,\mathrm{d}\left(\frac{\nu_k(x)}{C_d\kappa^d}\right) \qquad (3.3.14)\\ &\to \int_{\mathbb{R}^d} \tilde{\phi}(x) \cdot \rho^2(x)\,\mathrm{d}x = \int_\Omega \phi(x) \cdot \rho^2(x)\,\mathrm{d}x \quad \text{as } \kappa \to 0,\end{aligned}$$

where the second term in (3.3.14) converges to 0 because $\tilde{\phi} - (\eta_\kappa * \tilde{\phi})$ converges uniformly to 0 on Ω and $\|\nu_k/(C_d\kappa^d)\| \le 4\|\rho\|^2_{L^2(\Omega)}$ by means of (3.3.11). Hence, $\nu_\kappa/(C_d\kappa^d) \rightharpoonup^* \rho^2\mathcal{L}^d$ in $\mathcal{M}_+(\Omega)$ as $\kappa \to 0$. □

If ρ is substituted by a weak* approximation $\hat{\rho}$, Corollary 3.2.3 implies that minimizers $\hat{\nu}$ for $\hat{\rho}$ converge (up to subsequences) to a minimizer ν for ρ as $\hat{\rho} \overset{*}{\to} \rho$. When ρ is atomless, for small κ, by Proposition 3.3.10 and Lemma 3.3.11 we conclude that ν (and thus also $\hat{\nu}$) is close to the zero measure. If we are now interested in the "residuals" of ν and $\hat{\nu}$ (i.e., if we re-scale them such that their mass is on the order 1), then Proposition 3.3.13 tells us that we can only expect the residual of $\hat{\nu}$ to be close to that of ν when $\hat{\rho}$ approximates ρ well in an L^2-sense. Therefore, if we were interested in using the HK-barycenter to obtain a quantization or clustering of some measure ρ at a small κ-scale, but only an approximation $\hat{\rho}$ is available, then the approximate solution $\hat{\nu}$ will only be useful, if $\hat{\rho}$ is a good approximation in an L^2-sense. (Intuitively, we expect that it is sufficient if the L^2-approximation holds after an optional convolution with a mollifier at a scale less than κ.)

3.4 Numerical examples

To obtain a better understanding of the behavior of the HK barycenter between Dirac measures and to illustrate the theoretical results of the previous section we now consider some numerical approximations.

3.4.1 Lagrangian optimization scheme

Let us first consider the discrete barycenter problem between $r \in \mathbb{N}$ unit Dirac measures on Ω with $\mu_i = \delta_{x_i}$, $x_i \in \Omega$, and weights $\lambda_i > 0$ for $i = 1, \dots, r$ such that $\sum_{i=1}^r \lambda_i = 1$. This corresponds to $\Lambda := \sum_{i=1}^r \lambda_i \delta_{\mu_i}$ in $(\mathcal{P}_{\Lambda,\kappa})$ or equivalently $\rho := \sum_{i=1}^r \lambda_i \delta_{x_i}$ in $(\mathcal{P}_{\rho,\kappa})$.

For optimization over ν we employ a Lagrangian discretization, i.e. we optimize over the ansatz $\nu^s = \sum_{j=1}^s m_j \delta_{y_j}$ with locations $y_j \in \Omega$ and masses $m_j \ge 0$ for $j = 1, \dots, s$ for some $s \in \mathbb{N}$. The number of points in the ansatz s may change during optimization due to merging or addition of new particles. The resulting optimization problem can be written as

$$\min_{m_j \ge 0,\, y_j} J_{\rho,\kappa}(\nu^s) = \min_{m_j \ge 0,\, y_j} 1 + \sum_{j=1}^s m_j - 2\sum_{i=1}^r \lambda_i \sqrt{\sum_{j=1}^s m_j \operatorname{Cos}^2\left(\frac{|x_i - y_j|}{\kappa}\right)}. \qquad (3.4.1)$$

For gradient-based minimization we determine the partial derivatives with respect to mass coefficients m_j and locations y_j. The components of the gradient in mass are

$$\frac{\partial J_{\rho,\kappa}(\nu^s)}{\partial m_j} = 1 - \sum_{i=1}^{r} \lambda_i \frac{\operatorname{Cos}^2\left(\frac{|x_i - y_j|}{\kappa}\right)}{\sqrt{\sum_{l=1}^{s} m_l \operatorname{Cos}^2\left(\frac{|x_i - y_l|}{\kappa}\right)}}.$$

If we set $\psi^s(x_i) := 1 - \sqrt{\sum_{j=1}^{s} m_j \operatorname{Cos}^2\left(\frac{|x_i - y_j|}{\kappa}\right)}$ (which would be the optimal dual ψ if ν^s is primal optimal, see Proposition 3.3.4), then we obtain

$$\frac{\partial J_{\rho,\kappa}(\nu^s)}{\partial m_j} = 1 - F_{\rho,\kappa}(\psi^s)(y_j),$$

i.e. masses need to be increased when the corresponding constraint function at y_j is less than 1 (the constraint is inactive) or decreased when the constraint is violated. A vanishing gradient corresponds to the optimality condition $F_{\rho,\kappa}(\psi)(y_j) = 1$ on the support of ν.

For the gradient in coordinates y_j one obtains

$$\frac{\partial J_{\rho,\kappa}(\nu^s)}{\partial y_j} = \sum_{i=1}^{r} \lambda_i \frac{2m_j \operatorname{Cos}\left(\frac{|x_i - y_j|}{\kappa}\right) \operatorname{Sin}\left(\frac{|x_i - y_j|}{\kappa}\right) \frac{y_j - x_i}{\kappa |x_i - y_j|}}{\sqrt{\sum_{l=1}^{s} m_l \operatorname{Cos}^2\left(\frac{|x_i - y_l|}{\kappa}\right)}},$$

where $\operatorname{Sin}(x) := \sin(x)$ for $x \in [0, \pi/2]$ and 0 otherwise. By comparison we find again a relation to the constraint function (with the same ψ as above),

$$\frac{\partial J_{\rho,\kappa}(\nu^s)}{\partial y_j} = -\frac{\partial F_{\rho,\kappa}(\psi^s)(y_j)}{\partial y_j},$$

i.e. the points y_j will move "upwards" on the constraint function $F_{\rho,\kappa}(\psi)$ and only be locally optimal when sitting at a maximum, which then, by the mass optimality condition, has to be at value 1.

Due to the Lagrangian ansatz, the resulting optimization problem is non-convex and may get stuck in non-optimal points. On the other hand, because the points are allowed to move, the spatial accuracy is not limited to a grid. The issue of poor local minima can be remedied by testing whether the value of $F_{\rho,\kappa}(\psi)$ exceeds one at points where no y_j is located. This testing can be performed numerically with reasonable accuracy since $F_{\rho,\kappa}(\psi)$ is $1/\kappa$-Lipschitz continuous. Thus it is possible to combine the strengths of Lagrangian and Eulerian schemes: We remove points from the ansatz when their mass drops to zero, and we may add points when the dual constraint is violated, thus adaptively determining the appropriate number of point masses.

A Eulerian discretization with entropic smoothing was used in [65] for numerical examples. In Proposition 3.3.7 and [65, Proposition 6.2] it was shown that discrete minimizers exist when ρ is discrete. The entropic Eulerian ansatz cannot approximate them with high accuracy due to the fixed grid, entropic blur, and since we observe that minima are often "shallow" or non-unique. Therefore, to study these discrete minimizers the non-entropic Lagrangian method is more appropriate.

The problem formulated above is then solved with (preconditioned) gradient descent with gradient steps performed simultaneously in coordinates and in masses; inexact line search as described in [71].

Naturally, we are also interested in examples where ρ is not discrete, representing an uncountable infinite number of input measures. While Corollary 3.2.3 suggests that this case can be approximated with discrete ρ, we deduce from Proposition 3.3.13 that for small κ it will be difficult to obtain good approximations for the "residual" of ν_κ (which will be close to the zero measure). Therefore, for this regime we employ a slightly different numerical scheme where ν is also approximated in a (discrete) Lagrangian fashion but the integral over ρ in (3.3.2) is approximated more accurately by adaptive Gauss–Kronrod quadrature instead of individual Dirac sums as in (3.4.1). The corresponding formulas for the gradients are derived in analogy. For optimization of this functional we applied the quasi-Newton BFGS algorithm.

3.4.2 Finite number of input measures

For a discrete number of input Dirac measures, for κ sufficiently close to zero, it is easy to see that the resulting HK barycenter will be a superposition of Dirac measures, one per input measure (cf. Prop. 3.3.10). For κ sufficiently large, it will be one single Dirac measure. For κ increasing from 0 to ∞, a gradual merging of Diracs in the barycenter was observed in [65] on various examples. Here we demonstrate numerically that the general behavior is more complex. As κ increases, Diracs may merge and split, disappear and reappear, and the total number of Diracs may even temporarily increase.

Figure 3.1 illustrates the HK barycenter for

$$\rho := 0.4 \cdot \delta_0 + 0.1 \cdot \delta_{0.4} + 0.1 \cdot \delta_{0.6} + 0.4 \cdot \delta_1 \text{ on } \Omega = [0, 1], \qquad \text{for } \kappa \in [0.08, 0.8], \tag{3.4.2}$$

and the constraint function $F_{\rho,\kappa}(\psi)$ for the corresponding dual optimal ψ.

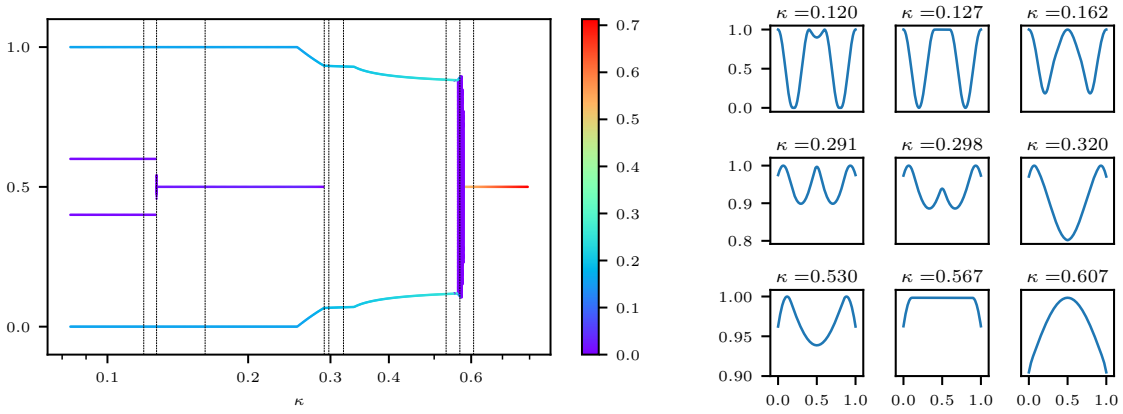

Figure 3.1: Left: the HK barycenter on $\Omega = [0, 1]$ for ρ and κ as in (3.4.2). For each κ, position of points indicates positions of Dirac measures, color code indicates the amount of mass. The vertical lines show locations for which the constraint function $F_{\rho,\kappa}(\psi)$ is shown on the right.

For small κ, as expected, the barycenter consists of four individual Dirac masses at the same locations as in ρ. Eventually the two middle masses merge (where the constraint function briefly exhibits an extended plateau of value 1, as analyzed in [65]). At some point, the outer masses

"see" the inner masses (i.e. their relative distance drops below $\kappa\pi/2$). Since their λ-weights are much higher, the joint Dirac in the barycenter remains much closer to the outer masses until the Dirac at the center even vanishes. Note that after this vanishing, the constraint function briefly even exhibits a local maximum at 0.5 which is strictly below 1. Eventually, all masses merge into one cluster. During the merging the constraint function exhibits an extended plateau of value 1 for an extended interval of κ values and in this regime a non-discrete barycenter exists (again, this was already analyzed in [65]).

Figure 3.2 shows a similar example with 6 masses in ρ, given by

$$\rho := 0.3 \cdot (\delta_0 + \delta_1) + 0.16 \cdot (\delta_{0.24} + \delta_{0.76}) + 0.03 \cdot (\delta_{0.45} + \delta_{0.55}). \tag{3.4.3}$$

.

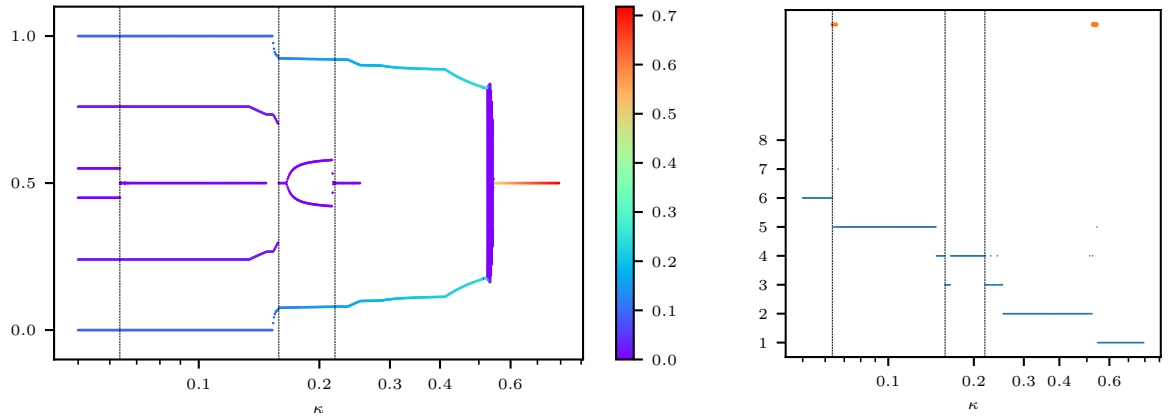

Figure 3.2: Left: the HK barycenter on $\Omega = [0, 1]$ for ρ with six Dirac masses (see (3.4.3)), visualized as in Figure 3.1. Right: the number of masses in the barycenter ν, which is not decreasing over κ. The regimes with a seemingly diffuse solution are marked by orange points.

The trajectory of HK barycenters over κ exhibits an even more intricate behavior with the mass at the center appearing and disappearing several times and even the number of masses temporarily increasing as κ increases. For at least two regions of scales, a diffuse barycenter seems to be admissible.

3.4.3 A continuum of input measures

Now we consider a continuum of input measures. Let $\Omega = [0, 1]$ and $\rho = \mathcal{L}^1 \llcorner \Omega$. Following Section 3.4.1 we consider the functional

$$J^s_{\rho,\kappa}(y_1, \dots, y_s, m_1, \dots, m_s) = J_{\rho,\kappa}(\nu^s) = 1 + \sum_{j=1}^{s} m_j - 2 \int_\Omega \sqrt{\sum_{j=1}^{s} m_j \operatorname{Cos}^2(|x - y_j|/\kappa)}\, \mathrm{d}x, \tag{3.4.4}$$

which corresponds to (3.4.1) with continuous ρ. The integral is approximated by adaptive Gauss–Kronrod quadrature and minimized via projected BFGS in positions and masses.

Figure 3.3 shows barycenters obtained for $\kappa \in [1/12, 1]$ and the evolution of the total mass with a detailed view presented in Figure 3.4.

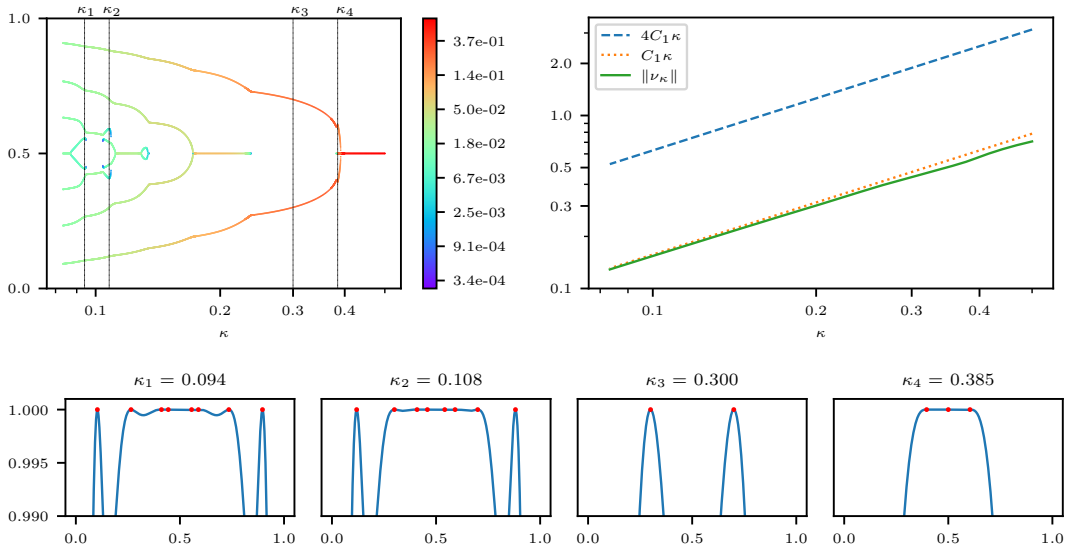

Figure 3.3: Top-Left: the HK barycenter on $\Omega = [0, 1]$ for $\rho = \mathcal{L} \llcorner [0, 1]$ visualized as in Figure 3.1 (here with a logarithmic color map). Top-Right: total mass of the HK barycenter, in comparison with bound and asymptotic expansion from Proposition 3.3.13. Bottom: the constraint function $F_{\rho,\kappa}(\psi)$ for some values of κ (as marked in the top-left), with positions $(y_j)_j$ of the primal masses marked by red points (note the range of the vertical axis, which only shows a very small interval close to one).

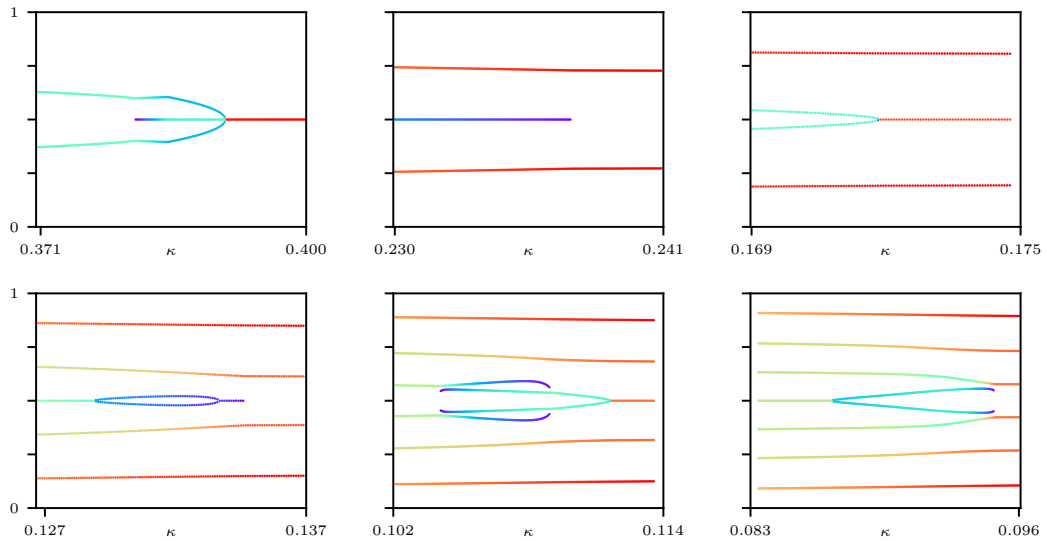

Figure 3.4: Zoom on the transition regions for Figure 3.3 for the given values of κ. For better visibility the color scale is adjusted to the mass range in each sub-figure.

In agreement with Proposition 3.3.13 the latter decreases linearly to 0 as $\kappa \to 0$. The evolution seems to consist of intervals in which the number of Dirac masses gradually decreases by one step at a time until only a single mass is left. However, these intervals are separated by transition regions, during which the behavior is more complicated and the number of masses also temporarily increases again.

Due to the uniformity of ρ this problem proved to be quite challenging numerically, as the constraint function for the optimal ψ was very close to one, almost throughout the entire bulk of the interval, see Figure 3.3. In particular the transition regions required detailed manual inspection. It is possible to solve the problem analytically for very small and very large κ, but the full spectrum seems to be beyond reach. Therefore, it seems ultimately impossible to prove that the true minimizers have the same structure as our numerical approximations. But via the primal-dual optimality conditions we can at least guarantee that the numerical approximations must be very close in terms of objective value. In particular, the observed complicated transitions seem to outperform simpler variants without additional particles. These transitions are shown in more detail in Figure 3.4.

It seems that each of the shown transitions follows a different pattern: From one to two particles, first a fork into three particles is observed, and then the middle particle vanishes (numerically it seems that in this region also a diffuse solution would be admissible, but we were unable to find a solution with less than three particles). In the transition from two to three, the third particle simply appears in the middle. From three to four, the middle particle splits into two. From four to five, a new particle first appears, then splits, and finally re-merges. From five to six, a particle first splits, but the two fragments then vanish and are replaced by appearing new particles. From six to seven, two particles appear and eventually merge. We did not anticipate such a complicated structure in a convex functional.

3.4.4 Comparison with empirical measures

Next, we study the convergence of the HK barycenter as ρ is approximated through sampling. For the previous example with ρ being the uniform measure on $[0, 1]$ we now generate $\hat{\rho}$ by drawing n points from ρ and using the obtained empirical measure. Corresponding empirical barycenters are shown in Figure 3.5.

By Corollary 3.2.3 we expect convergence of the empirical barycenter to the true one, as $n \to \infty$. However, as $\kappa \to 0$, the true barycenter will converge to the zero measure (Proposition 3.3.10). Convergence of the "residual" part is analyzed in Proposition 3.3.13 for the case when ρ has an L^2-density. From this we expect that the residual of the empirical barycenter will be close to the real residual, when $\hat{\rho}$ is a good L^2-approximation of ρ. As $\hat{\rho}$ is an empirical measure, it has no L^2-density. Intuitively, we expect the result to still hold when a small convolution on a length scale below κ is applied to $\hat{\rho}$, and when this mollified version of $\hat{\rho}$ is close to ρ in an L^2-sense. In a nutshell, we expect the residuals to become worse as κ decreases and better as n increases. This is confirmed by the examples in Figure 3.5.

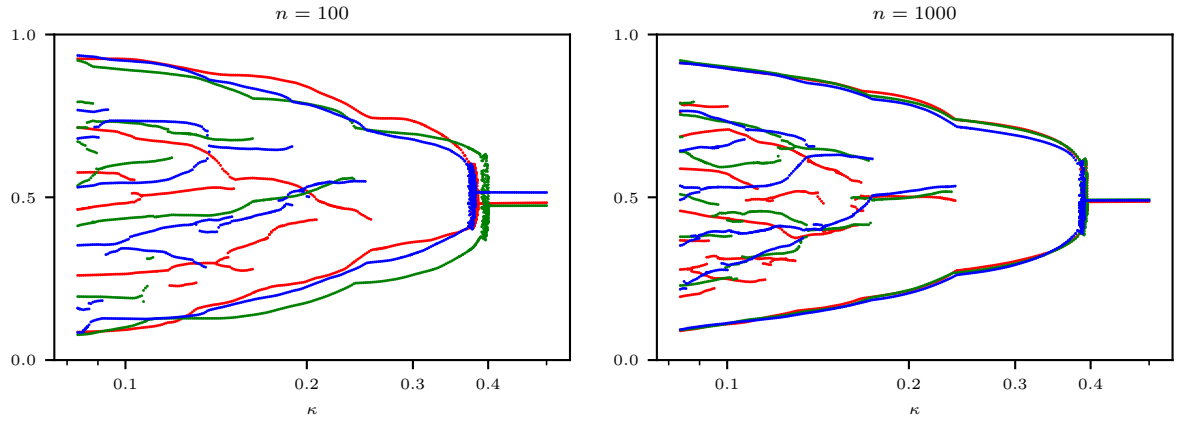

Figure 3.5: Barycenters for the input measures sampled from uniform distribution on $[0, 1]$. Points mark support of masses, mass itself is not visualized. Three different instances are shown in different colors to visualize the variance between them. Left: 100 points sampled. Right: 1000 points sampled.

Convergence of the dual solution and constraint function (Prop. 3.3.6) is visualized in Figure 3.6.

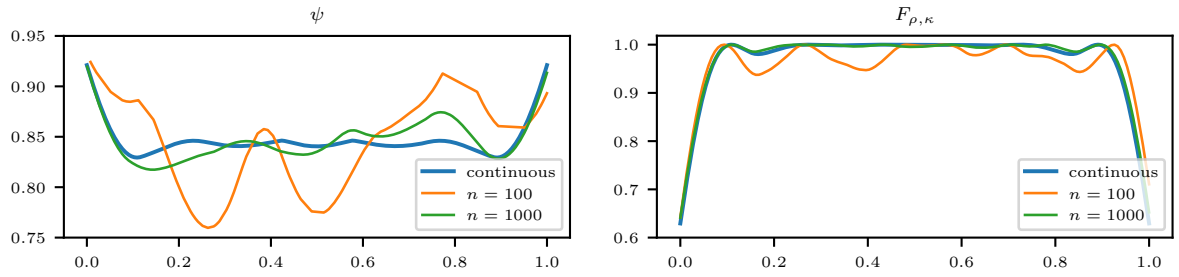

Figure 3.6: The dual (left) and the constraint function (right) at $\kappa = 0.1$, for two sampled solutions with $n = 100$ and $n = 1000$ together with the solution obtained from the continuous ansatz of Section 3.4.3 for comparison.

The uniformity of ρ in the example above makes it not quite clear what kind of "clustering" to expect for smaller κ. Therefore, we perform similar experiments on mixtures of Gaussian distributions. That is, ρ is given by a mixture of 5 Gaussians with means and standard deviations given by

$$(0.15, 0.05),\ (0.30, 0.03),\ (0.46, 0.08),\ (0.71, 0.03),\ (0.81, 0.06).$$

Figure 3.7 shows the corresponding numerical results for sampling $n_i = 100$ and $n_i = 1000$ points from each Gaussian. The coarse structure of the resulting HK barycenters seems to indicate five major clusters, one per Gaussian, that gradually merge. As expected, cluster masses are higher for more concentrated Gaussians, and the second cluster seems to absorb some points from the first Gaussian. As in the earlier examples, the transition between the major cluster intervals is more complicated, and differs between instances. Occasionally, smaller spurious particles with low mass are present. The rough structure of the barycenters seems consistent between all four experiments, in agreement with the proven stability results.

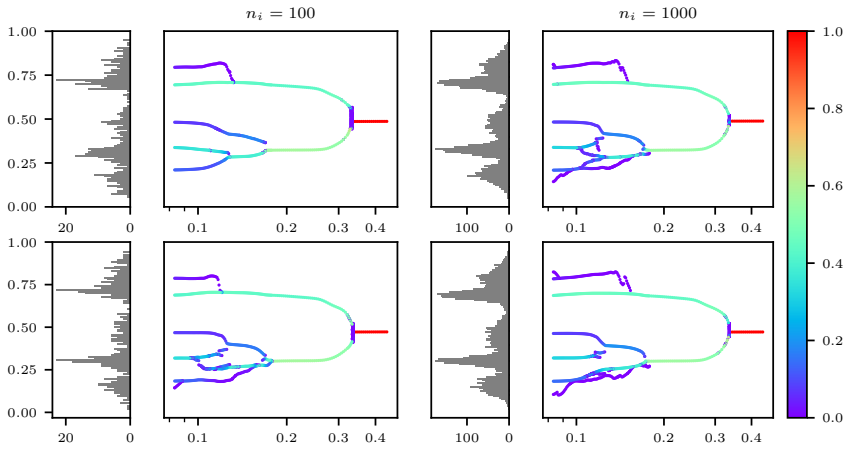

Figure 3.7: HK barycenters for samples from a mixture of 5 Gaussian distributions. Each column shows two instances. Left column: 100 samples per Gaussian. Right column: 1000 samples per Gaussian. Empirical distribution is visualized by the vertical histograms for each instance (50 bins in left column, 100 bins in right column). The mass of the barycenter for each κ is re-normalized to 1 for better visibility.

Figure 3.8 shows the regions of the constraint function $F_{\rho,\kappa}(\psi)$ close to one for the examples presented in Figure 3.7. The threshold is scaled with parameter κ as this seems to produce lines of approximately constant width over the scales. By Proposition 3.3.6 we know that these regions are unique and convergent as $n \to \infty$. They seem to be in good correspondence with the primal pictures and are reasonably stable under repeated experiments.

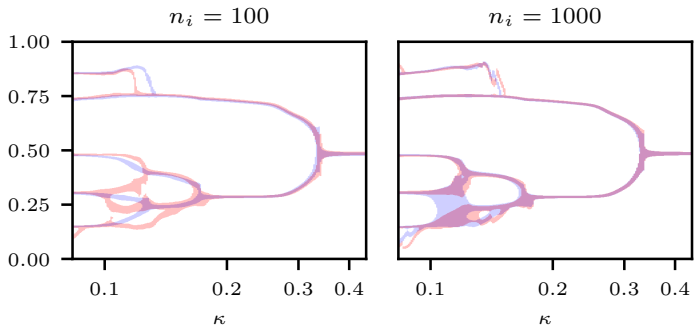

Figure 3.8: Thresholded regions where the dual constraints are close to being active, i.e. where $F_{\rho,\kappa}(\psi) \geq 1 - \frac{\exp(-9.5)}{\kappa}$, for 2 instances with $n_i = 100$ (left) and $n_i = 1000$ (right) sampled points in each Gaussian presented in Figure 3.7. The re-scaling with κ was done based on the empirical observation that it yielded approximately consistent widths of the lines.

For visual comparison, Figure 3.9 shows four single linkage cluster dendrograms for the sampled points presented above, computed with the algorithm described in [101].

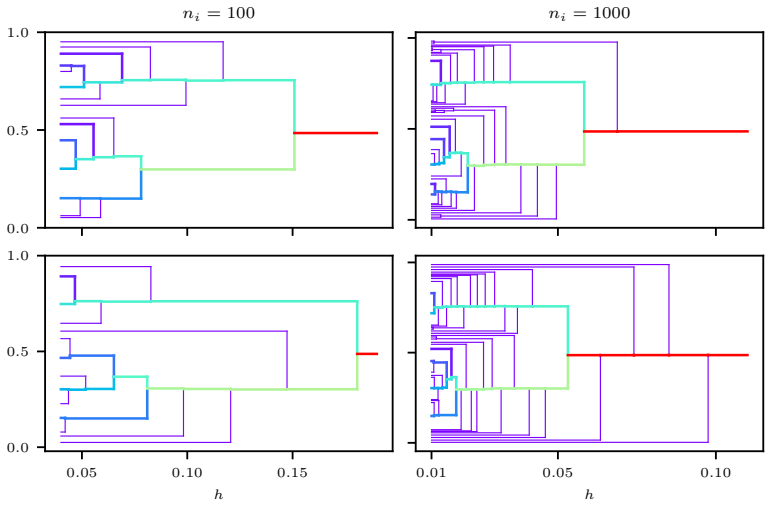

Figure 3.9: Single linkage cluster dendrograms for two instances with $n_i = 100$ (left) and for 2 instances with $n_i = 1000$ (right) sampled points in each Gaussian presented in Figure 3.7, truncated for better visibility. Colorscale represents the number of points clustered in a branch. Branches carrying less then 2% of the total mass are shown with narrower lines for better visibility.

The obtained major clusters are qualitatively similar to the results obtained by the HK barycenter. It differs from the HK barycenter figure in some important features: First, it is well known that single linkage produces many spurious outliers that show up in the figure as dark, thin lines. The HK barycenter seems less prone to such outliers. Second, the obtained dendrograms for $n_i = 100$ and $n_i = 1000$ appear to be shifted horizontally against each other, since the expected distances between pairs of points are different in both cases. For the HK barycenter the behavior is qualitatively consistent at a fixed κ for different n_i. Third, of course the dendrogram does provide a strict hierarchical clustering of the data, unlike the HK barycenters, and it can be computed with simple and efficient algorithms. These are features that the HK barycenter does not offer. For an approach to making the dendrograms more robust to spurious outliers, see for instance [39].

3.4.5 A two-dimensional example

Finally, we present some two-dimensional examples. We start with the uniform density on the square $[0, 1]^2$, discretized by 131^2 discrete Dirac masses, see Figure 3.10. As in one dimension, an intricate sequence of transitions between relatively regular intervals is observed.

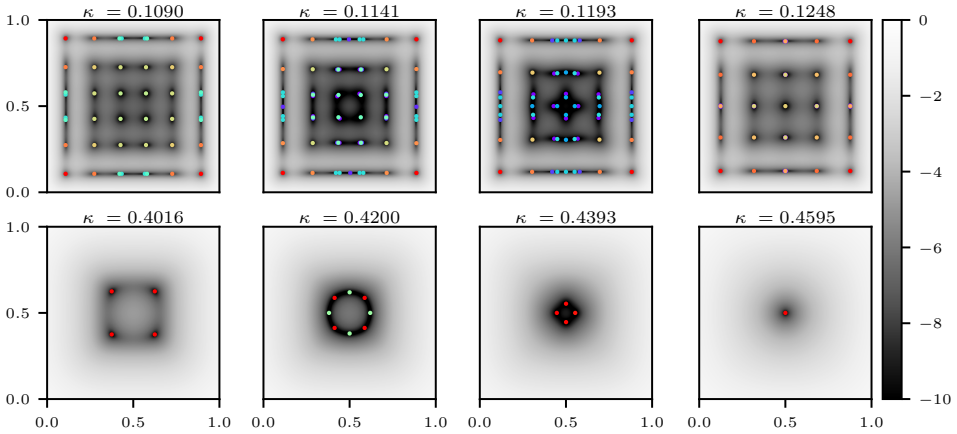

Figure 3.10: HK barycenter for various κ for uniform ρ. In grayscale: dual feasibility residual in log-scale ($\ln(|1 - F_{\rho,\kappa}(\psi)|)$). In color scale: locations and masses of barycenter with maximal mass in each barycenter re-normalized to 1 for better visibility.

An example with a mixture of 3 Gaussians is shown in Figure 3.11. For this experiment, 50 points were sampled from each 2D Gaussian distribution. The sampled points are shown in the plot in grey, the barycenter for selected κ is shown in color. The HK barycenter in this case presents a "clustering" behavior similar to the one shown for a mixture of Gaussians in 1D.

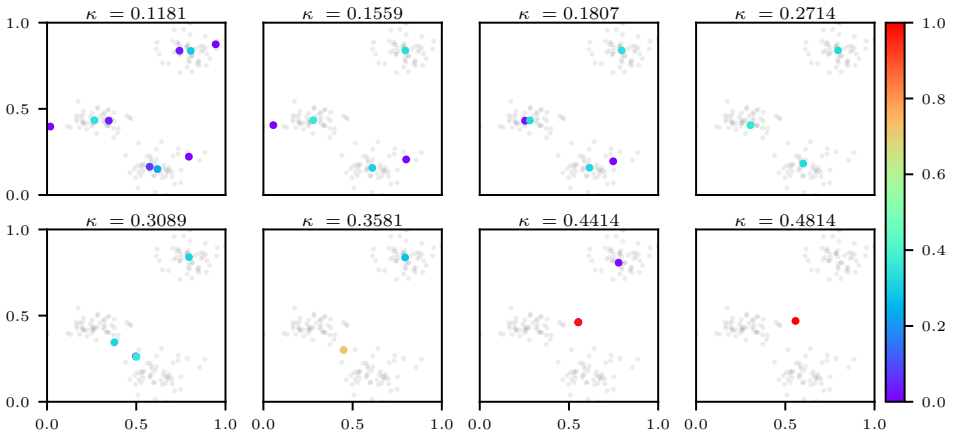

Figure 3.11: HK barycenter for various κ for ρ sampled from mixture of Gaussians in 2D. Sampled points are shown in grey. The mass of the barycenter for each κ is re-normalized to 1 for better visibility.

3.5 Conclusion

In this chapter we have studied in more detail the barycenter between an uncountable number of input measures with respect to the Hellinger–Kantorovich distance, with a particular focus on Dirac input measures. We have shown existence and stability with respect to input data and length scale parameter and derived a corresponding dual problem. For Dirac input measures we have shown existence of continuous dual maximizers, their uniqueness (ρ-a.e.) and primal-dual optimality conditions. The behavior of the solutions as $\kappa \to 0$ was studied in more detail, including the limit solution and asymptotic mass and density estimates. We showed that in some cases no discrete minimizers can exist. A numerical scheme based on Lagrangian discretization was introduced and it was shown numerically that the evolution of the minimizer with respect to the length scale does not correspond to a simple gradual merging of "clusters". With these two properties (non-existence of discrete minimizers, no simple merging behavior) the HK barycenter does not induce a simple hierarchical clustering of data points in the conventional sense. Instead, a wide variety of transition behaviors is observed numerically. However, it still provides a one-parameter family of measures, interpolating between the input data and a single Dirac measure, which can be interpreted as a gradual coarse graining. It is reasonably robust under empirical approximation by sampling, as demonstrated theoretically and with numerical examples, and comes with a corresponding family of dual problems that provide additional interpretation and information.

As such Hellinger–Kantorovich barycenter might be an interesting tool for the structure analysis of point clouds and application to real data would be a possible direction for future research. This would lead to related questions such as the interpretation of the trajectory of barycenters in high dimensions and their reliable numerical approximation.

4 Branched and Multimaterial Transport

This chapter is devoted to the branched transport problem and its convex relaxation in terms of the multimaterial transport problem. This chapter is mainly based on a manuscript in preparation to be submitted for publication [24].

As already mentioned in the introduction, various natural and human-made objects have branching structures. Figure 4.1 shows a simple example which demonstrates the difference between the standard optimal transport and branching behavior: The transport is performed between two sources (at A and B) and a sink (at C); the standard optimal transport is shown on the left, and the branched transport, where the flows are joining at an intermediate point and then continue together to the sink, is shown on the right. The observed behavior is modelled through a special construction of the transportation cost: In the case of branched transport, the cost is selected to be a concave, subadditive function, which encourages the transport network ramification.

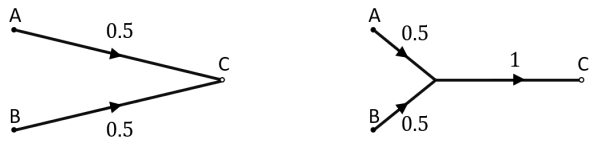

Figure 4.1: Sketch of Wasserstein-1 transport (left) and branched transport (right) from $\frac{1}{2}(\delta_A + \delta_B)$ to δ_C.

The concave cost function, however, makes the problem difficult to analyze. Alternative models promoting network ramification exist, for example, in the setting of multimaterial transport, where the input is assigned to be of different "materials", and the (convex) cost function gives reduction for transporting different materials together. An example of multimaterial transport problem giving the same network as in the branched transport problem in Figure 4.1 (right) is given in Figure 4.2.

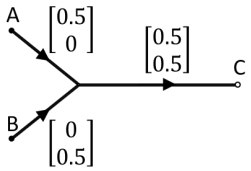

Figure 4.2: Sketch of multimaterial transport from $\frac{1}{2}\left(\begin{bmatrix}1\\0\end{bmatrix}\delta_A + \begin{bmatrix}0\\1\end{bmatrix}\delta_B\right)$ to $\frac{1}{2}\begin{bmatrix}1\\1\end{bmatrix}\delta_C$.

Therefore, in this chapter, we first recall the branched transport problem and the multimaterial transport problem and then study the latter in more detail. This chapter is organized as follows.

In Section 4.1 we recall the statement of the branched transport problem in Eulerian description.

In Section 4.2 we state the multimaterial transport problem and discuss the properties of the multimaterial transport cost. We also give a dual formulation and an alternative primal formulation and analyze the primal-dual optimality conditions. While the results presented in Section 4.2 are in principle known in the literature, they appear to be scattered throughout a number of publications and are given in different mathematical formulations. We collected the results and presented them in a mathematically homogeneous language suitable for the subsequent research.

Section 4.3 is devoted to a study of a special subclass of multimaterial problems that admits a single topology. Here we formally state the single topology problem and a free vertex optimization problem and then show the connection between the two. We also consider the minimal example of a single topology problem, namely the problem with 2 sources and 1 sink, and provide a full characterization of its solution set.

Section 4.4 is devoted to the numerical experiments. We present numerical schemes of two different classes: a scheme based on optimization over graphs and a scheme based on finite-element-like discretization of the original infinite-dimensional problem. We present some examples, which support our findings from the previous sections and make observations about the structure of the problem (especially the seemingly simple structure of the dual), which motivate the upcoming research.

In Section 4.5 we study a multimaterial problem with 3 sources and 1 sink in a special setting when two solutions with different topologies are optimal. We study the properties of the problem from geometric and algebraic perspectives. We discover and prove an interesting connection between the transportation costs of the candidates and the bisectors of the angles of the graphs. We then study the set of solutions of the prescribed type and discover some configurations, which can a priori (i.e. based only on the initial data of the problem) be shown not to admit solutions of the prescribed type.

Conclusions and discussion of presented results and potential future work are given in Section 4.6.

Author's contribution
The author has made major contributions to all the sections. In particular, the author contributed to adaptation and presentation of the results given in Sections 4.2 and 4.3. The author proposed the research question on the properties of the 3-vertex problem presented in Subsection 4.3.3 and then contributed to formulating conjectures and proving them. The author contributed to implementing and adapting the numerical schemes discussed in Section 4.4, conducted the numerical experiments and analyzed the results. The author helped to define the general direction of the research in Section 4.5 and participated in discovering, formulating and proving the conjectures on the properties of the specific 4-vertex problem.

4.1 Branched transport

Here we follow the branched transport problem in Eulerian description introduced by Xia [122] and generalized to concave transportation costs by Brancolini and Wirth [32]. The alternative Lagrangian model of branched transport was proposed by Maddalena, Solimini and Morel [90].

Let transportation cost $\tau(m) : [0,\infty) \mapsto [0,\infty)$ be a non-decreasing concave function with $\tau(0) = 0$.

Definition 4.1.1. A *polyhedral mass flux* between discrete probability measures $\mu_+ \in \mathcal{P}(\Omega)$ and $\mu_- \in \mathcal{P}(\Omega)$ is a vector-valued Radon measure $\mathcal{F} \in \mathcal{M}(\Omega, \mathbb{R}^n)$ which satisfies in the distributional sense the Kirchhoff's mass preservation law

$$\mathrm{div}\mathcal{F} = \mu_+ - \mu_- \tag{4.1.1}$$

and can be written as

$$\mathcal{F} = \sum_e m_e \sigma_e \mathscr{H}^1 \llcorner e, \tag{4.1.2}$$

where $e = x_e + [0,1](y_e - x_e) \subset \Omega$ are the non-overlapping edges with orientation $\sigma_e = (y_e - x_e)/|y_e - x_e|$, coefficients $m_e \in \mathbb{R}_{++}$ are positive real weights, and $\mathscr{H}^1 \llcorner e$ is the restriction of the one-dimensional Hausdorff measure to the edge e.

Definition 4.1.2. The *branched transport cost* of polyhedral mass flux $\mathcal{F}$ with respect to a transportation cost τ is defined as

$$\mathcal{J}^{\tau,\mu_+,\mu_-}[\mathcal{F}] = \sum_e \tau(m_e)\mathscr{H}^1(e). \tag{4.1.3}$$

The problem of minimizing functional (4.1.3) in the case $\tau(m) = m^\alpha$ with $\alpha \in (0,1)$

$$\inf \left\{ \sum_e m_e^\alpha \mathscr{H}^1(e) \;\middle|\; \mathcal{F} \text{ as in Definition 4.1.1} \right\} \tag{4.1.4}$$

is called the *Gilbert–Steiner problem* [68]. Note that the limit case $\alpha = 0$ corresponds to the Steiner problem (see [46, 34] for historical overview), while case $\alpha = 1$ corresponds to Wasserstein-1 transport.

The branched transport cost of a general Radon measure satisfying Kirchhoff's mass preservation law (4.1.1) is defined via relaxation of the branched transport cost of polyhedral mass fluxes $\mathcal{J}^{\tau,\mu_+,\mu_-}$:

Definition 4.1.3. A vector-valued Radon measure $\mathcal{F} \in \mathcal{M}(\Omega, \mathbb{R}^n)$ is called *mass flux* between the probability measures μ_+ and μ_- if there exist two sequences of discrete probability measures $(\mu_+^k), (\mu_-^k) \subset \mathcal{P}(\Omega)$ and a sequence of polyhedral mass fluxes $\mathcal{F}_k$ with $\mathrm{div}(\mathcal{F}_k) = \mu_+^k - \mu_-^k$ such that $\mathcal{F}_k \rightharpoonup^* \mathcal{F}$ and $\mu_\pm^k \rightharpoonup^* \mu_\pm$, where $\rightharpoonup^*$ indicates the weak-$*$ convergence in duality with continuous functions. In this case, we write $(\mathcal{F}_k, \mu_+^k, \mu_-^k) \rightharpoonup^* (\mathcal{F}, \mu_+, \mu_-)$.

If $\mathcal{F} \in \mathcal{M}(\Omega, \mathbb{R}^n)$, then the *branched transport cost* of $\mathcal{F}$ is defined as

$$\mathcal{J}^{\tau,\mu_+,\mu_-}[\mathcal{F}] = \inf \left\{ \liminf_k \mathcal{J}^{\tau,\mu_+^k,\mu_-^k}[\mathcal{F}_k] \;\middle|\; (\mathcal{F}_k, \mu_+^k, \mu_-^k) \rightharpoonup^* (\mathcal{F}, \mu_+, \mu_-) \right\},$$

the lower semicontinuous envelope or relaxation of the branched transport cost on polyhedral mass fluxes.

The corresponding *branched transport problem* is the optimization problem

$$\inf \left\{ \mathcal{J}^{\tau,\mu_+,\mu_-}[\mathcal{F}] \;\middle|\; \mathcal{F} \in \mathcal{M}(\Omega, \mathbb{R}^n),\ \mathrm{div}(\mathcal{F}) = \mu_+ - \mu_- \right\}.$$

As already mentioned above, the concave cost function makes it difficult to solve the problem both analytically and numerically. Therefore, it has been suggested in [92] to consider a relaxation of the branched transport, called *multimaterial* (or sometimes multi-material) transport problem. The rigorous arguments and results on the relaxation can be found in [92, 87].

4.2 Multimaterial transport problem

4.2.1 Problem statement

The multimaterial transport problem seeks to find the best (joint) displacement for m materials between their prescribed initial and final distributions in convex, compact set $\Omega \subset \mathbb{R}^n$ (with non-empty interior). Let initial and final distributions of material i be denoted by $\mu_+^i, \mu_-^i \in \mathcal{M}_+(\Omega)$, $i = 1, \dots, m$ where we impose the consistency condition $\|\mu_+^i\| = \|\mu_-^i\|$. These distributions can be summarized into two vector-valued measures $\mu_+ = (\mu_+^1, \dots, \mu_+^m)^\intercal, \mu_- = (\mu_-^1, \dots, \mu_-^m)^\intercal \in \mathcal{P}(\Omega)^m$.

The *multimaterial transport problem* consists in finding a collection of fluxes $\omega = [\omega^1, \dots, \omega^m] \in \mathcal{M}(\Omega)^{m\times n}$, where row $\omega^i \in \mathcal{M}(\Omega)^n$ is the flux of material i, minimizing the total cost of transporting all the mass from μ_+ to μ_-, given by

$$\mathcal{P}(\mu_+, \mu_-) = \inf_{\omega} \left\{ \int_\Omega H\left(\frac{\mathrm{d}\omega}{\mathrm{d}|\omega|}\right) \mathrm{d}|\omega| \;\middle|\; \omega \in \mathcal{M}(\Omega)^{m\times n} :\ \mathrm{Div}(\omega) = \mu_+ - \mu_- \right\}, \tag{4.2.1}$$

where the divergence operator Div acts row-wise on each material flux ω^i and is to be understood in a weak sense. That is, we impose

$$\int_\Omega \nabla\phi \mathrm{d}\omega^i + \int_\Omega \phi \mathrm{d}(\mu_+^i - \mu_-^i) = 0 \tag{4.2.2}$$

for all test functions $\phi \in \mathrm{C}^1(\Omega)$ and for all $i = 1, \dots, m$. This constraint describes mass conservation (also known as Kirchoff's law) for the material fluxes.

The multimaterial transport cost $H : \mathbb{R}^{m\times n} \to [0, \infty)$ is assumed to be convex, lower semi-continuous and positively 1-homogeneous. More specifically, we assume that H is constructed from a unit transport cost $h : \mathbb{R}^m \to [0, \infty)$ in way described below, where $h(\theta)$ gives the cost of transporting a combination of materials $\theta \in \mathbb{R}^m$ along one unit of length (signed vectors θ can represent the case where some materials travel in opposite directions). Further, we assume that the unit transport cost h is constructed from some prescribed prototypical material combinations $\Theta \subset \mathbb{R}^m$ and corresponding cost coefficients $c_\theta \in \mathbb{R}_+$ for $\theta \in \Theta$. Given Θ and $(c_\theta)_{\theta\in\Theta}$ we set h to be the largest convex, positively 1-homogeneous, even symmetric function $\mathbb{R}^m \to [0, \infty]$ that satisfies $h(\theta) \le c_\theta$ for $\theta \in \Theta$. More explicitly, one first introduces a preliminary function $\hat{h}$ by

$$\hat{h}(p) := \begin{cases} |\lambda| c_\theta, & \text{if } p = \lambda\theta \text{ for some } \lambda \in \mathbb{R},\ \theta \in \Theta, \\ +\infty, & \text{else.} \end{cases} \tag{4.2.3}$$

and then sets h to be its convex lower-semicontinuous envelope $h := \hat{h}^{**}$. h inherits positive 1-homogeneity and evenness from $\hat{h}$. For the convex conjugate $\hat{h}^*$ one finds

$$\hat{h}^*(q) = \sup_\lambda \max_{\theta\in\Theta} \lambda\langle q, \theta\rangle - |\lambda| c_\theta = \begin{cases} 0, & \text{if } |\theta^\intercal q| \le c_\theta \quad \forall \theta \in \Theta, \\ +\infty, & \text{else.} \end{cases} \tag{4.2.4}$$

Since $\hat{h}^* = \hat{h}^{***} = h^*$ and by positive 1-homogeneity one has that $\hat{h}^*$ is the indicator function of the subdifferential $\partial h(0)$.

Throughout this chapter we assume that

$$0 \in \operatorname{int} \partial h(0). \tag{4.2.5}$$

This implies, for instance, that $\inf_{\theta\in\Theta} c_\theta > 0$.

Given h we then construct H, by specifying that the cost for a collection of materials $p \in \mathbb{R}^m$ being transported in direction $e \in \mathbb{R}^n$ is given by $h(p) \cdot \|e\|$. For a general multimaterial flux matrix the cost is defined again via the convex, positively 1-homogeneous envelope. Analogous to above introduce a preliminary function $\hat{H}$ via

$$\hat{H}(P) := \begin{cases} h(p)\ \|e\|, & \text{if } P = p \otimes e, \\ +\infty, & \text{else.} \end{cases} \tag{4.2.6}$$

Note that because unit cost function h and the norm $\|\cdot\|$ are 1-homogeneous, the function $\hat{H}(P)$ is defined consistently even if the decomposition $P = p \otimes e$ is not unique. Then, like before, we set $H = \hat{H}^{**}$. For $\hat{H}^*$ one finds

$$\begin{aligned} \hat{H}^*(Q) &= \sup_{p,e} \langle Q, p \otimes e \rangle - h(p)\,\|e\| = \sup_{p,e} p^\intercal Q e - h(p)\,\|e\| = \sup_{e:\|e\|=1} \sup_{p} p^\intercal Q e - h(p) \\ &= \sup_{e:\|e\|=1} h^*(Qe) = \begin{cases} 0, & \text{if } |\theta^\intercal Q e| \le c_\theta \quad \forall \theta \in \Theta, e : \|e\| = 1, \\ +\infty, & \text{else} \end{cases} \\ &= \begin{cases} 0, & \text{if } \|\theta^\intercal Q\| \le c_\theta \quad \forall \theta \in \Theta, \\ +\infty, & \text{else.} \end{cases} \end{aligned} \tag{4.2.7}$$

As above, the function $H^* = \hat{H}^*$ is the indicator function of the subdifferential $\partial H(0)$. H inherits convexity, positive 1-homogeneity, and lower-semicontinuity from h. Assumption (4.2.5) implies that

$$0 \in \operatorname{int} \partial H(0). \tag{4.2.8}$$

Further, one has

$$H(p \otimes e) = h(p)\,\|e\| \tag{4.2.9}$$

for any mass vector p and direction e [87, Lemma 4.1.0.1]. Note also that by Carathéodory's theorem, for any flow $P \in \mathbb{R}^{m\times n}$ there exists a decomposition $P = \sum_i \lambda_i\ \theta_i \otimes e_i$ for some $\{\lambda_i\} \subset \mathbb{R}$, $\{\theta_i\} \subset \Theta$ and $\{e_i\} \subset \mathbb{R}^n$, such that $H(P) = \sum_i \lambda_i\ h(\theta_i)\ \|e_i\| = \sum_i |\lambda_i|\ c_{\theta_i}\ \|e_i\|$.

Example 4.2.1. Let $m = 2$, $\Theta = \{[0,1]^\intercal, [1,0]^\intercal, [1,1]^\intercal\}$, fix a parameter $q \in [0,1]$ and set the cost of transporting a unit of mass of each individual material to 1, and the cost of transporting one unit of each material *jointly* to $1+q$, so $c_{[0,1]} = 1, c_{[1,0]} = 1, c_{[1,1]} = 1+q$. Figure 4.3 illustrates $\hat{h}$ and its envelope (or extension) h.

Figure 4.4 shows the subdifferential $\partial h(0)$. As can be seen from (4.2.4), each vector $\theta \in \Theta$ induces two bounding hyperplanes with orientations given by $\pm\theta$ and offsets from the origin given by c_θ.

Finally, note that the choice $q = 1$ corresponds to the regular Wasserstein-1 cost, and $q = 0$ to the Steiner cost.

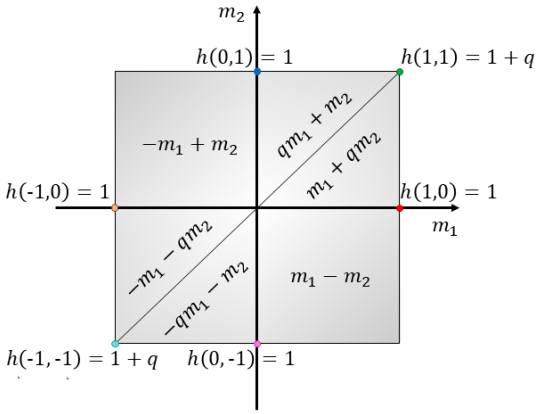

Figure 4.3: Unit cost function $h(m_1, m_2)$.

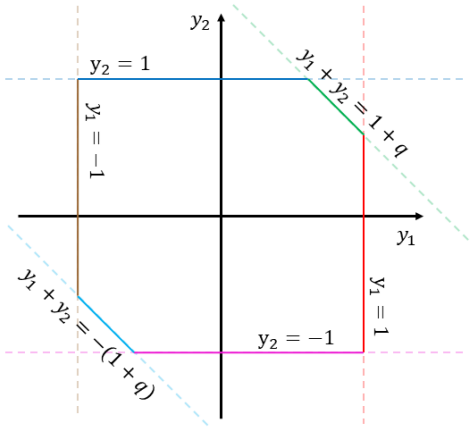

Figure 4.4: Subdifferential $\partial h(0)$.

4.2.2 Dual formulation

A dual multimaterial problem can be obtained via Fenchel–Rockafellar duality.

Proposition 4.2.2 (Dual multimaterial problem)**.** *Let*

$$\mathcal{D}(\mu_+, \mu_-) = \sup \left\{ \int_\Omega \varphi \cdot \mathrm{d}(\mu_+ - \mu_-) \;\middle|\; \varphi \in \mathrm{C}^1(\Omega)^m,\, -D\varphi(x) \in \partial H(0)\, \forall\, x \in \Omega \right\}. \tag{4.2.10}$$

Then one has $\mathcal{D}(\mu_+, \mu_-) = \mathcal{P}(\mu_+, \mu_-)$.

The dual problem (4.2.10) can be interpreted as a generalization of the Kantorovich–Rubinstein formula. Recalling the characterization (4.2.7) of $\partial H(0)$ the dual constraint can be written as

$$-D\varphi(x) \in \partial H(0) \Leftrightarrow \|\nabla \theta^\intercal \varphi(x)\| \leq c_\theta \quad \forall \theta \in \Theta, \tag{4.2.11}$$

where $\theta^\intercal \varphi = \sum_{i=1}^m \theta_i \varphi_i \in \mathrm{C}^1(\Omega)$. That is for every prescribed combination of materials (given by the set Θ), the norm of the gradient of the corresponding (weighted) sum of dual potentials is limited by the corresponding unit cost.

Proof. The primal problem (4.2.1) can be written as

$$\inf_{\omega \in \mathcal{M}(\Omega)^{m \times n}} \{F(A\omega) + G(\omega)\}$$

for $F = \iota_{\{\mu_+ - \mu_-\}}$, $G(\omega) = \int_\Omega H(\frac{\mathrm{d}\omega}{\mathrm{d}|\omega|})\mathrm{d}|\omega|$, and $A = \mathrm{Div}$. By Fenchel–Rockafellar duality theorem 2.2.6, the dual problem is then formally given by

$$\inf_{\omega \in \mathcal{M}(\Omega)^{m \times n}} \{F(A\omega) + G(\omega)\} = \sup_{\varphi \in \mathrm{C}^1(\Omega)^m} \{-F^*(-\varphi) - G^*(A^*\varphi)\}, \tag{4.2.12}$$

where A^* is the adjoint of operator A and F^*, G^* are convex conjugates of respectively F and G. One obtains

$$F^*(\phi) = \int_\Omega \phi \mathrm{d}(\mu_+ - \mu_-), \qquad G^*(\psi) = \begin{cases} 0, & \text{if } \psi(x) \in \partial H(0)\, \forall\, x, \\ +\infty, & \text{else.} \end{cases} \tag{4.2.13}$$

and $A^* = -D$ where D is the component-wise gradient of functions $\varphi \in \mathrm{C}^1(\Omega)^m$.

To show that the duality gap in (4.2.12) is in fact zero, we need to verify the constraint qualifications. In this case, this has to be done on the dual side. Indeed, for the function $\varphi : x \mapsto 0$ one has that F^* is finite at $-\varphi$ and G^* is finite and continuous at $A^*\varphi$, where the latter follows from the fact that $0 \in \operatorname{int} \partial H(0)$, (4.2.8). □

Remark 4.2.3. In problem (4.2.10) the dual potentials φ have to be continuously differentiable, which is a strong restriction, and solutions to this problem do not always exist. It can be shown that the class of admissible functions can be relaxed to Lipschitz functions $\varphi \in \mathrm{C}^{0,1}(\Omega, \mathbb{R}^m)$ and that the relaxed dual problem admits an optimizer [87].

We also note that the dual solutions are often referred to as *certificates*, as an optimal dual solution can be used to confirm optimality of a feasible primal candidate. In the context of multimaterial transport the gradients of the dual potentials are also called *calibrations* [92].

4.2.3 Primal-dual optimality conditions

It is also useful to consider the primal-dual optimality conditions for this problem. We state here again the primal-dual optimality conditions (Theorem 2.2.9): For a primal-dual pair of optimization problems written in the form (4.2.12), candidates (ω, φ) are optimal if and only if

$$\begin{aligned} A\omega \in \partial F^*(-\varphi) \quad &\Leftrightarrow \quad -\varphi \in \partial F(A\omega) \quad \Leftrightarrow \quad F(A\omega) + F^*(-\varphi) = -\langle A^*\varphi, \omega \rangle , \\ A^*\varphi \in \partial G(\omega) \quad &\Leftrightarrow \quad \omega \in \partial G^*(A^*\varphi) \quad \Leftrightarrow \quad G(\omega) + G^*(A^*\varphi) = \langle A^*\varphi, \omega \rangle . \end{aligned} \tag{4.2.14}$$

For the multimaterial transport problem specifically, the two conditions become

$$\operatorname{Div}(\omega) = \mu_+ - \mu_-, \tag{4.2.15}$$

$$-D\varphi(x) \in \partial H(0) \ \forall\, x \in \Omega, \tag{4.2.16}$$

$$\left\langle -D\varphi(x), \frac{\mathrm{d}\omega}{\mathrm{d}|\omega|}(x) \right\rangle_F = H\left(\frac{\mathrm{d}\omega}{\mathrm{d}|\omega|}(x)\right) \quad |\omega| - \text{almost everywhere.} \tag{4.2.17}$$

Conditions (4.2.15) and (4.2.16) are the primal and dual feasibility constraints and the latter can be rewritten as in (4.2.11). To interpret condition (4.2.17), assume for simplicity that

$$\frac{\mathrm{d}\omega}{\mathrm{d}|\omega|}(x) = \lambda\theta \otimes e \text{ for some } \lambda \in \mathbb{R},\ \lambda > 0,\ \theta \in \Theta \text{ with } h(\theta) = c_\theta, \text{ and } e \in \mathbb{R}^n, \|e\| = 1$$

$|\omega|$-almost everywhere on some closed set $S \subset \Omega$. Then on S we can rewrite the left side of (4.2.17) as

$$\left\langle -D\varphi(x), \frac{\mathrm{d}\omega}{\mathrm{d}|\omega|}(x) \right\rangle_F = -\lambda \langle \theta^\intercal D\varphi(x), e \rangle.$$

The cost on the right side is by definition (4.2.9)

$$H\left(\frac{\mathrm{d}\omega}{\mathrm{d}|\omega|}(x)\right) = \lambda h(\theta)\, \|e\| = \lambda\, c_\theta,$$

and so condition (4.2.17) becomes

$$-\langle \theta^\intercal D\varphi(x), e \rangle = h(\theta)\, \|e\| = c_\theta.$$

Comparing this with the feasibility condition (4.2.11), which can be written as

$$\|\theta^\intercal D\varphi(x)\| \leq c_\theta,$$

and using that e has unit length, we conclude that $|\omega|$-almost everywhere on S it must hold that

$$\theta^\intercal D\varphi = -c_\theta\, e. \tag{4.2.18}$$

This means that if a certain mass combination $\theta \in \Theta$ is flowing in direction $e \in \mathbb{R}^n$ in an optimal primal solution, then the corresponding dual optimal solution must satisfy that the weighted combination of dual potentials $\theta^\intercal\varphi$ is maximally decreasing in direction e, as far as the constraint (4.2.11) is concerned. This is a natural generalization of the classical Wasserstein-1 problem where the single dual potential must be decrease with slope 1.

4.2.4 Momentum condition

The primal-dual optimality conditions become particularly strong when several material fluxes meet at a vertex. In the following we give a formal discussion and refer to [87], in particular Remark 4.2.0.4, for rigorous treatment of regularity. Let $x \in \Omega$ and consider a flow ω where several fluxes form a vertex at x (see Figure 4.5):

$$\omega = \sum_i (\theta_i \otimes e_i)\sigma_i\, \mathscr{H}^1 \llcorner \ell_i.$$

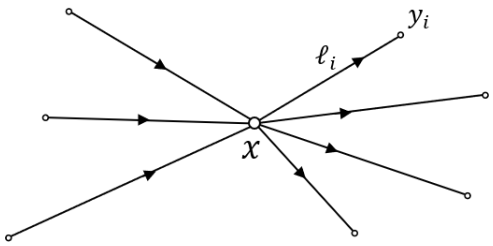

Figure 4.5: Incoming and outgoing flow at a point x.

Here vectors $\theta_i \in \Theta$ are the material compositions, $\ell_i \subset \mathbb{R}^n$ denote the straight line segments between the vertex x and the end-points $y_i \neq x$ with $e_i = (x - y_i)/\|x - y_i\|$ giving their orientation (pointing towards the vertex), and scalars $\sigma_i \in \{\pm 1\}$ determine the direction of flows along these edges ($+1$ for incoming, -1 for outgoing). One finds

$$\mathrm{Div}(\omega) = \sum_i \theta_i\sigma_i\, \mathrm{div}(e_i \mathscr{H} \llcorner \ell_i) = \sum_i \theta_i\sigma_i\, (\delta_{y_i} - \delta_x).$$

As source and sink measure we choose

$$\mu_+ = \sum_{i:\sigma_i=+1} \theta_i\, \delta_{y_i}, \qquad\qquad \mu_- = \sum_{i:\sigma_i=-1} \theta_i\, \delta_{y_i}$$

and thus from the constraint $\mathrm{Div}(\omega) = \mu_+ - \mu_-$ we obtain the mass preservation condition at the vertex x that

$$\sum_i \theta_i\sigma_i = 0. \tag{4.2.19}$$

Applying now condition (4.2.18) to each edge (and taking into account the orientation σ_i), at x we obtain for each i that

$$\sigma_i \, \theta_i^\intercal D\varphi = -c_{\theta_i} \, e_i.$$

Summing this over i and using (4.2.19) one then obtains the condition

$$\sum_i c_{\theta_i} e_i = 0, \tag{4.2.20}$$

known as the balance formula [122, Example 2.1] or momentum conservation condition [87, Remark 4.2.0.4].

In the special case when three edges meet, often referred to as a *triple junction*, this means that the relative magnitude of the costs $(c_{\theta_i})_{i=1}^3$ fully determines the relative angles between the edges, since they prescribe the edge lengths of a triangle. This is visualized in Figure 4.6.

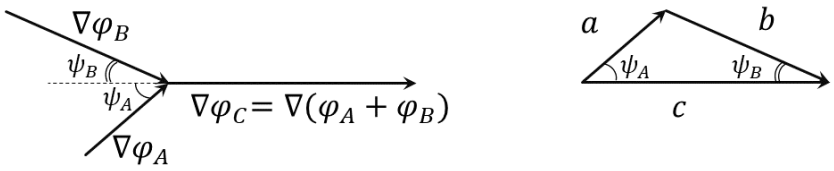

Figure 4.6: Junction with 3 flows (left) and the corresponding cost triangle (right).

For discrete distributions of sources and sinks, the optimal primal solution has been shown to be represented by a network [92], and so these local optimality conditions provide guidance for constructing solutions or at least solution candidates.

When fixing three boundary points y_i, mass compositions v_i and costs c_{v_i} it may however still be that the optimal solution does not exhibit a single vertex x with the implied angles, since either the three costs might not allow for the formation of a triangle (e.g. one being more expensive than the other two combined), or since the vertex x cannot be placed between the y_i with the prescribed angles. In these cases the optimal solution will be *degenerate* and x will coincide with one of the y_i. This is examined in more detail in Section 4.3.3.

For more than three edges there are more degrees of freedom and the relative angles are therefore not fully fixed by the costs.

4.2.5 Alternative primal formulation

After rewriting the dual constraint of (4.2.10) as in (4.2.11) we can obtain an alternative primal problem. We use again the Fenchel–Rockafellar duality theorem and pick $-F^*(-\varphi)$ as in (4.2.13) but we now choose $G^*(\phi) = \imath_{\|\phi_\theta\| \le c_\theta \ \forall \theta \in \Theta}$ and $(A^*\varphi)_\theta = \theta^\intercal D\varphi$. Note that in this case each of the dual constraints (4.2.11) induces a corresponding scalar flux in the primal problem. We denote the component corresponding to vector $\theta \in \Theta$ by ω_θ. Then the alternative primal formulation is given by

$$\inf \left\{ \sum_{\theta \in \Theta} c_\theta \, \|\omega_\theta\| \;\middle|\; \omega \in (\mathcal{M}(\Omega)^n)^{|\Theta|} : \sum_{\theta \in \Theta} \theta \, \mathrm{div}\,(\omega_\theta) = \mu_+ - \mu_- \right\}. \tag{4.2.21}$$

Each scalar flux ω_θ represents a potential combination of masses θ. It makes the joint fluxes more explicit, provides another intuitive view on the multimaterial transport problem, and is

also practical from a numerical perspective. We refer to [25, Section 4.2] for a more detailed study of this alternative formulation and numerical examples.

4.3 Problems with single topology

4.3.1 Setting and vertex optimization problem

A challenging aspect of the multimaterial transport problem is the exponentially large number of potential topologies of the optimal network. Therefore, we defer the problem of identifying the optimal network topology for now and consider in this section special problem instances where only a single topology is admissible to gain some insight in the remaining problem of finding the optimal location of network vertices. We show that this can be re-written as a finite-dimensional convex optimization problem. Such single-topology problems can be designed by suitable choice of sources and sinks, as well as the function h via the choice of prototypical material combinations Θ. We now outline this construction.

Definition 4.3.1 (Single-topology setting)**.** A *single-topology* multimaterial transport problem is specified by the following components:

(i) Initial and final distributions are given by

$$\mu_+ = \sum_{i=1}^{m} e_i \, \delta_{x_i}, \qquad \mu_- = 1_m \delta_y, \tag{4.3.1}$$

i.e. the source consists of m distinct Dirac measures of different materials i (e_i is the canonical i-th basis vector in $\mathbb{R}^m$), located at positions $x_i \in \Omega$, that all move to a common sink at $y \in \Omega$ (1_m being the vector with all entries 1).

(ii) The network topology is encoded by an abstract directed graph (V, E) with $V = \{1, \dots, K\}$ where $K \geq m + 1$ and each vertex has an associated position $z_i \in \Omega$. Vertices $1, \dots, m$ and K are *fixed vertices* and their positions are given by $z_i = x_i$ for $i = 1, \dots, m$ and $z_K = y$. The positions of the other *free vertices* will have to be determined by optimization. When $(i, j) \in E$, then all mass from z_i will flow to z_j, and so the graph (V, E) describes the gradual merging of materials from the sources towards the sink. In particular K is the unique root with no outgoing edge, all other vertices have precisely one outgoing edge, and the set of leaves with no incoming edges is $\{1, \dots, m\}$. A simple example for such a graph is shown in Figure 4.7.

(iii) We denote by $\mathrm{ch}(i) = \{j \in V : (j, i) \in E\}$ the children of i and by $\mathrm{pa}(i)$ the unique parent for $i \in V \setminus \{K\}$ for which one has $(i, \mathrm{pa}(i)) \in E$.

(iv) For $(i, j) \in E$ we denote by $\theta_{(i,j)}$ the material vector that is flowing on the edge (i, j). By construction, on the edges emerging from the source vertices, the material vector is the corresponding unit vector. At each free vertex, the material vector of the outgoing edge must be the sum of the material vectors of the incoming edges. And the sum of the material

vectors that are incoming on the sink/root node must be 1_m. This means, we have

$$\begin{aligned}\theta_{(i,\mathrm{pa}(i))} &= e_i \quad \text{for } i = 1,\dots,m,\\ \theta_{(i,\mathrm{pa}(i))} &= \sum_{k\in\mathrm{ch}(i)} \theta_{(k,i)} \quad \text{for} \quad i = m+1,\dots,K-1,\\ 1_m &= \sum_{j\in\mathrm{ch}(K)} \theta_{(j,K)}. \end{aligned} \tag{4.3.2}$$

So the material combinations are fixed by the topology E and independent of the vertex positions.

(v) Let $\Theta := \{\theta_e | e \in E\}$ and associate with each material combination θ_e, $e \in E$ a cost coefficient c_e for the transport per unit length. Together they induce the unit transport cost function h, as described in Section 4.2. The choice of coefficients c_e is arbitrary except for being strictly positive, which together with finiteness of E implies that (4.2.5) holds.

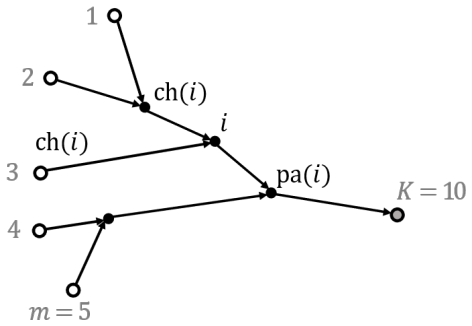

Figure 4.7: A network with fixed topology, sources represented by white circles, sink represented by grey-filled circle, free vertices represented by black-filled circle.

We will show that minimizing over the position of the free vertices is equivalent to solving the flow problem (4.2.1).

Definition 4.3.2 (Primal and dual vertex optimization problem)**.**

(i) The primal vertex optimization problem is given by

$$\inf_{z\in\mathbb{R}^{K\times n}} \sum_{(i,j)\in E} c_{(i,j)}\|z_j - z_i\| + \sum_{i=1}^{m} \imath_{\{x_i\}}(z_i) + \imath_{\{y\}}(z_K), \tag{4.3.3}$$

where $z \in \mathbb{R}^{K\times n}$ denotes the collection of all vertex positions, $z_i \in \mathbb{R}^n$ denoting the position of vertex i. This means, we sum over the edges in the network the respective edge lengths weighted by the unit cost coefficients for the corresponding material combination. The positions of the fixed vertices are enforced by the indicator functions.

(ii) The corresponding dual problem is (sketch for duality given below)

$$\begin{aligned}\sup\Big\{ \sum_{i=1}^{m} x_i\cdot\phi_{(i,\mathrm{pa}(i))} - \sum_{j\in\mathrm{ch}(K)} y\cdot\phi_{(j,K)} \,\Big|\, (\phi_e)_{e\in E}\in\mathbb{R}^{|E|\times n},\\ |\phi_e| \le c_e \text{ for } e\in E,\ \sum_{j\in\mathrm{ch}(i)} \phi_{(j,i)} = \phi_{(i,\mathrm{pa}(i))} \text{ for all } i = m+1,\dots,K-1.\Big\}\end{aligned} \tag{4.3.4}$$

The dual variable $\phi \in \mathbb{R}^{|E|\times n}$ has one n-dimensional component per edge of the network, indexed by $(i,j) \in E$.

Problems (4.3.3) and (4.3.4) are finite-dimensional non-smooth convex optimization problems and can, for instance, be tackled with proximal splitting methods [45, 29].

Remark 4.3.3 (Sketch of duality)**.** Duality between (4.3.3) and (4.3.4) can again be obtained via Fenchel–Rockafellar duality (Theorem 2.2.6), similar to (4.2.12). Equation (4.3.3) can be brought into the canonical form by choosing

$$\begin{aligned} &F : \mathbb{R}^{|E|\times n} \to \mathbb{R}, && F(w) = \sum_{e\in E} c_e \|w_e\|, \\ &G : \mathbb{R}^{K\times n} \to \{0,\infty\}, && G(z) = \begin{cases} 0, & \text{if } z_i = x_i \text{ for } i = 1,\dots,m \text{ and } z_K = y, \\ +\infty, & \text{else,} \end{cases} \\ &A : \mathbb{R}^{K\times n} \to \mathbb{R}^{|E|\times n}, && (Az)_{(i,j)} = z_j - z_i \quad \text{for } (i,j) \in E. \end{aligned}$$

The function F is the sum of scaled norm terms that can be conjugated separately. The convex conjugate of the Euclidean norm is the indicator function of the unit ball $\imath_{\|\cdot\|\le 1}$, and so we obtain

$$F^*(w) = \sum_{e\in E} \imath_{\|\cdot\|\le 1}\left(\frac{w_e}{c_e}\right) = \begin{cases} 0, & \text{if } \|w_e\| \le c_e \quad \forall e \in E, \\ +\infty, & \text{else.} \end{cases}$$

The conjugate of G, an indicator of a singleton for fixed vertex positions and not depending on the free vertex positions, is

$$G^*(w) = \begin{cases} \sum_{i=1}^m w_i \cdot x_i + w_K \cdot y, & \text{if } w_j = 0 \text{ for } j = m+1,\dots,K-1, \\ +\infty, & \text{else.} \end{cases}$$

For the adjoint of A we find

$$(A^*\phi)_i = -\sum_{j:(i,j)\in E} \phi_{(i,j)} + \sum_{j:(j,i)\in E} \phi_{(j,i)} = -\phi_{(i,\mathrm{pa}(i))} + \sum_{j\in \mathrm{ch}(i)} \phi_{(j,i)}$$

with the convention that the first term is zero for $i = K$ where $\mathrm{pa}(i)$ is not defined. By Fenchel–Rockafellar theorem 2.2.6, the dual problem is then formally given by

$$\sup_{\phi\in\mathbb{R}^{|E|\times n}} -F^*(-\phi) - G^*(A^*\phi),$$

where we find that the F^* term yields the constraints $\|\phi_e\| \le c_e$ for $e \in E$. The G^* term yields the constraints $\phi_{(i,\mathrm{pa}(i))} = \sum_{j\in\mathrm{ch}(i)} \phi_{(j,i)}$ for $i = m+1,\dots,K-1$ and the terms

$$\sum_{i=1}^m \phi_{(i,\mathrm{pa}(i))} \cdot x_i - \sum_{j\in\mathrm{ch}(K)} \phi_{(j,K)} \cdot y,$$

which together yields (4.3.4). Strong duality holds because F and G satisfy the constraint qualifications on an affine subspace.

The primal-dual optimality conditions, analogous to (4.2.14), in this case yield:

$$A^*\phi \in \partial G(z), \qquad -\phi \in \partial F(Az),$$

where the former simply implies the dual constraint

$$(A^*\phi)_i = 0 \quad \Leftrightarrow \quad \phi_{(i,\mathrm{pa}(i))} = \sum_{j\in \mathrm{ch}(i)} \phi_{(j,i)} \text{ for } i = m+1,\dots,K-1, \tag{4.3.5}$$

and the latter becomes

$$-\phi_{(i,j)} \in c_{(i,j)} \cdot (\partial\|\cdot\|)((Az)_{(i,j)}) \quad \text{for } (i,j)\in E, \tag{4.3.6}$$

which implies the dual constraint $\|\phi_{(i,j)}\| \le c_{(i,j)}$, and in particular for $z_i \neq z_j$ that $\phi_{(i,j)} = c_{(i,j)}\cdot(z_i - z_j)/\|z_i - z_j\|$. This means that when an edge in the transport network is not degenerate (i.e. its start and endpoint are different), then the corresponding dual variable must be aligned against the flux in that edge and its length must be the material cost $c_{(i,j)}$.

We will show below that any configuration of vertex positions $z \in \mathbb{R}^{K\times n}$ with appropriate fixed vertex locations in (4.3.3) induces an admissible candidate for the primal multimaterial flux problem (4.2.1). Further, from any admissible dual candidate $\phi \in \mathbb{R}^{|E|\times n}$ in (4.3.4) one can construct an admissible candidate for the dual multimaterial problem (4.2.10) by identifying $\phi_{(i,j)}$ with $\theta_{(i,j)}^{\mathsf{T}}\nabla\varphi$, i.e. we assume that the dual potentials φ are linear functions with constant derivatives. This will imply the equivalence between the problems. Before turning to the general equivalence proof, we study an explicit example to gain some intuition.

Example 4.3.4. Consider the case $m = 3$, $K = 6$, with four fixed vertices $\{z_1 = x_1, z_2 = x_2, z_3 = x_3, z_6 = y\}$ and two free vertices $\{z_4, z_5\}$, with edges $E = \{(1,4),(2,4),(4,5),(3,5),(5,6)\}$, as illustrated in Figure 4.8.

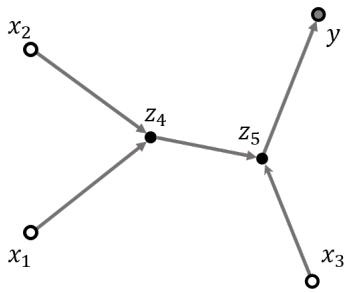

Figure 4.8: Vertex optimization: Sources denoted with x_i, sink denoted with y, free vertices denoted with z_j.

The corresponding collection of material vectors is given by

$$\Theta = \{\theta_{(1,4)} = e_1, \theta_{(2,4)} = e_2, \theta_{(4,5)} = e_1 + e_2, \theta_{(3,5)} = e_3, \theta_{(5,6)} = 1_3\}.$$

Flattening the collection of vertex positions $z \in \mathbb{R}^{K\times n}$ into a vector of dimension $K\cdot n$, the corresponding matrix representation of the operator A can be written as

$$A = \begin{bmatrix} -I & 0 & 0 & I & 0 & 0 \\ 0 & -I & 0 & I & 0 & 0 \\ 0 & 0 & 0 & -I & I & 0 \\ 0 & 0 & -I & 0 & I & 0 \\ 0 & 0 & 0 & 0 & -I & I \end{bmatrix},$$

where $I \in \mathbb{R}^{n\times n}$ is the identity matrix and 0 here denotes the $n \times n$ zero matrix. The primal problem (4.3.3) for this graph becomes

$$\inf_{z_4,z_5} \left[c_{(1,4)} \|z_4 - x_1\| + c_{(2,4)} \|z_4 - x_2\| + c_{(4,5)} \|z_5 - z_4\| + c_{(3,5)} \|z_5 - x_3\| + c_{(5,6)} \|y - z_5\| \right],$$

and the dual problem (4.3.4) is to maximize the objective function

$$\sup_{\phi\in\mathbb{R}^{|E|\times n}} \left[\sum_{i=1}^{m} \langle x_i, \phi_{(i,\mathrm{pa}(i))} \rangle - \langle y, \phi_{(5,6)} \rangle \right] \tag{4.3.7}$$

under the constraints

$$\phi_{(1,4)} + \phi_{(2,4)} - \phi_{(4,5)} = 0, \qquad \phi_{(4,5)} + \phi_{(3,5)} - \phi_{(5,6)} = 0 \tag{4.3.8}$$

and

$$\|\phi_{(i,j)}\| \le c_{(i,j)} \quad \text{for } (i,j) \in E. \tag{4.3.9}$$

In the discrete vertex problem (4.3.3) we have not constrained the free vertices to lie in Ω. This allowed for a simpler dualization. The following Lemma shows that if all fixed vertices lie in Ω and the latter is convex, this restriction is not required. The Lemma will also be convenient in the rest of the chapter.

Lemma 4.3.5. *Primal optimal free vertex locations in (4.3.3) lie in the convex hull of the fixed vertices.*

Proof. The coefficients c_e, $e \in E$, in (4.3.3) are strictly positive, and each $(Az)_e$ is the difference of two vertex positions. Let C be the convex hull of the fixed vertices, and denote by $P_C : \mathbb{R}^n \to C$ the projection onto C. Then for $x, y \in \mathbb{R}^n$ one has

$$\|P_C x - P_C y\| \le \|x - y\|$$

and for $x \in \mathbb{R}^n \setminus C$ and $y \in C$ one has

$$\|P_C x - \underbrace{P_C y}_{=y}\| < \|x - y\|.$$

Thus in (4.3.3), for admissible fixed vertex positions, if any one of the free vertices are not in C, the objective can be strictly decreased by projecting them onto C. □

4.3.2 Equivalence of vertex optimization and multimaterial transport problem

Proposition 4.3.6. *Let z and ϕ be primal and dual optimal in (4.3.3) and (4.3.4). Let*

$$\omega = \sum_{(i,j)\in E} \theta_{(i,j)} \otimes e_{(i,j)} \cdot \mathscr{H}^1 \llcorner l_{(i,j)}, \qquad \varphi_i(x) = \langle x, \phi_{(i,\mathrm{pa}(i))} \rangle \text{ for } i = 1, \dots, m.$$

Here for $(i,j) \in E$ we set

$$e_{(i,j)} = \begin{cases} \frac{z_j - z_i}{\|z_j - z_i\|} & \text{if } z_i \neq z_j, \\ 0 & \text{else,} \end{cases}$$

and $l_{(i,j)} \subset \mathbb{R}^n$ is the straight line segment between z_i and z_j. Then ω and φ are primal and dual optimal for (4.2.1) and (4.2.10).

Proof. We first show primal feasibility of ω by evaluating $\mathrm{Div}(\omega)$. We find

$$\begin{aligned}\mathrm{Div}(\omega) &= \sum_{(i,j)\in E} \theta_{(i,j)} \cdot \mathrm{div}(e_{(i,j)} \cdot \mathscr{H}^1 \llcorner l_{(i,j)}) = \sum_{(i,j)\in E} \theta_{(i,j)}(\delta_{z_i} - \delta_{z_j}) \\ &= \sum_{i=1}^{K} \left(\sum_{j:(i,j)\in E} \theta_{(i,j)} - \sum_{j:(j,i)\in E} \theta_{(j,i)} \right) \cdot \delta_{z_i} \\ &= \sum_{i=1}^{m} e_i \delta_{x_i} - 1_m \delta_y = \mu_+ - \mu_-.\end{aligned}$$

In the forth equality we use (4.3.2), which implies that in the sum the terms for $i = m+1, \dots, K-1$ are zero, for $i = 1, \dots, m$ there are no incoming edges, and for $i = K$ no outgoing edges.

For the objective we find

$$\int_{\mathbb{R}^n} H\left(\frac{\mathrm{d}\,\omega}{\mathrm{d}\,|\omega|}\right) \mathrm{d}\,|\omega| = \sum_{(i,j)\in E} H(\theta_{(i,j)} \otimes e_{(i,j)}) \cdot \|z_j - z_i\| = \sum_{(i,j)\in E} h(\theta_{(i,j)}) \|(Az)_{(i,j)}\| \le \sum_{(i,j)\in E} c_{(i,j)} \|(Az)_{(i,j)}\|,$$

where A as in Remark 4.3.3. Here we used (4.2.9) and the inequality $h(\theta_e) = \hat{h}^{**}(\theta_e) \le \hat{h}(\theta_e) = c_e$ for $e \in E$. From this we conclude that (4.2.1) $\le$ (4.3.3).

Now we consider the dual problems. Based on the tree structure (V, E) introduced in Definition 4.3.1, for $i = m+1, \dots, K-1$ denote by $\mathrm{desc}(i)$ the *descendants* of i, i.e. the vertices $j \in \{1, \dots, m\}$ for which the unique path to K passes through i. Condition (4.3.2) then implies that $\theta_{(i,\mathrm{pa}(i))} = \sum_{j\in\mathrm{desc}(i)} e_j$ for $i = m+1, \dots, K-1$, and in particular $(\theta_{(i,\mathrm{pa}(i))})_j = 1$ if and only if $j \in \mathrm{desc}(i)$. In a similar way, for $i = m+1, \dots, K-1$, recursive application of the constraint $\sum_{j\in\mathrm{ch}(i)} \phi_{(j,i)} = \phi_{(i,\mathrm{pa}(i))}$ of (4.3.4) yields $\phi_{(i,\mathrm{pa}(i))} = \sum_{j\in\mathrm{desc}(i)} \phi_{(j,\mathrm{pa}(j))}$. With this we get for the dual candidate as constructed above and $(i, \mathrm{pa}(i)) \in E$ that

$$\theta_{(i,\mathrm{pa}(i))}^{\intercal} D\varphi(x) = \sum_{j=1}^{m} (\theta_{(i,\mathrm{pa}(i))})_j \phi_{(j,\mathrm{pa}(j))} = \sum_{j\in\mathrm{desc}(i)} \phi_{(j,\mathrm{pa}(j))} = \phi_{(i,\mathrm{pa}(i))}.$$

Therefore, the constraint $\|\phi_{(i,\mathrm{pa}(i))}\| \le c_{(i,\mathrm{pa}(i))}$ in (4.3.4) implies the constraint for the dual flow problem (4.2.11). As for the dual objective, one gets in (4.2.10) for the above φ that

$$\begin{aligned}\int \varphi \,\mathrm{d}(\mu_+ - \mu_-) &= \sum_{i=1}^{m} \int \varphi_i \mathrm{d}(\delta_{x_i} - \delta_y) = \sum_{i=1}^{m} \langle \phi_{(i,\mathrm{pa}(i))}, x_i - y \rangle \\ &= \sum_{i=1}^{m} \langle \phi_{(i,\mathrm{pa}(i))}, x_i \rangle - \sum_{j\in\mathrm{ch}(K)} \langle \phi_{(j,K)}, y \rangle.\end{aligned}$$

With this we find (4.2.10) $\ge$ (4.3.4), and using strong duality for both problems we eventually get equality between (4.2.1) = (4.2.10) = (4.3.3) = (4.3.4) and thus optimality of the candidates constructed above. □

4.3.3 Example: Three vertex problem

As an instructive and popular example of a problem with only a single admissible topology we will now study the three vertex problem and in particular the explicit solution (or characterization of solutions) of the corresponding vertex optimization problem, based on the chosen locations of sources and sink, and of the material costs.

For this concrete case, a slightly more explicit notation will be convenient. Consider two sources $z_1 = A = x_1$, $z_2 = B$ and a sink $z_4 = C$ in Ω, with a single free vertex $z_3 = S$, i.e. $m = 2$ and $K = 4$, and $E = \{(1,3),(2,3),(3,4)\}$. For simplicity, we assume that A, B and C are distinct. Degenerate special cases can be solved separately. According to Lemma 4.3.5, point S lies in the convex hull of points A, B, C, and the convex hull of 3 distinct points is contained in a 2-dimensional plane; therefore we restrict our discussion in this section to $A, B, C, S \in \mathbb{R}^2$ without the loss of generality. For more compact notation we will refer to the cost coefficients by

$$a = c_{(1,3)}, \qquad b = c_{(2,3)}, \qquad c = c_{(3,4)}.$$

As above, we assume that all three coefficients are strictly positive. Then the primal vertex optimization problem (4.3.3) can be written as

$$\inf_{S \in \mathbb{R}^n} a \cdot \|A - S\| + b \cdot \|B - S\| + c \cdot \|C - S\|,$$

where again $S = z_3$ denotes the single free vertex. The dual corresponding dual problem (4.3.4) can be written as

$$\sup \Big\{ \langle A, \phi_a \rangle + \langle B, \phi_b \rangle - \langle C, \phi_c \rangle \;\Big|\; \phi_a, \phi_b, \phi_c \in \mathbb{R}^n : \phi_a + \phi_b = \phi_c,\ \|\phi_x\| \le x \text{ for } x \in \{a, b, c\} \Big\}.$$

Note that the constraint in principle allows to eliminate one of the ϕ_x, for instance ϕ_c, which would correspond to the continuous dual problems (4.2.10), see also Prop. 4.3.6. In particular, the latter proposition suggests that we can identify ϕ_x with the gradient of the corresponding (sum of) dual potential(s), which in turn can be assume to be globally linear, and we will do so in the following discussion.

In addition, the primal-dual optimality condition implies for the optimal S and $(\phi_x)_{x \in \{a,b,c\}}$ that

$$\begin{aligned} S \neq A \quad &\Rightarrow \quad \phi_a = a \cdot \frac{A - S}{\|A - S\|}, \\ S \neq B \quad &\Rightarrow \quad \phi_b = b \cdot \frac{B - S}{\|B - S\|}, \\ S \neq C \quad &\Rightarrow \quad \phi_c = c \cdot \frac{S - C}{\|C - S\|}. \end{aligned} \tag{4.3.10}$$

In addition, Lemma 4.3.5 implies that S must lie in the convex hull of $\{A, B, C\}$. When $S \notin \{A, B, C\}$, then the constraint $\phi_a + \phi_b = \phi_c$ and (4.3.10) fix the relative angles at which the segments AS, BS and CS meet in S via the "cost triangle" with lengths a, b, c, see Figures 4.6 and 4.9. In particular none of the angles between any two segments can be π, implying that S cannot lie on the edges of the triangle ABC.

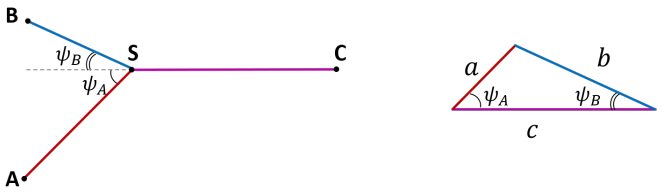

Figure 4.9: Angles at the free vertex S and the corresponding cost triangle.

This means that one of four cases must occur: S coincides with one of the three vertices, A, B, C, or S lies in the interior of the convex hull, see Figure 4.10.

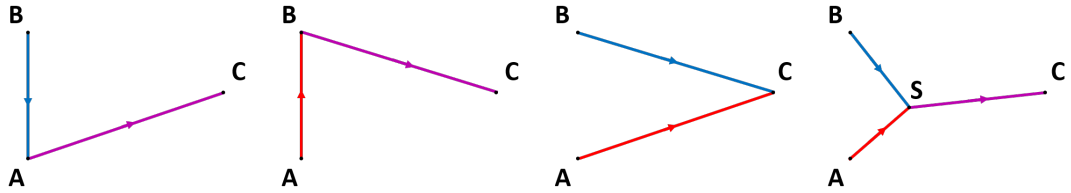

Figure 4.10: Possible types of graph (left to right): L-graph ($S = A$), L-graph ($S = B$), V-graph ($S = C$), Y-graph.

It is easy to see that when a point S can be found, under which the three segments have the correct relative angles, it will be a primal solution and the segment orientations induce the corresponding dual solution. In the following we discuss under which conditions such a point S does or does not exist.

Problem normalization.

Since three vertices A, B, C always lie in a two-dimensional plane and since the optimal S will lie in the convex hull of A, B, C, it is sufficient to consider the case $n = 2$ in the following. Further, the optimization problem is invariant under isometries of $\mathbb{R}^2$ such as translations, rotations, and reflections; and also under scaling in the sense that the solution of the transformed problem is the transformation of the original problem. We can therefore consider the case where $A = (0, 0)^\intercal$, $B = (0, 1)^\intercal$ and C having a non-negative first coordinate.

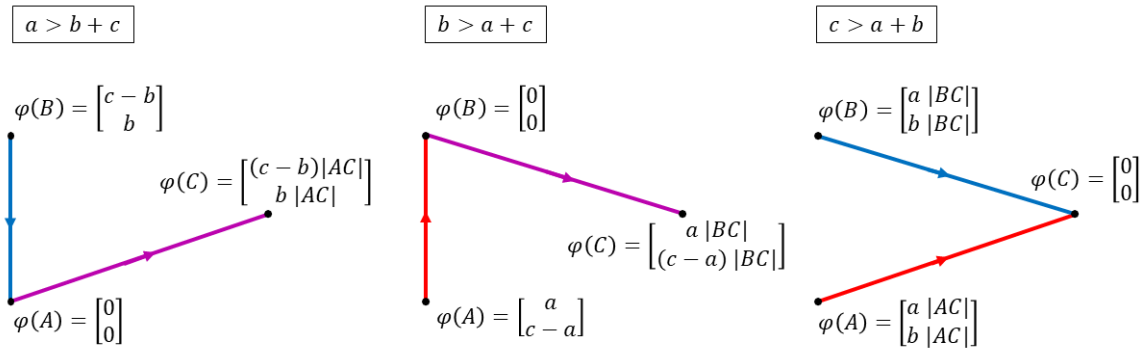

Figure 4.11: Calibrations for the cases when the cost triangle does not exist.

Cost constraints. A point S with the appropriate angles cannot exist when the angles do

not exist, i.e. when one cannot form a triangle with edge lengths a, b, c because one of them is longer than the sum of the other two. Assume $c > a + b$ (the other cases are symmetric). Then for any ϕ_a, ϕ_b with $\|\phi_x\| \leq x$, $x \in \{a, b\}$ one automatically has for $\phi_c = \phi_a + \phi_b$ that $\|\phi_c\| \leq c$. So flows a and b can move independently and be calibrated independently by the dual potentials without having to worry about the constraint for ϕ_c. The primal solution will therefore be the V-graph. This and the other two cases are illustrated in Figure 4.11.

Domain constraints. The other situation when a point S with the suitable angles does not exist is, informally, when the triangle is too small or oblong. For a normalized problem (see above), the set of points S under which A and B appear under the correct relative angle lies on a circle (by the inscribed angle theorem, see Figure 4.12). Fixing S on that circle, the set of points C for which this point S is primal optimal then lie on a ray emerging from S with a prescribed angle. Thus, with A and B fixed, the set of all C for which a S with the proper angles exists is the union of all these rays. This set is bounded by the circle and the two extremal rays when S coincides with either A or B. An elementary geometric consideration yields that these angles are given by ψ_B and ψ_A respectively (see Figure 4.9 for the notation of the angles ψ_A, ψ_B).

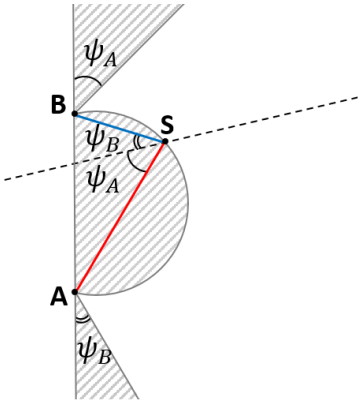

Figure 4.12: Domain constraints: if the angles at S have to agree with the cost triangle angles, then C cannot lie in the dashed regions.

One can then verify that when C lies within the circle, that the optimal solution will again be the V-graph where a and b flows can be calibrated independently without violating the ϕ_c-constraint (intuitively, because now ϕ_a and ϕ_b meet at an angle larger than the one prescribed by the cost triangle, and thus the resulting ϕ_c edge is shorter than c). Likewise, when C lies beyond one of the two rays, one of the two L-graphs will turn out to be optimal.

4.4 Numerical approximation

This section is devoted to the numerical approximation of solutions of multimaterial transport problem. Namely, we consider two different discretization options with dedicated numerical solvers, show some numerical experiments and discuss the observations.

4.4.1 Graph optimization

The fact that for discrete sources and sinks optimal solutions are supported on graphs [92] and the rich structure of the primal-dual optimality conditions suggest numerical approximation

schemes that directly optimize over graphs. While this is a simple problem for a fixed topology where only the locations of the vertices need to be found (as discussed in Section 4.3), interesting problems usually exhibit a number of potential topologies that is exponential in the number of sources and sinks. Here we consider a numerical scheme of the described type.

The scheme consists in alternating optimization: First we optimize over the flows on a fixed graph (the topology and the vertex positions are fixed), and then, for a fixed flow on the graph, we optimizer over the positions of the free vertices of the graph. While both optimization problems separately are convex, the joint problem (i.e., optimization over flows and vertex positions) is a non-convex discretization of the original multimaterial transport problem. Therefore, in general we cannot expect global optimality of the solutions we obtain from this scheme. As a heuristic remedy, in both steps with add regularization: in the flow updates we add quadratic regularization of the flows to encourage spread of the flow over multiple edges and thus to increase the chances of "observing" more potential topologies. in the vertex position updates we add a quadratic fidelity to the previous step to avoid that the graph collapses to a degenerate configuration immediately. We now describe the two alternating steps in more detail.

The first component of the scheme is a numerical solver for a regularized multimaterial minimum-cost flow problem on a given graph (see for example [2, Chapter 17] or [18, Section 8.3]; compare with the flow minimization problem in Section 4.2.5): Let (V, E) be a connected directed graph, let m be the number of materials and $\Theta \subset \mathbb{R}^m$ be the finite set of material combinations with given cost coefficients $c_\theta \in \mathbb{R}_{++}$ for $\theta \in \Theta$. Let $\mu_+, \mu_- \in \mathbb{R}_+^{|V|\times m}$ be the matrices associated with the vertices of the graph V giving respectively the initial and the final distribution of masses for each material such that $\sum_i (\mu_+)_{ik} = \sum_j (\mu_-)_{jk} = 1$, $k = 1, \dots, m$. The regularized multicommodity flow problem is as follows:

$$\min \left\{ \sum_{\theta\in\Theta} \sum_{e\in E} c_\theta \; h_\varepsilon(\omega_{e,\theta}) \; l_e \;\middle|\; \omega \in \mathbb{R}^{|E|\times|\Theta|} : \sum_{\theta\in\Theta} \theta \; \mathrm{div}(\omega_\theta) = \mu_+ - \mu_- \right\}, \tag{4.4.1}$$

where l_e is the length of edge e, $h_\varepsilon : \mathbb{R} \to \mathbb{R}_+$ is the Huber function

$$h_\varepsilon(z) = \begin{cases} |z| - \varepsilon/2, & \text{if } |z| \geq \varepsilon, \\ \frac{z^2}{2\varepsilon}, & \text{else,} \end{cases}$$

and $\mathrm{div} : \mathbb{R}^{|E|} \mapsto \mathbb{R}^{|V|}$ denotes the graph divergence operator.

We introduce the regularization to "spread" the flow over the edges, as the quadratic term of the Huber function encourages the transport of smaller amounts of mass through a larger subset of edges. For $\varepsilon = 0$ one recovers the original problem. We will use gradual reduction of ε as an annealing technique.

We use Chambolle–Pock proximal splitting method [38] to solve problem (4.4.1), as the necessary proximal operators $\mathrm{prox}_{\sigma,G}$ and prox_{τ,F^*} can be evaluated efficiently. Namely, by selecting

$$\begin{aligned} G &: \mathbb{R}^{|E|\times|\Theta|} \to \mathbb{R}_+, & G(\omega) &= \sum_{\theta\in\Theta} \sum_{e\in E} c_\theta \; h_\varepsilon(\omega_{e,\theta}), \\ F &: \mathbb{R}^{|V|\times m} \to \{0, \infty\}, & F(v) &= \imath_{\cdot=\mu_+-\mu_-}(v), \\ A &: \mathbb{R}^{|E|\times|\Theta|} \to \mathbb{R}^{|V|\times m}, & A\omega &= \sum_{\theta\in\Theta} \theta \; \mathrm{div}(\omega_\theta), \end{aligned}$$

we find that problem (4.4.1) can be written as $\min_{\omega\in\mathbb{R}^{|E|\times|\Theta|}} G(\omega)+F(A\omega)$. For the proximal operators (Definition 2.2.2), we get

$$(\mathrm{prox}_{\sigma,G}(\omega))_{e,\theta}=\begin{cases}\omega_{e,\theta}-\sigma c_\theta|e|, & \text{if } \omega_{e,\theta}>\varepsilon+\sigma c_\theta|e|,\\ \omega_{e,\theta}+\sigma c_\theta|e|, & \text{if } \omega_{e,\theta}<-\varepsilon-\sigma c_\theta|e|,\\ \frac{\varepsilon}{\varepsilon+\sigma c_\theta|e|}\omega_{e,\theta}, & \text{otherwise,}\end{cases}\qquad \forall e\in E,\theta\in\Theta,$$

$$\mathrm{prox}_{\tau,F^*}(v)=v-\tau(\mu_+-\mu_-).$$

The second component of the alternating optimization scheme is the vertex optimization. We formulate the problem similar to Definition 4.3.1 as minimization over the free vertex positions while keeping the flow on edges ω fixed. Let $z\in\mathbb{R}^{k_z\times n}$ denote the free vertex positions (with $k_z\in\mathbb{N}$ standing for the number of free vertices) and $x\in\mathbb{R}^{(|V|-k_z)\times n}$ denote the fixed vertex positions. Let $A^{free}:\mathbb{R}^{k_z\times n}\mapsto\mathbb{R}^{|E|\times n}$ and $A^{fixed}:\mathbb{R}^{(|V|-k_z)\times n}\mapsto\mathbb{R}^{|E|\times n}$ be two operators which give for each edge the adjacent vertex positions, with the convention that $(A^{free}z)_e=z_i$ if edge e ends in vertex i and $(A^{free}z)_e=-z_i$ if edge e starts in vertex i (and similar for operator A^{fixed}). Let $w\in\mathbb{R}_+^{|E|}$ be the set of non-negative weights; we choose $w_e=\sum_{\theta\in\Theta}c_\theta\,h_\varepsilon(\omega_{e,\theta})$. Then the vertex optimization problem can be formulated as follows:

$$\min\left\{\sum_{e\in E}w_e\left\|(A^{free}z)_e+(A^{fixed}x)_e\right\| \;\middle|\; z\in\mathbb{R}^{k_z\times n}\right\}.$$

For instance, for the problem presented in Example 4.3.4, with $n=2$, for the 2 free vertices $z\in\mathbb{R}^{2\times2}=\{z_4,z_5\}$, four fixed vertices $x\in\mathbb{R}^{4\times2}=\{x_1,x_2,x_3,y\}$, and the edges defined as before, the operators A^{free} and A^{fixed} are as follows:

$$A^{free}=\begin{bmatrix}I & 0\\ I & 0\\ -I & I\\ 0 & I\\ 0 & -I\end{bmatrix},\qquad A^{fixed}=\begin{bmatrix}-I & 0 & 0 & 0\\ 0 & -I & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & -I & 0\\ 0 & 0 & 0 & I\end{bmatrix},$$

where $I\in\mathbb{R}^{2\times2}$ is the identity matrix and $0\in\mathbb{R}^{2\times2}$ is the zero matrix.

We also add a quadratic regularization term to the vertex optimization problem to introduce damping into the system (i.e. to prevent the vertices from moving too far from their original positions):

$$\min\left\{\sum_{e\in E}w_e\left\|(A^{free}z)_e+(A^{fixed}x)_e\right\|+\lambda\sum_{i=1}^{k_z}\|z_i-\bar z_i\| \;\middle|\; z\in\mathbb{R}^{k_z\times n}\right\},\tag{4.4.2}$$

where $\lambda\in\mathbb{R}_{++}$ is the regularization parameter and $\bar z_i$ is the constant denoting the original location of point z_i.

We also use the Chambolle–Pock method for this problem, and select

$$\begin{aligned}&G:\mathbb{R}^{k_z\times n}\to\mathbb{R}_+, && G(z)=\lambda\sum_{i=1}^{k_z}\|z_i-\bar z_i\|,\\ &F:\mathbb{R}^{|E|}\to\mathbb{R}_+, && F(v)=\sum_{e\in E}w_e\left\|v_e+(A^{fixed}x)_e\right\|,\\ &A:\mathbb{R}^{k_z\times n}\to\mathbb{R}^{|E|}, && Az=A^{free}z,\end{aligned}$$

for which we obtain

$$\operatorname{prox}_{\sigma,G}(z) = \frac{1}{1+2\lambda\sigma}\,(z + 2\lambda\sigma\bar{z}),$$
$$(\operatorname{prox}_{\tau,F^*}(v))_e = \begin{cases} v_e + \tau(A^{fixed}x)_e, & \text{if } \left\|v_e + \tau(A^{fixed}x)_e\right\| \le w_e, \\ \frac{v_e+\tau(A^{fixed}x)_e}{\left\|v_e+\tau(A^{fixed}x)_e\right\|} w_e, & \text{otherwise.} \end{cases}.$$

We perform the numerical optimization as follows. Given the initial data (distribution of sources and sinks and cost coefficients for material combinations), we first obtain an initial graph by triangulating the convex hull of the support of sources and sinks. We then select the parameters of the algorithm (list of values of ε, regularization parameter λ, parameters of the Chambolle–Pock algorithm) and stopping criteria thresholds. For the stopping criteria, we use a standard combination of feasibility and optimality conditions. We then proceed to perform a preset number of alternating optimization steps (edge flows and vertex positions) for the largest selected value of ε, and then pass the resulting graph with a flow on its edges into the alternating optimization for the next ε. After going through all the ε in the list, we perform one more round of alternating minimization steps with unregularized flow problem. Here we show some examples of the application of the numerical scheme we discussed.

Example 4.4.1 (2 sources, 1 sink)**.** Consider the problem described in Section 4.3.3 with two material types with a single Dirac source for each and a common sink. We place the sources at $[0,0]^\intercal$ and $[0,1]^\intercal$ for the materials 1 and 2 respectively, and the sink at $[2,0.5]$. We select the cost coefficients for the individual flows as $a = 1$, $b = 0.5$ and for the joint flow as $c = 1.3$ (see Section 4.3.3 for the notation). The initial graph generated by area triangulation is shown in Figure 4.13.

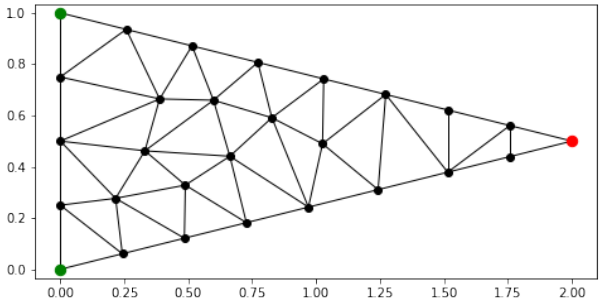

Figure 4.13: Initial graph used in Example 4.4.1 with sources in green and common sink in red.

We use the numerical scheme described above to solve the problem; the evolution of numerical approximation is shown in Figure 4.14: Each row shows the flows of the individual materials through the graph (left for material 1, right for material 2) after the alternating optimization for the given ε, the mass is shown with the α-channel; values p and d are the primal and dual score respectively.

We compute the angles ψ_A and ψ_B at the Steiner point analytically (as in Section 4.3.3), and

compare them to the values computed numerically. The theoretical values are

$$\psi_A = \arccos\left(\frac{a^2+c^2-b^2}{2\,a\,c}\right) \approx 0.352648, \qquad \psi_B = \arccos\left(\frac{b^2+c^2-a^2}{2\,b\,c}\right) \approx 0.762550,$$

while from the numerical experiment we obtain

$$\widetilde{\psi_A} \approx 0.352878, \qquad \widetilde{\psi_B} \approx 0.760649,$$

which shows a good correspondence with the theoretical values (with 0.07% and 0.25% error respectively).

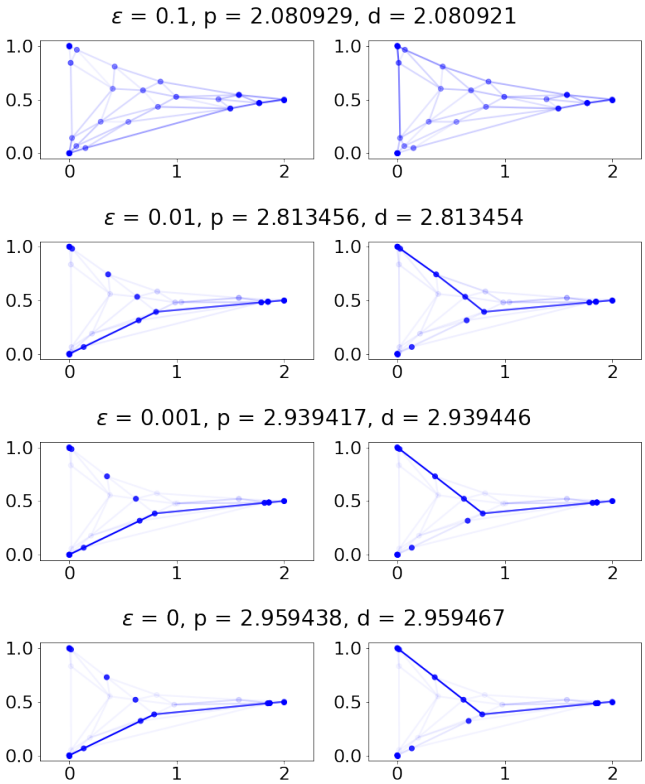

Figure 4.14: Numerical approximations at the end of ε-scaling iteration for the 2 materials.

Example 4.4.2 (3 sources, 1 sink)**.** Next we consider a similar problem but with 3 materials with separate sources and common sink. Sources of the materials $1, 2, 3$ are located respectively at $[0, 0]^\intercal, [0, 0.7]^\intercal, [0, 1]^\intercal$, and the common sink is located at $[2, 0.4]^\intercal$. The cost coefficients for single material flows is set to 1, for any combination of 2 materials to 1.4, and of all 3 materials to 1.9. The initial graph (obtained by domain triangulation) is shown in Figure 4.15. The results are shown in Figure 4.16.

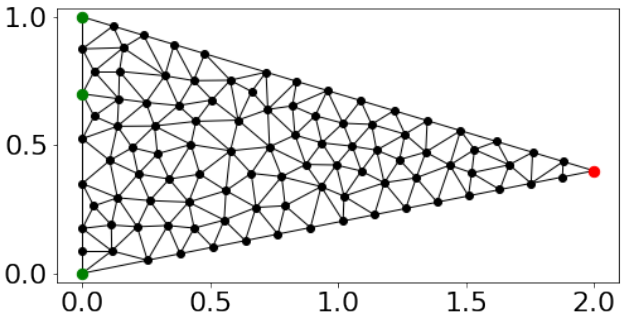

Figure 4.15: Initial graph used in Example 4.4.2 with sources in green and common sink in red.

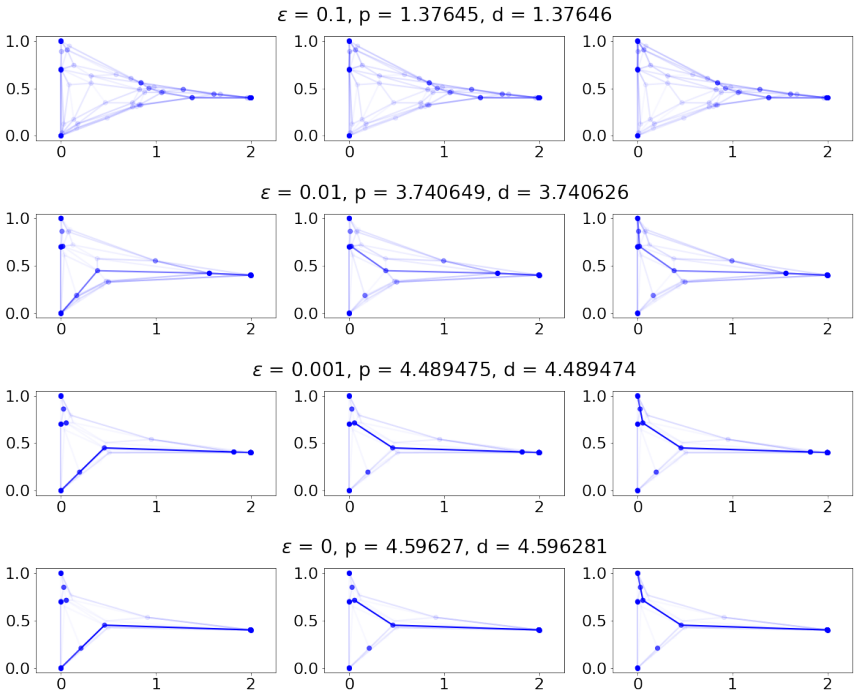

Figure 4.16: Numerical approximations at the end of ε-scaling iteration for the 3 materials.

We can observe the solution of expected type: flows of materials 2 and 3 meet in a Steiner point (which does not coincide with any of the fixed points), the joint flow with materials 2 and 3 then meets the single material flow of type 1 to form another non-degenerate Steiner point, and from there the joint flow of all 3 materials goes to the common sink.

Example 4.4.3. Next we consider an example very similar to Example 4.4.2, with only a minor change: The location of the common sink will now be at $[2, 0.3]^\intercal$. The evolution of approximations is shown in Figure 4.17. We observe here a different topology of the solution, namely, the Steiner point of materials 2 and 3 is now degenerate (as it coincides with the source of material 2).

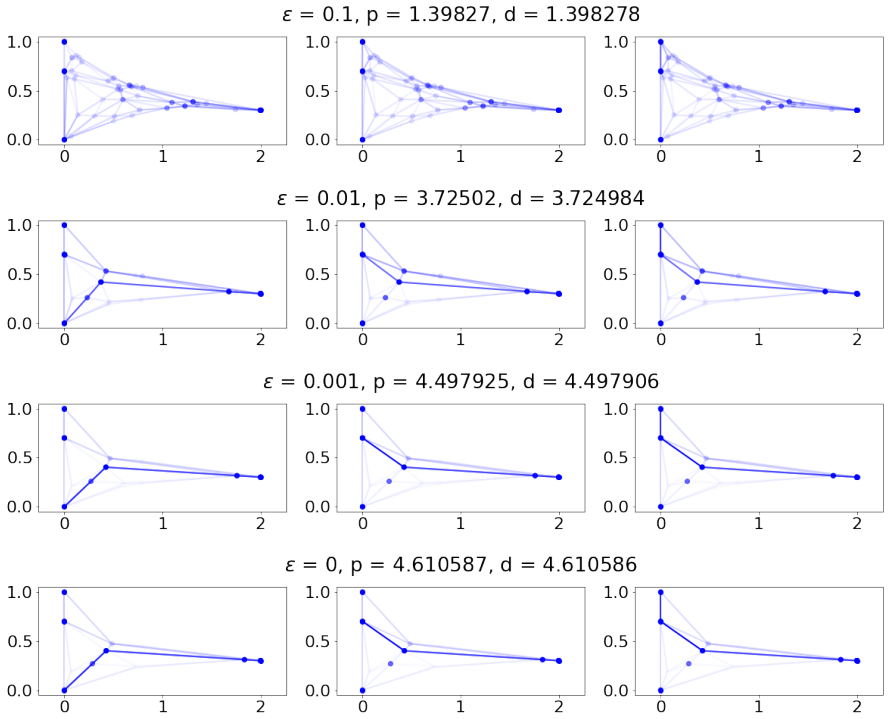

Figure 4.17: Numerical approximations at the end of ε-scaling iteration for the 3 materials.

However, by adjusting just one parameter of the numerical scheme (namely, changing the vertex optimization regularization parameter λ from 0.01 to 0.1), one can obtain a different solution, shown in Figure 4.18. We note that in this solution (after the last unregularized step) both of the Steiner points are non-degenerate, and the primal value here ($p = 4.606984$) is lower than in the previous solution ($p = 4.610587$). As discussed above, this is a consequence of the non-convexity of the chosen problem discretization.

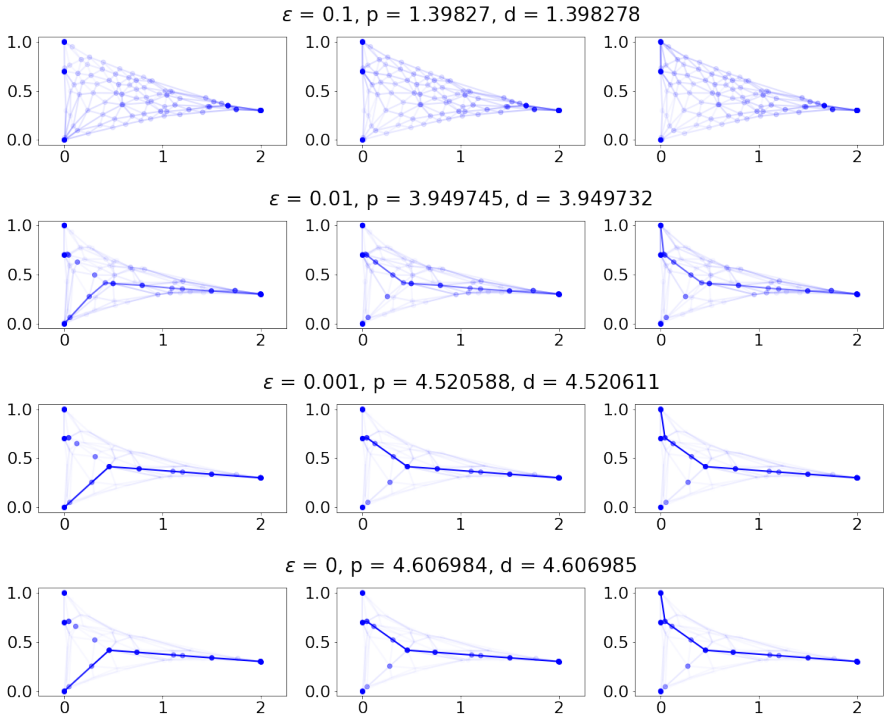

Figure 4.18: Numerical approximations at the end of ε-scaling iteration for the 3 materials: adjusted parameters of the numerical scheme.

The presented numerical experiments confirm that the developed graph optimization scheme does not guarantee global optimality of the approximate solutions even in simple cases. Similar behavior has been observed in other topology optimization approaches, see for example [25, 44].

4.4.2 Finite element discretization

As an alternative, one can apply a finite element discretization to the problem, typically with a weak divergence constraint on the fluxes. There is a huge body of literature on different discretization schemes non-convex branched transport and convex multimaterial transport problems with different choices for discretization, regularization, adaptive mesh refinement et cetera [93, 27].

This avoids explicit optimization over topologies as they can now be implicitly encoded as fluxes on the mesh. On the other hand, for a sharp resolution of the resulting network structures typically a very large number of mesh triangles is required as the weak divergence constraint induces a strong spatial blur. Here we consider the following approach.

Recalling the dual multimaterial problem (4.2.10) we relax the class of admissible dual functions from $\mathrm{C}^1(\Omega)^m$ to $\mathrm{C}^{0,1}(\Omega, \mathbb{R}^m)$ (see Remark 4.2.3). Assuming there is an underlying triangle mesh for the $\mathrm{spt}(\mu_+) \cup \mathrm{spt}(\mu_-)$ we discretize the dual problem on it and enforce the dual constraints on the elements of the mesh (i.e., on the faces of the triangles). We use the standard finite

element technique for obtaining the matrix which extracts from the values of the dual potentials at the nodes their gradients on the elements (see for example [64, Section 7.2.3]). Note that the dual problem obtained in this manner will be conformal to the original dual problem discussed in (4.2.10), as its solution will be a feasible candidate for the original problem (taking into account the relaxation to Lipschitz functions). However, the primal problem obtained by dualization of the conformal dual will not be conformal to the original primal problem ((4.2.1) or (4.2.21)), as the new primal will have the flow supported on the faces of the triangles, with only a weak notion of divergence constraint enforced, and so its solutions cannot directly be used as candidates for the original primal problem.

We use the Chambolle–Pock proximal splitting method for optimization and adaptive refinement of the mesh for improving the discretization accuracy. Namely, after computing the solutions to the conformal dual and non-conformal primal with the Chambolle–Pock method, we compute the divergence of the primal solution on the edges of the mesh and then refine the discretization in the vicinity of the edges with the highest absolute value of the weak divergence and repeat the process. For simplicity, we describe the case $n = 2$ (i.e., we work in $\mathbb{R}^2$), but with suitable generalizations the scheme can be extended to general n.

Here we show the application of this scheme to the problem from Example 4.4.2.

Example 4.4.4. We start by generating a simple triangulation of the convex hull of the support of the input measures. In this case, we only start with 2 triangles, as shown in Figure 4.19.

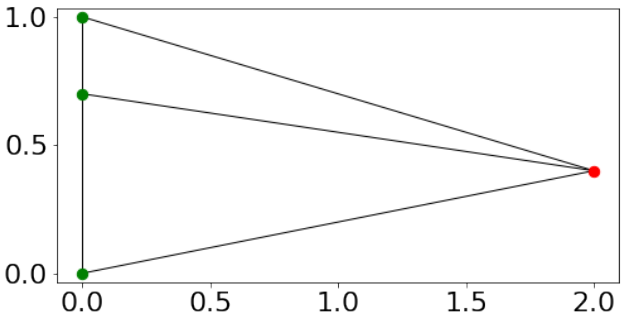

Figure 4.19: Initial mesh used in Example 4.4.4 with sources in green and common sink in red.

We then use the numeric scheme as described above and show the evolution of the primal and dual solutions, as well as the computed divergence residual and the triangles selected for refinement. The evolution is presented in Figures 4.20–4.24. The first row shows the absolute value of the primal flow for each of the 3 individual materials, the second row shows the dual potentials corresponding to each of the materials interpolated into the faces of the triangles, and the third row shows the divergence error computed on edges, the error propagated onto the faces of the elements, and the triangles selected for refinement on the next step.

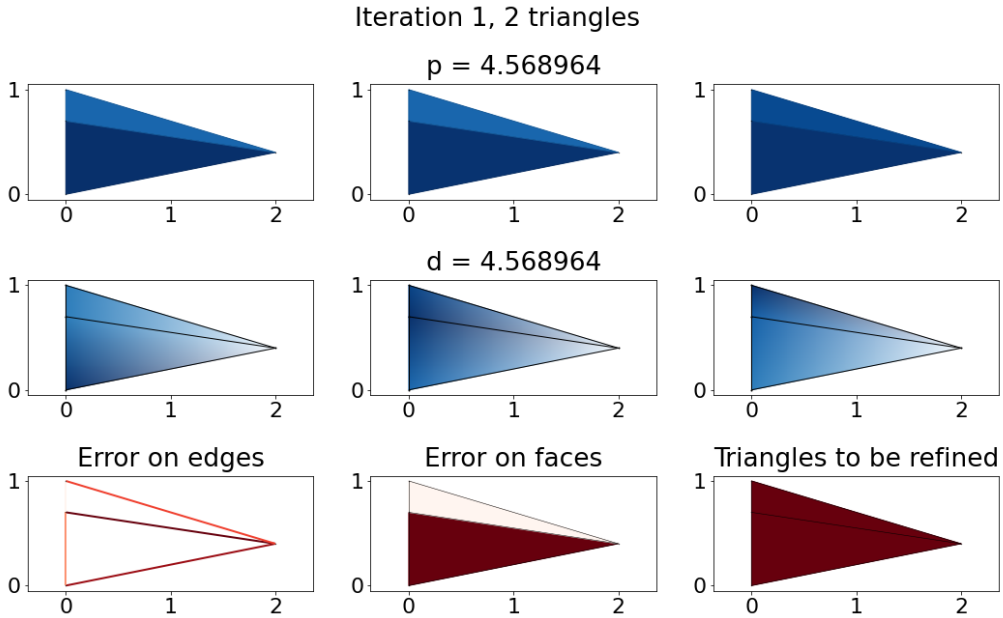

Figure 4.20: Numerical approximation of the non-conformal primal and conformal dual with divergence residual.

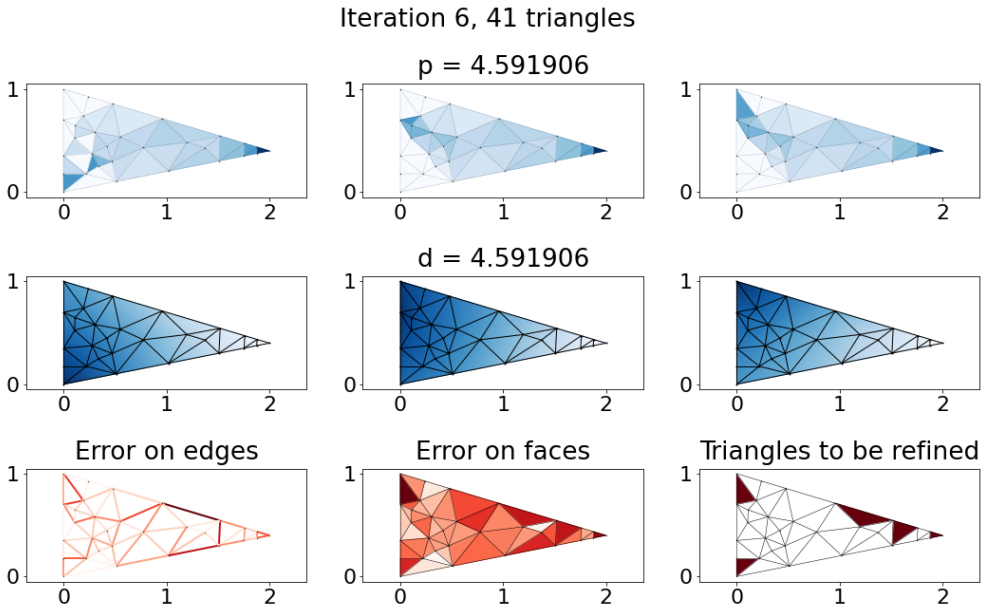

Figure 4.21: Numerical approximation of the non-conformal primal and conformal dual with divergence residual.

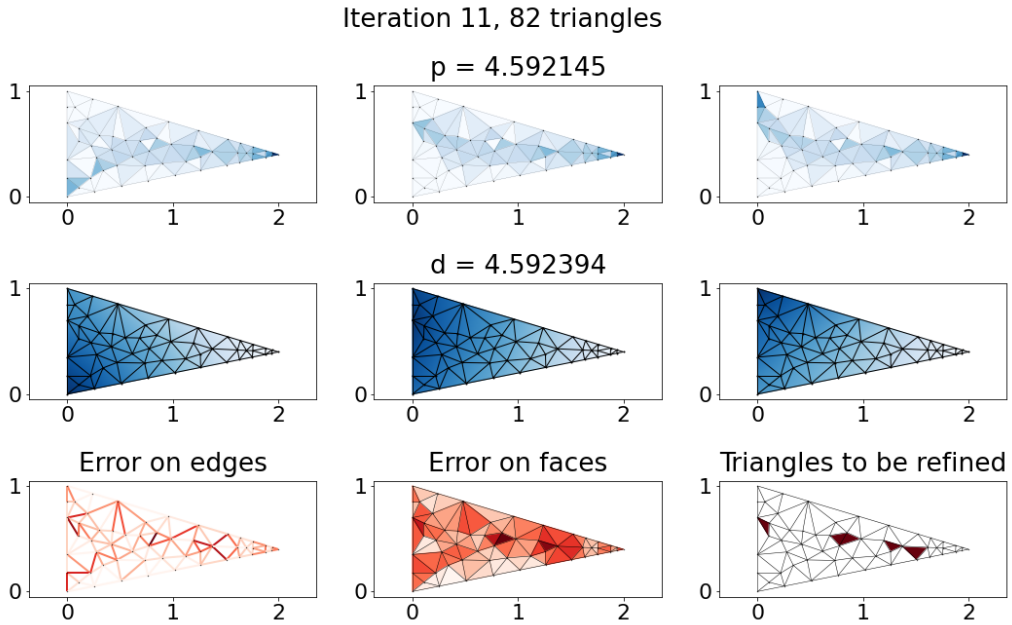

Figure 4.22: Numerical approximation of the non-conformal primal and conformal dual with divergence residual.

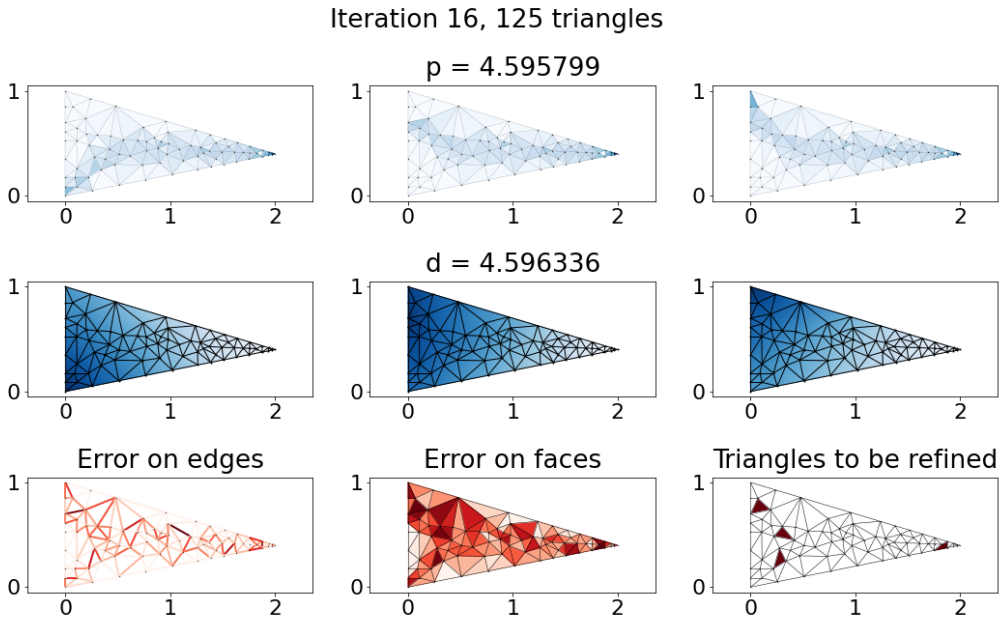

Figure 4.23: Numerical approximation of the non-conformal primal and conformal dual with divergence residual.

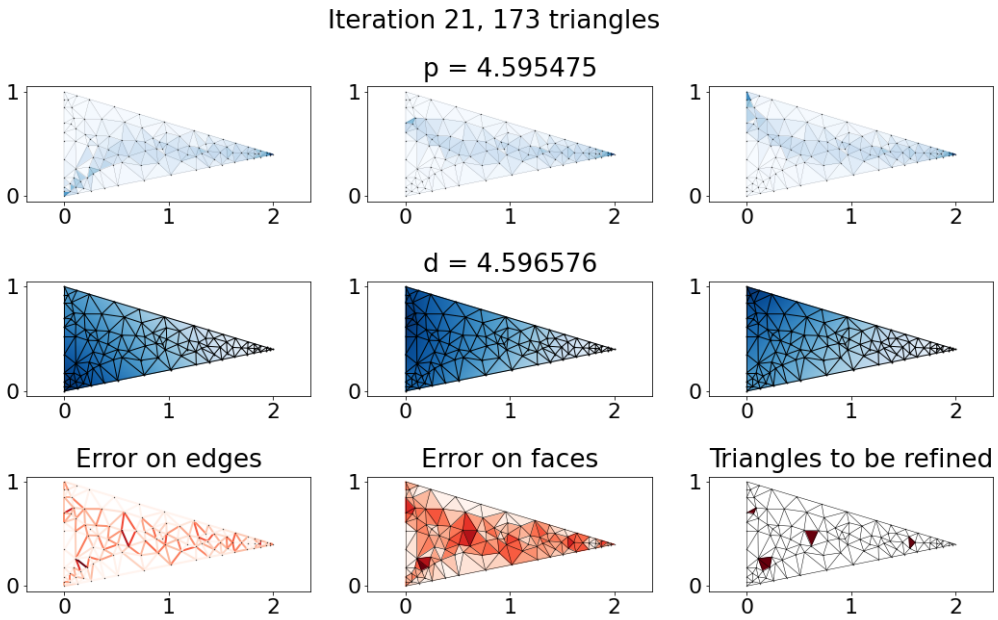

Figure 4.24: Numerical approximation of the non-conformal primal and conformal dual with divergence residual.

We can see that the primal flow follows the path we discovered in Example 4.4.2, however it is too diffused to reliably extract the topology of the true optimal graph. We can also see that the dual solution seems to be affine and remains stable (not seemingly dependent on the discretization) after further iterations.

Having performed numerical experiments with the two described approaches, we have confirmed that solving multimaterial problem in cases when a single topology is not prescribed can be challenging. We have also noted that the dual solution seemingly requires a much sparser discretization.

We therefore proceed to study in detail simple prototypical problems with more than one admissible topology to evaluate if more suitable discretization and mesh adaptation rules can be inferred from them.

4.5 Problem on 4 vertices

4.5.1 Motivation and overview

In this section we study the case where two networks with different topologies have the same transport cost. If both are also optimal, any dual solution must act as certificate for both (i.e. satisfy the corresponding primal-dual optimality conditions), and therefore the gradients of the dual potential must locally adapt to the direction of the fluxes of the two networks and thus cannot be globally linear. Already in the case where one network is optimal and the other one slightly sub-optimal, optimal dual potentials might not be globally linear, since they must still act as approximate certificates for the slightly sub-optimal network.

When both networks are indeed optimal, the additional primal-dual optimality conditions provide further information (or constraints) on the dual solution which might be helpful for their explicit construction. In future work one could then investigate whether there exist piecewise affine dual solutions on a simple mesh.

Throughout this section we will study the following example problem. Let there be $m = 3$ types of materials with sources and sinks as in the single topology case, (4.3.1), i.e.

$$\mu_+ = \sum_{i=1}^{m} e_i \, \delta_{x_i}, \qquad \mu_- = 1_m \delta_y \tag{4.5.1}$$

for $x_1, x_2, x_3, y \in \Omega$. It will be convenient to introduce the notation

$$A = x_1, \qquad B = x_2, \qquad C = x_3, \qquad D = y. \tag{4.5.2}$$

Assumption 4.5.1. Within this section, we consider $A, B, C, D \in \mathbb{R}^2$.

We will consider two potential network topologies. In in the first, flows from A and B merge first at a point S_1, and then with the flow from C at S_3, from which the combined materials flow to D. In the second, flows from A and C merge first at a point S_4, and then with the flow from B at S_2, and from there jointly flow to D (by Lemma 4.3.5, $S_1, S_2, S_3, S_4 \in \mathbb{R}^2$). The rationale behind the indexing of the merging points will become clear in Section 4.5.3. We will label the corresponding material vectors as

$$\Theta := \{\theta_a := e_1,\ \theta_b := e_2,\ \theta_c := e_3,\ \theta_d := 1_3,\ \theta_e := e_1 + e_2,\ \theta_f := e_1 + e_3\} \tag{4.5.3}$$

and denote by a, b, c, d, e, f the corresponding strictly positive cost coefficients, that together with Θ induce the cost function h, as outlined in Section 4.2. Networks for both topologies with corresponding material vectors are sketched in Figure 4.25.

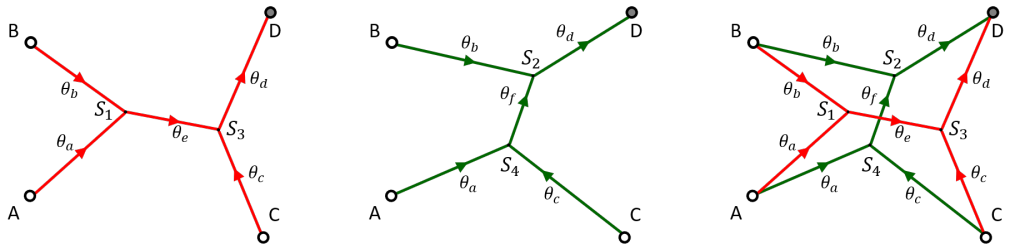

Figure 4.25: Two potential network topologies for 3 sources, 1 sink, with notation for merging points and the material vectors.

By removing θ_e or θ_f (or by setting the corresponding coefficients e or f to $+\infty$), the problem is reduced to the single topology setting of Section 4.3. In each case, there are two free vertices (S_1 and S_3 if $f = \infty$, S_2 and S_4 if $e = \infty$) and the optimal flow is given by a graph with straight edges between the appropriate fixed and free vertices.

We are interested in the case where the angles between the edges at the free vertices S_1 to S_4 are given by the corresponding cost triangles, as in (4.2.20), as this imposes the strictest constraints on the dual variables. All four relevant triangles and nomenclature for the involved angles are shown in Figure 4.26. Where these angles appear in the two networks is shown in Figure 4.27. We will be referring to these two networks as the e- and the f-graph. We will explicitly allow cases where some of the edges have zero length, as long as it is possible to consistently assign orientations to them that still match the cost triangle angles.

Assumption 4.5.2. We assume that $a, b, c, d, e, f > 0$ and that all four cost triangles exist and are not degenerate (i.e., each triangle has a strictly positive area). In particular, we have $\psi_X^i \in (0, \pi)$ for all appropriate $X \in \{A, B, C, D\}$, $i \in 1, 2, 3, 4$.

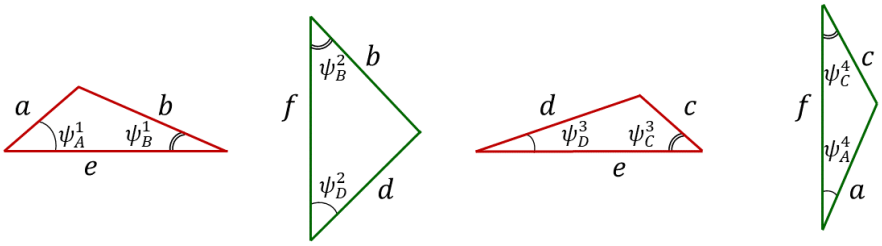

Figure 4.26: Cost triangles for the four free vertices $S_1, \ldots, S_4$. For each angle ψ_X^i the sup-script and sub-script indicate the associated free and fixed vertex, see also Figure 4.27.

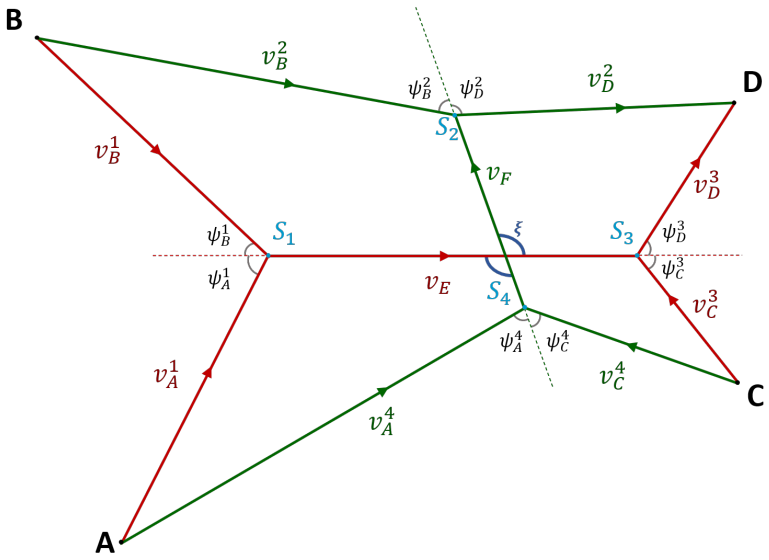

Figure 4.27: Free and fixed vertices and cost triangle vertices in both networks. The edges of the graphs follow the directions of the respective gradients of dual potentials to allow for primal-dual optimality conditions.

In Section 4.5.2 we examine whether both networks exist, with merging angles according to the cost triangles, and with the same transport cost. In Section 4.5.3 we show that imposing that both networks have the same transport cost is equivalent to a geometric condition that certain bisector lines intersect in a single point. These lines then provide a natural candidate mesh for defining piecewise affine dual solutions. In Section 4.5.4 we then derive necessary conditions for the cost coefficients (a, b, c, d, e, f) such that two equal cost networks with cost triangle angles exist.

4.5.2 Simultaneous existence of two networks with equal cost

We now consider the question for which choices of cost coefficients (a, b, c, d, e, f) there are positions A, B, C, D such that in both networks (the e- and f-graph) the edges at the free vertices meet with the angles prescribed by the cost triangles. The cost triangle angles fix the relative angles between any edges within the two graphs separately. By fixing an additional angle, for instance between the e and f segment, which we denote by $\xi \in (-\pi, \pi]$, see Figure 4.27, the angles between any two edges in both graphs are fixed. By choosing a suitable global coordinate system we can arrange that the e-segment (between vertices S_1 and S_3) is horizontal. The segments then have the following orientations:

$$\begin{aligned}
&\angle(S_1 - A) = \psi_A^1, && \angle(S_1 - B) = -\psi_B^1,\\
&\angle(S_2 - B) = \pi + \xi + \psi_B^2, && \angle(D - S_2) = \xi - \psi_D^2,\\
&\angle(D - S_3) = \psi_D^3, && \angle(S_3 - C) = \pi - \psi_C^3,\\
&\angle(S_4 - C) = \xi + \psi_C^4, && \angle(S_4 - A) = \xi - \psi_A^4,\\
&\angle(S_3 - S_1) = 0, && \angle(S_4 - S_2) = \xi.
\end{aligned} \tag{4.5.4}$$

We will also occasionally use the unit vectors v_X^i, v_E and v_F aligned with the respective edges, as shown in Figure 4.27, to denote the orientations.

We now show that for given ξ the question of existence of two networks can then be formulated as existence of non-negative solutions to a linear system of equations. To this end, denote by

$$L = (l_A^1, l_B^1, l_B^2, l_D^2, l_D^3, l_C^3, l_C^4, l_A^4, l_E, l_F) \in \mathbb{R}^{10} \tag{4.5.5}$$

the lengths of the corresponding segments in the two networks. We need to find non-negative values for these lengths such that the two networks both connect the same fixed vertices (A, B, C, D). This requires, for instance, that if we move along the top cycle (B, S_1, S_3, D, S_2, B) we must indeed return to the initial point B. Given the edge orientations (4.5.4) and edge lengths (4.5.5), this yields the condition

$$l_E \begin{bmatrix}1\\0\end{bmatrix} + l_D^3 \begin{bmatrix}\cos(\psi_D^3)\\ \sin(\psi_D^3)\end{bmatrix} - l_D^2 \begin{bmatrix}\cos(\xi - \psi_D^2)\\ \sin(\xi - \psi_D^2)\end{bmatrix} + l_B^2 \begin{bmatrix}\cos(\xi + \psi_B^2)\\ \sin(\xi + \psi_B^2)\end{bmatrix} + l_B^1 \begin{bmatrix}\cos(-\psi_B^1)\\ \sin(-\psi_B^1)\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}. \tag{4.5.6a}$$

Likewise, consistency along the cycles (A, S_1, S_3, C, S_4, A) (bottom) and (C, S_4, S_2, D, S_3, C) (right) leads to the equations

$$l_E \begin{bmatrix}1\\0\end{bmatrix} + l_A^1 \begin{bmatrix}\cos(\psi_A^1)\\ \sin(\psi_A^1)\end{bmatrix} - l_A^4 \begin{bmatrix}\cos(\xi - \psi_A^4)\\ \sin(\xi - \psi_A^4)\end{bmatrix} + l_C^4 \begin{bmatrix}\cos(\xi + \psi_C^4)\\ \sin(\xi + \psi_C^4)\end{bmatrix} + l_C^3 \begin{bmatrix}\cos(-\psi_C^3)\\ \sin(-\psi_C^3)\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}, \tag{4.5.6b}$$

and

$$l_F \begin{bmatrix}1\\0\end{bmatrix} + l_D^2 \begin{bmatrix}\cos(\psi_D^2)\\ \sin(\psi_D^2)\end{bmatrix} - l_D^3 \begin{bmatrix}\cos(\xi - \psi_D^3)\\ \sin(\xi - \psi_D^3)\end{bmatrix} + l_C^3 \begin{bmatrix}\cos(\xi + \psi_C^3)\\ \sin(\xi + \psi_C^3)\end{bmatrix} + l_C^4 \begin{bmatrix}\cos(-\psi_C^4)\\ \sin(-\psi_C^4)\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}. \tag{4.5.6c}$$

Adding equations for any other cycle in the two graphs will make the system linearly dependent.

For a given $\xi \in (-\pi, \pi]$ and a corresponding non-negative solution L to (4.5.6) one can then infer potential positions for all fixed and free vertices. Of course, the edge lengths are invariant under translation or rotation of all vertex positions. And for any solution L, any positive re-scaling $\lambda \cdot L$, $\lambda > 0$, will also induce a solution.

The two networks corresponding to (non-negative) solutions of (4.5.6) may still have the same cost. The following equation enforces that both networks have the same cost:

$$a\,l_A^1 + b\,l_B^1 + c\,l_C^3 + d\,l_D^3 + e\,l_E = a\,l_A^4 + b\,l_B^2 + c\,l_C^4 + d\,l_D^2 + f\,l_F. \tag{4.5.7}$$

The left and right side of this equation state the transport cost of the two respective networks (see Figure 4.27) and thus any $L \geq 0$ that satisfies (4.5.6) and (4.5.7) corresponds to two networks between the same fixed vertices with equal cost.

For the trivial solution $L = 0$ all vertices lie in the same point, and it is thus of little interest. The following definition clarifies the relevant set of solutions.

Definition 4.5.3 (Non-degenerate solutions and graphs)**.** We call a solution $L \in \mathbb{R}^{10}$ of (4.5.6) *non-degenerate* if $L \geq 0$ and if all of the implied fixed vertex positions are distinct from each other. This means that the lengths corresponding to each cycle in the two networks must contain at least one non-zero entry. If such a solution exists, we also say that two *non-degenerate* graphs exist. If the solution in addition satisfies the equal-cost condition (4.5.7), we say it is a non-degenerate same cost solution or there are two non-degenerate same cost graphs.

The following Lemma gives necessary conditions for ξ, such that non-degenerate solutions can exist.

Proposition 4.5.4 (Necessary conditions for angle ξ)**.** *For fixed ξ, for equations (4.5.6) to have non-degenerate solutions $L \in \mathbb{R}^{10}$ in the sense of Definition 4.5.3 it is necessary (but not sufficient) that either*

$$\xi \in [0,\pi] \cap [\pi - \max\{\psi_B^1, \psi_C^3\} - \max\{\psi_B^2, \psi_C^4\}, \max\{\psi_A^1, \psi_D^3\} + \max\{\psi_A^4, \psi_D^2\}] \tag{4.5.8}$$

or

$$\xi \in (-\pi,0) \cap [\pi - \max\{\psi_B^1, \psi_C^3\} - \max\{\psi_B^2, \psi_C^4\}, \max\{\psi_A^1, \psi_D^3\} + \max\{\psi_A^4, \psi_D^2\} - 2\pi]. \tag{4.5.9}$$

Proof. For (4.5.6) to have a non-degenerate solution, the equations corresponding to the diagonal cycles $(A, S_1, S_3, D, S_2, S_4, A)$ and $(B, S_1, S_3, C, S_4, S_2, B)$ must have non-trivial solutions. Since the relative orientation of the edges in these cycles will depend on ξ, in the rest of the proof we will consider the cases $\xi \in [0,\pi]$ and $\xi \in (-\pi, 0)$ separately.

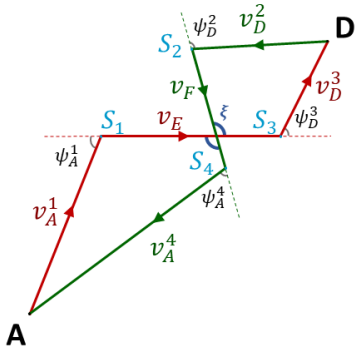

Figure 4.28: Loop $(A, S_1, S_3, D, S_2, S_4, A)$.

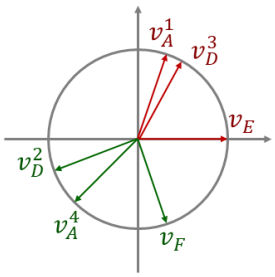

Figure 4.29: Edge orientations in Figure 4.28.

Case 1: $\xi \in [0,\pi]$**.** The cycle $(A, S_1, S_3, D, S_2, S_4, A)$ is shown in Figure 4.28 with notation for unit vectors v_X^i aligned with each of the segments. The orientation of these unit vectors is given by

$$\begin{aligned} &\angle v_E = 0, && \angle v_F = \xi - \pi \in [-\pi, 0], \\ &\angle v_A^1 = \psi_A^1 \in (0,\pi), && \angle v_A^4 = \xi - \pi - \psi_A^4 \in (\xi - 2\pi, \xi - \pi), \\ &\angle v_D^3 = \psi_D^3 \in (0,\pi), && \angle v_D^2 = \xi - \pi - \psi_D^2 \in (\xi - 2\pi, \xi - \pi). \end{aligned} \tag{4.5.10}$$

The orientations are illustrated in Figure 4.29 and their relative ordering is as shown in the figure, up to swapping v_A^1, v_D^3 and/or v_A^4, v_D^2.

For the equation

$$l_A^1 \cdot v_A^1 + l_E \cdot v_E + l_D^3 \cdot v_D^3 + l_D^2 \cdot v_D^2 + l_F \cdot v_F + l_A^4 \cdot v_A^4 = 0$$

to have a non-trivial non-negative solution $(l_A^1, l_E, l_D^3, l_D^2, l_F, l_A^4)$ the cone spanned by the corresponding vectors $(v_A^1, v_E, v_D^3, v_D^2, v_F, v_A^4)$ (i.e. the set of points spanned by these vectors with non-negative coefficients) must contain at least a half-space. This could be shown, for instance, by an application of Farkas' lemma (see Lemma 4.5.20 and Corollary 4.5.21).

By (4.5.10) the only angle between two adjacent vectors that can be larger than π is the one between $v_{A/D}^{4/2}$ and $v_{A/D}^{1/3}$ (depending on the relative orientation of v_A^4, v_D^2 and v_A^1, v_D^3). So for a non-trivial solution to exist we must impose that

$$\max\{\angle v_A^1, \angle v_D^3\} - \min\{\angle v_A^4, \angle v_D^2\} \geq \pi,$$

which is equivalent to

$$\xi \leq \max\{\psi_A^1, \psi_D^3\} + \max\{\psi_A^4, \psi_D^2\}.$$

Now apply the same argument to the cycle $(C, S_3, S_1, B, S_2, S_4, C)$. By symmetry this corresponds to substituting

$$\psi_A^1 \to \psi_C^3, \qquad \psi_A^4 \to \psi_C^4, \qquad \psi_D^2 \to \psi_B^2, \qquad \psi_D^3 \to \psi_B^1, \qquad \xi \to \pi - \xi$$

and thus yields the bound

$$\pi - \xi \leq \max\{\psi_C^3, \psi_B^1\} + \max\{\psi_C^4, \psi_B^2\}.$$

Together, the two conditions imply that in the case $\xi \in [0,\pi]$ one has

$$\xi \in [0,\pi] \cap \left[\pi - \max\{\psi_B^1, \psi_C^3\} - \max\{\psi_B^2, \psi_C^4\}, \max\{\psi_A^1, \psi_D^3\} + \max\{\psi_A^4, \psi_D^2\}\right]. \tag{4.5.11}$$

Case 2: $\xi \in (-\pi, 0)$**.** In this case the formulas (4.5.10) for the edge orientations remain valid, but the bounds and their relative orientation, as shown in Figure 4.29, may change. The updated bounds can be expressed as

$$\begin{aligned} &\angle v_E = 0, && \angle v_F = \pi + \xi \in (0,\pi), \\ &\angle v_A^1 = \psi_A^1 \in (0,\pi), && \angle v_A^4 = \pi + \xi - \psi_A^4 \in (\xi, \pi + \xi), \\ &\angle v_D^3 = \psi_D^3 \in (0,\pi), && \angle v_D^2 = \pi + \xi - \psi_D^2 \in (\xi, \pi + \xi). \end{aligned}$$

The largest angle (in $(-\pi,\pi]$) will be either that of v_F, v_A^1 or v_D^3. The smallest angle will be one of v_E, v_A^4 or v_D^2. From largest to smallest, the difference between any two adjacent angles will always be strictly less than π. Thus, the only condition to impose for existence of a solution is

$$\max\{\angle v_F, \angle v_A^1, \angle v_D^3\} - \min\{\angle v_E, \angle v_A^4, \angle v_D^2\} \geq \pi,$$

which becomes

$$\max\{\pi-\xi,\psi_A^1,\psi_D^3\}-\min\{0,\pi+\xi-\psi_A^4,\pi+\xi-\psi_D^2\}\geq\pi.$$

One can see that if $\pi-\xi$ is maximal in the first term, or 0 is minimal in the second, then the condition will be false. So we can simplify this to

$$\max\{\psi_A^1,\psi_D^3\}-\min\{\pi+\xi-\psi_A^4,\pi+\xi-\psi_D^2\}\geq\pi,$$

which is finally equivalent to

$$\xi\leq\max\{\psi_A^1,\psi_D^3\}+\max\{\psi_A^4,\psi_D^2\}-2\pi.$$

Applying the same variable substitutions as in the first case (but this time we have to substitute $\xi\to-\xi-\pi$ to remain in the same interval) for the other diagonal cycle, we obtain the bound

$$-\pi-\xi\leq\max\{\psi_B^1,\psi_C^3\}+\max\{\psi_B^2,\psi_C^4\}-2\pi.$$

Together, we obtain

$$\xi\in(-\pi,0)\cap\left[\pi-\max\{\psi_B^1,\psi_C^3\}-\max\{\psi_B^2,\psi_C^4\},\max\{\psi_A^1,\psi_D^3\}+\max\{\psi_A^4,\psi_D^2\}-2\pi\right]. \quad \square$$

We conjecture that in fact no solutions can exist for $\xi\in(-\pi,0)$, however this would require a more careful analysis of combinations of values for the cost triangle angles ψ_X^i.

4.5.3 Equal cost and bisectors

For fixed ξ and cost coefficients (a,b,c,d,e,f), (4.5.6) and (4.5.7) are a linear system of 7 equations for 10 length variables. Unfortunately, the non-linear and coupled dependency between the angles ψ_X^i and the cost coefficients (a,b,c,d,e,f), as well as the appearance of the parameter ξ make the analysis of existence of (non-degenerate) solutions to this system quite difficult.

In this section we provide an equivalent condition for the last equation (4.5.7), with a geometric interpretation, which will then later allow to derive necessary conditions on the coefficients (a,b,c,d,e,f) for the simultaneous existence of two equal-cost solutions, as well as a candidate mesh for simple piecewise affine dual solutions.

Consider again the challenge of constructing a piecewise affine dual optimal potential. Assume that both networks were known, had equal cost and were optimal, arranged in a way similar to Figure 4.27. Focus now on the dual potential ϕ_1 for material type 1, which has its source $\mu_+^1=\delta_{x_1}$ at point $A=x_1$. Dual feasibility dictates that ϕ_1 must be Lipschitz, with Lipschitz constant a. The primal-dual optimality conditions imply ϕ_1 must decrease with maximal slope a from A towards S_1 and S_4. If ϕ_1 is to be piecewise affine on a simple mesh (i.e. with few triangles), then there must be at least two triangles touching the point A and, for symmetry reasons, one may assume the boundary between the two triangles to run along the *bisector* between the segments $[A,S_1]$ and $[A,S_4]$. The same consideration applies to the vertices B and C for potentials ϕ_2 and ϕ_3, and to vertex D for the sum of potentials $\sum_{i=1}^3\phi_i$. These four bisector lines therefore seem to be natural candidates for the basis of the sought-after simple mesh and the question arises how the central region, where the bisector lines meet, can be triangulated. Figure 4.30 illustrates the four bisector lines and introduces some related notation to be used in the following. The main result of this section is that (under minor technical assumptions) for solutions $L\in\mathbb{R}^{10}$ of (4.5.6) the four bisectors will intersect in a single point if and only if L also solves (4.5.7), i.e. when both networks have the same transport cost. This implies that the four lines do indeed offer a simple and very tempting candidate mesh for dual solutions. Studying the feasibility and optimality of solutions on such meshes could offer an interesting direction for the future work.

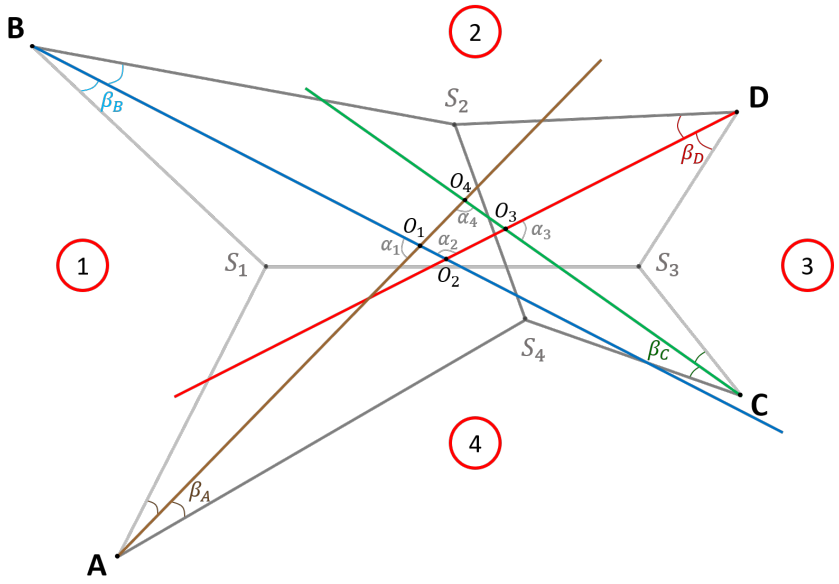

Figure 4.30: Two prototypical flow networks, the bisectors between the two edges connected to each of the four vertices A, B, C, D and with some relevant intersections marked by O_1 to O_4. In addition, angles between the bisectors are labeled α_1 to α_4 and the angles between bisectors and the flow edges are labeled β_A to β_D.

The following angles, also shown in Figure 4.30, will be instrumental for the rest of this section. Their values can be determined by elementary geometric considerations.

Definition 4.5.5 (Vertex and splitting angles)**.** We call

$$\begin{aligned} \beta_A &= \frac{-\xi + \psi_A^1 + \psi_A^4}{2}, & \beta_B &= \frac{\xi + \psi_B^1 + \psi_B^2 - \pi}{2}, \\ \beta_C &= \frac{\xi + \psi_C^3 + \psi_C^4 - \pi}{2}, & \beta_D &= \frac{-\xi + \psi_D^2 + \psi_D^3}{2} \end{aligned} \tag{4.5.12}$$

the *vertex angles* between the flow segments and the bisectors at each vertex, and

$$\begin{aligned} \alpha_1 &= \frac{\psi_A^1 + \psi_B^1 - \psi_A^4 - \psi_B^2 + \pi}{2}, & \alpha_2 &= \frac{\psi_B^2 + \psi_D^2 - \psi_B^1 - \psi_D^3 + \pi}{2}, \\ \alpha_3 &= \frac{\psi_C^3 + \psi_D^3 - \psi_C^4 - \psi_D^2 + \pi}{2}, & \alpha_4 &= \frac{\psi_A^4 + \psi_C^4 - \psi_A^1 - \psi_C^3 + \pi}{2} \end{aligned} \tag{4.5.13}$$

the *splitting angles* (since they split the whole domain into four sectors if all bisectors meet in one point). Note that the splitting angles no not depend on ξ but are solely determined by the cost triangle angles (and hence, the cost coefficients).

Remark 4.5.6 (Sign convention for vertex and splitting angles)**.** Some care has to be taken with the sign convention of these angles. We will show below that $\beta_X \in (-\pi/2, \pi)$ (Proposition 4.5.8). In an arrangement as shown in Figure 4.30 the convention is that all vertex and splitting angles are positive. If, for instance, β_A were negative (and therefore in $(-\pi/2, 0)$) then the free vertices S_1 and S_4 would lie on the opposite side of the bisector line $(A - O_1)$, respectively. If

$\alpha_1 > 0$, then the orientation of the B-bisector $(B - O_1)$ is obtained from the orientation of the A-bisector $(A - O_1)$ by a clockwise rotation by $|\alpha_1| = \alpha_1$.

Formulas (4.5.12) and (4.5.13) also allow for negative values for these angles, which may appear counter intuitive and it is not obvious for all configurations what the corresponding transport networks would look like. In the following, we establish some non-existence results, that clarify that "too negative angle configurations" do not exist (for any admissible choice of cost coefficients) or that no transport networks with equal cost do exist in these cases. See for instance Propositions 4.5.8, 4.5.9 and several of the statements of Section 4.5.4.

Remark 4.5.7 (Sectors)**.** If all $\alpha_i > 0$ for $i = 1, \dots, 4$ and all four bisectors meet in a single point, then they divide the space $\mathbb{R}^2$ into four wedges which we will refer to as sectors, enumerated by 1 to 4. If all $\beta_X \in [0, \pi]$, then the free vertices S_1 to S_4 will lie in the sectors with the same index, and the same index is accordingly also used for the ψ-angles and segments of the transport networks associated with these free vertices.

Finally, when bisectors do not meet in a single point, we label some of their pairwise intersections according to the sector that they correspond to.

Proposition 4.5.8. *Let $\xi \in (-\pi, \pi]$ and consider the vertex angles β_A, β_B, β_C and β_D defined in (4.5.12). Assume there exist two non-degenerate graphs. Then*

$$\beta_A, \beta_B, \beta_C, \beta_D \in \left(-\frac{\pi}{2}, \pi\right).$$

Proof. From the definition of the vertex angles (4.5.12), using that angles ψ_X^i are defined via cost triangles, we immediately observe the following boundaries:

$$\beta_A, \beta_D \in \left(-\frac{\pi}{2}, \frac{3\pi}{2}\right), \qquad \beta_B, \beta_C \in (-\pi, \pi).$$

Observe now that, again via cost triangles, $\beta_A + \beta_B = (\psi_A^1 + \psi_B^1 + \psi_A^4 + \psi_B^2 - \pi)/2 < \pi$ and $\beta_B + \beta_D = (\psi_B^2 + \psi_D^2 + \psi_B^1 + \psi_D^3 - \pi)/2 < \pi$. In particular, if $\beta_A \geq \pi$, then $\beta_B < 0$. Assume now that $\beta_A \in [\pi, 3\pi/2)$. Then, we have

$$\xi \in (\psi_A^1 + \psi_A^4 - 3\pi, \psi_A^1 + \psi_A^4 - 2\pi],$$

and in particular $\xi \leq \psi_A - 2\pi < 0$. So in order for two non-degenerate graphs to exist, using Proposition 4.5.4, we must have in addition

$$\xi \in (-\pi, 0) \cap [\pi - \max\{\psi_B^1, \psi_C^3\} - \max\{\psi_B^2, \psi_C^4\}, \max\{\psi_A^1, \psi_D^3\} + \max\{\psi_A^4, \psi_D^2\} - 2\pi].$$

Therefore, we can use $\psi_A^1 + \psi_A^4 - 2\pi$ for the upper bound and for the range to be non-empty it will be necessary that

$$\psi_A^1 + \psi_A^4 + \max\{\psi_B^1, \psi_C^3\} + \max\{\psi_B^2, \psi_C^4\} \geq 3\pi.$$

As the sum of any two angles from the same cost triangle $\psi_X^i + \psi_Y^i < \pi$, to have a possibility of satisfying this inequality we must have $\psi_C^3 > \psi_B^1$ and $\psi_B^2 > \psi_C^4$, so that it is necessary that

$$\psi_A^1 + \psi_A^4 + \psi_C^3 + \psi_B^2 \geq 3\pi. \tag{4.5.14}$$

We then use a technique similar to the one presented in the proof of Proposition 4.5.4 and consider construction of the right loop (D, S_2, S_4, C, S_3, D) (see Figure 4.31). The angles for the

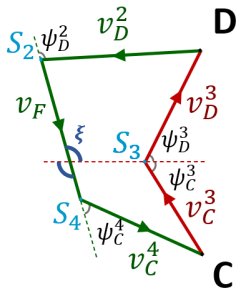

Figure 4.31: Loop (D, S_2, S_4, C, S_3, D).

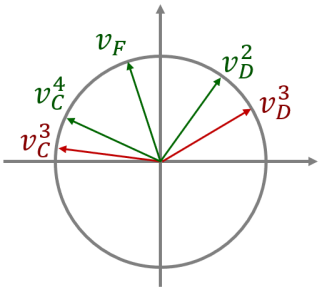

Figure 4.32: Edge orientations in Figure 4.31.

unit vectors indicating the segment orientations are

$$\begin{aligned}
&\angle v_F = \xi + \pi \in (0, \pi) \\
&\angle v_D^2 = \xi + \pi - \psi_D^2 \in (\xi, \xi + \pi) && \angle v_D^3 = \psi_D^3 \in (0, \pi) \\
&\angle v_C^4 = \xi + \pi + \psi_C^4 \in (\xi + \pi, \xi + 2\pi) && \angle v_C^3 = \pi - \psi_C^3 \in (0, \pi).
\end{aligned}$$

Note that $\angle v_C^3 > \angle v_D^3$ (from cost triangles) and $\angle v_C^4 > \angle v_F > \angle v_D^2$, that is, the order of the green vectors and the order of the red vectors separately is as shown in Figure 4.32. We also can see from the cost triangles that $\angle v_C^3 - \angle v_D^3 < \pi$ and using $\psi_B^2 > \psi_C^4$ (see above) and the cost triangles we find $\angle v_C^4 - \angle v_D^2 < \pi$. Therefore, in order for these 5 vectors to be able to form a non-degenerate a closed loop with positive coefficients, we need that

$$\angle v_C^3 - \angle v_D^2 \geq \pi \qquad \text{or} \qquad \angle v_C^4 - \angle v_D^3 \geq \pi. \tag{4.5.15}$$

Consider the first option of (4.5.15). From $\xi \geq \psi_A - 3\pi$ and (4.5.14) we then estimate the left side of the expression above as

$$-\psi_C^3 - \xi + \psi_D^2 \leq -\psi_C^3 - \psi_A + 3\pi + \psi_D^2 \leq \psi_B^2 + \psi_D^2 < \pi,$$

with the last inequality given by the cost triangle. I.e. this condition is not satisfied.

Now consider the second option of (4.5.15). Using $\xi \leq \psi_A - 2\pi$, we estimate the left side as

$$\xi + \pi + \psi_C^4 - \psi_D^3 \leq \psi_A^4 + \psi_C^4 + \psi_A^1 - \psi_D^3 - \pi < \pi,$$

as the sum of the first two terms $\psi_A^4 + \psi_C^4 < \pi$ and the next term $\psi_A^4 < \pi$ by the cost triangles.

So neither of the two conditions is satisfied and therefore the loop cannot be constructed. Hence, we cannot have $\beta_A \in (\pi, 3\pi/2]$. With analogous arguments it can be shown that $\beta_D \notin (\pi, 3\pi/2]$ as well as $\beta_B, \beta_C \notin (-\pi, -\pi/2]$. □

Proposition 4.5.9. *The splitting angles $\alpha_1, \ldots, \alpha_4$, (4.5.13), lie in $(-\pi/2, \pi)$. At least three splitting angles are strictly positive.*

Proof. By Assumption 4.5.2 the cost triangle angles ψ_X^i all lie in $(0, \pi)$. The sum of two cost triangle angles that appear in the same cost triangle must be less than π. This yields the bound $\alpha_i \in (-\pi/2, \pi)$.

The sum of all the splitting angles is 2π. Let $P \subset \{1, 2, 3, 4\}$ be the indices of strictly positive angles. One has

$$\sum_{i \in P} \alpha_i \geq \sum_{i=1}^{4} \alpha_i = 2\pi.$$

Since each $\alpha_i < \pi$, the inequality above can be satisfied only if P contains at least three indices. □

Based on the definition of the vertex angles we can assign orientations to the bisector segments as follows:

$$\begin{aligned} &\angle(A - O_1) = \angle(A - O_4) = \psi_A^1 - \beta_A, && \angle(B - O_1) = \angle(B - O_2) = -\psi_B^1 + \beta_B, \\ &\angle(C - O_3) = \angle(C - O_4) = \pi - \psi_C^3 + \beta_C, && \angle(D - O_3) = \angle(D - O_2) = \pi + \psi_D^3 - \beta_D. \end{aligned} \tag{4.5.16}$$

Intuitively, these point "outwards" from the intersection toward the vertices. Using these orientations we can then assign lengths to the bisector segments, from the vertices towards the intersection points O_i, as follows.

Definition 4.5.10 (Bisector lengths)**.** Assume $\alpha_1 \neq 0$. For given $l_A^1, l_B^1 \in \mathbb{R}^2$, as introduced in (4.5.5), we define the (potentially negative) *bisector lengths* (b_A^1, b_B^1) as the unique coefficients satisfying

$$b_B^1 \begin{bmatrix} \cos(\beta_B - \psi_B^1) \\ \sin(\beta_B - \psi_B^1) \end{bmatrix} - l_B^1 \begin{bmatrix} \cos(-\psi_B^1) \\ \sin(-\psi_B^1) \end{bmatrix} + l_A^1 \begin{bmatrix} \cos(\psi_A^1) \\ \sin(\psi_A^1) \end{bmatrix} - b_A^1 \begin{bmatrix} \cos(\psi_A^1 - \beta_A) \\ \sin(\psi_A^1 - \beta_A) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which corresponds to closing the oriented loop (O_1, B, S_1, A, O_1) in Figure 4.30. The solution is given explicitly by

$$\begin{bmatrix} b_A^1 \\ b_B^1 \end{bmatrix} = \frac{1}{\sin(\alpha_1)} \begin{bmatrix} \sin(\beta_A + \alpha_1) & \sin(\beta_B) \\ \sin(\beta_A) & \sin(\beta_B + \alpha_1) \end{bmatrix} \begin{bmatrix} l_A^1 \\ l_B^1 \end{bmatrix}. \tag{4.5.17}$$

In the same way we define the bisector lengths (b_B^2, b_D^2), (b_D^3, b_C^3) and (b_C^4, b_A^4) whenever the corresponding splitting angles are nonzero.

These bisector lengths can be used to analyze whether all four bisector meeting points O_1 to O_4 coincide, and they can also be used to parametrize solutions to the loop equations (4.5.6).

Lemma 4.5.11. *Assume $\alpha_1, \alpha_2 \neq 0$ and $\beta_C \notin \{0, \pi/2\}$. Then there is a linear bijection between solutions $L \in \mathbb{R}^{10}$ of (4.5.6) and the four bisector lengths $(b_A^1, b_B^1, b_B^2, b_D^2)$ as given by Definition 4.5.10.*

This means that in this case the system of equations (4.5.6) has rank 6 and its four-dimensional space of solutions can be parametrized by the named bisector lengths. Of course, analogous results hold for any other pairs of bisector lengths from adjacent sectors.

Proof. Since $\alpha_1 \neq 0$, the two bisector lengths (b_A^1, b_B^1) are in linear one-to-one correspondence with the two lengths (l_A^1, l_B^1) by Definition 4.5.10, and likewise for (b_B^2, b_D^2) and (l_B^2, l_D^2) since $\alpha_2 \neq 0$.

Now we show that if $\beta_C \neq 0$ and $(l_A^1, l_B^1, l_B^2, l_D^2)$ are given, the other lengths of the networks (i.e. entries of L) are uniquely determined by linear equations. Using the orientations of (4.5.4) we find that

$$S_2 - A = l_A^1 \cdot v_A^1 - l_B^1 \cdot v_B^1 + l_B^2 \cdot v_B^2 = l_A^4 \cdot v_A^4 + l_F \cdot v_F.$$

By Assumption 4.5.2 the vectors v_A^4 and v_F are not (anti-)parallel. Hence l_A^4 and l_F are uniquely determined through the above equation. The same applies to the two lengths (l_D^3, l_E). At this point the vector $S_3 - S_4$ is known. Since by assumption $\beta_C \notin \{0, \pi/2\}$ and with the bound of Proposition 4.5.8, the vectors v_C^3 and v_C^4 are linearly independent and $S_3 - S_4$ uniquely determines the two lengths (l_C^3, l_C^4). □

Lemma 4.5.12. *Assume $\alpha_1, \alpha_3, \alpha_4 \neq 0$ and $\beta_B, \beta_D \notin \{0, \pi/2\}$. Let $L \in \mathbb{R}^{10}$ be a solution of (4.5.6). Then $[O_1 = O_4] \Leftrightarrow [O_3 = O_4]$.*

Proof. By Lemma 4.5.11 there is a linear one-to-one correspondence between the bisector lengths $(b_C^4, b_A^4, b_A^1, b_B^1)$, solutions $L \in \mathbb{R}^{10}$ to (4.5.6), and $(b_D^3, b_C^3, b_C^4, b_A^4)$. So in particular there is a linear bijection between the two 4-tuples of bisector lengths. This means that there are coefficients $(u_1, u_2, u_3, u_4) \in \mathbb{R}$ such that

$$b_A^1 - b_A^4 = u_1 \cdot b_D^3 + u_2 \cdot b_C^3 + u_3 \cdot b_C^4 + u_4 \cdot b_A^4. \tag{4.5.18}$$

These coefficients can be determined by computing the lengths of intersecting line segments, as described in the proof of Lemma 4.5.11, leading to formulas similar to (4.5.17), and nested versions thereof. (We recommend a computer algebra tool to keep track of these coefficients.) For instance, directly after plugging in all subsequent identities one finds that

$$\begin{aligned} u_1 = \frac{1}{\sin(\psi_A^4+\psi_C^4)\sin(\alpha_1)} \Bigg(& \frac{\sin(\alpha_1+\beta_A)\sin(\beta_A)\sin(\psi_C^4+\xi)}{\sin(\psi_A^1)} + \frac{\sin(\psi_C^4)\sin(\psi_B^1+\psi_D^2)\sin(\beta_A)\sin(\beta_B)}{\sin(\psi_D^2)\sin(2\beta_B)} \\ & + \frac{\sin(\beta_A)\sin(\psi_A^1-\psi_C^4-\xi)\sin(\psi_B^2+\xi)\sin(\beta_B)}{\sin(\psi_A^1)\sin(2\beta_B)} - \frac{\sin(\alpha_1+\beta_A)\sin(\psi_A^4-\xi)\sin(\alpha_4+\beta_C)}{\sin(\psi_A^1)} \\ & + \frac{\sin(2\beta_A)\sin(\psi_B^2+\xi)\sin(\beta_B)\sin(\alpha_4+\beta_C)}{\sin(\psi_A^1)\sin(2\beta_B)} \Bigg). \end{aligned}$$

Careful manipulation with extensive use of trigonometric identities and the law of sines on the cost triangles, which implies for example that

$$\frac{\sin\psi_A^1}{b} = \frac{\sin\psi_B^1}{a} = \frac{\sin(\psi_A^1+\psi_A^1)}{e},$$

reveal that $u_1 = u_4 = 0$ and $u_2 = -u_3$. Therefore

$$b_A^1 - b_A^4 = u_2 \cdot (b_C^3 - b_C^4).$$

Since the linear map from $(b_D^3, b_C^3, b_C^4, b_A^4)$ to $(b_C^4, b_A^4, b_A^1, b_B^1)$ is full rank one must have $u_2 \neq 0$. Therefore $[b_A^1 - b_A^4 = 0] \Leftrightarrow [b_C^3 - b_C^4 = 0]$. These two conditions correspond to the condition $O_1 = O_4$ and $O_3 = O_4$ respectively, which proves the claim. □

The following analysis will at times be greatly simplified by the assumption that no two adjacent bisectors are (anti-)parallel and neither are two graph segments emerging from the same vertex. This is encoded by the following assumption. We consider this to be the "generic" case. The remaining special cases could again be studied in more detail if they become relevant.

Assumption 4.5.13. One has $\beta_X \notin \{0, \pi/2\}$ for $X \in \{A, B, C, D\}$ and $\alpha_i \neq 0$ for $i \in \{1, 2, 3, 4\}$.

Lemma 4.5.14. *Under Assumption 4.5.13, if any two of the bisector meeting points O_1 to O_4 coincide, then all four coincide.*

Proof. If two "opposite" meeting points coincide (i.e. $O_i = O_j$ for $(i, j) \in \{(1, 3), (2, 4)\}$), then all four bisectors clearly meet in a single point (cf. Figure 4.30). If two "adjacent" meeting points coincide (i.e. for $(i, j) \in \{(1, 2), (2, 3), (3, 4), (4, 1)\}$) then the result follows by repeated application of Lemma 4.5.12. □

Finally, we are able to state the main result of this section.

Theorem 4.5.15. *Under Assumption 4.5.13 and assuming that* $\psi_A^1 + \psi_B^1 + \psi_A^4 + \psi_B^2 \neq \pi$*, for any solution* $L \in \mathbb{R}^{10}$ *of* (4.5.6) *one has that* L *also solves* (4.5.7)*, the equal cost condition, if and only if all four bisectors corresponding to* L *meet in a single point.*

This holds in particular for non-zero, non-degenerate solutions (Definition 4.5.3) and therefore it implies that two transport networks have equal cost if and only if their implied bisectors meet in a single point. We will see in Section 4.5.4 that for $\psi_A^1 + \psi_B^1 + \psi_A^4 + \psi_B^2 < \pi$ no non-degenerate same cost solutions L can exist and thus the above assumption excludes the (potentially delicate) boundary case.

Proof. By Assumption 4.5.13 and Lemma 4.5.11 the lengths $(b_D^3, b_C^3, b_C^4, b_A^4)$ are in one-to-one linear correspondence with L. By Lemma 4.5.14, all bisectors meet if and only if any two of them meet. We can thus restrict ourselves, for example, to the points O_3 and O_4. The condition $O_3 = O_4$ is equivalent to $b_C^3 = b_C^4$. Parametrizing the solution L by $(b_D^3, b_C^3, b_C^4, b_A^4)$ and plugging this into (4.5.7) yields an equation of the form

$$0 = v_1 \cdot b_D^3 + v_2 \cdot b_C^3 + v_3 \cdot b_C^4 + v_4 \cdot b_A^4.$$

similar as (4.5.18). Again, by careful manipulation similar as in Lemma 4.5.12 one finds that $v_1 = v_4 = 0$ and

$$v_2 = -v_3 = \frac{2a \sin(\alpha_4)}{\cos(\beta_B)} \cdot \cos\left((\psi_A^1 + \psi_A^4 + \psi_B^1 + \psi_B^2)/2\right).$$

From the cost triangles it follows that $\psi_A^1 + \psi_B^1 + \psi_A^4 + \psi_B^2 \in (0, 3\pi)$ and thus by the assumptions made in this theorem the above expression is non-zero.

It therefore follows that L solves (4.5.7) if and only if $b_C^3 = b_C^4$, which completes the proof. □

4.5.4 Excluding graph configurations a priori

Theorem 4.5.15 gives a simple geometric interpretation for solutions L to (4.5.6), relating the equal cost condition to the bisectors. But throughout Section 4.5.3 the non-degeneracy of L, and potentially negative values for some splitting or vertex angles were not discussed. Figures 4.33 and 4.34 show hypothetical configurations with a negative vertex angle and a negative splitting angle, respectively. In this section we give some necessary conditions on the cost coefficients such that non-degenerate solutions with equal costs can exist. These then pose some limits on the negativity of these angles and thus rule out some particularly counter-intuitive configurations.

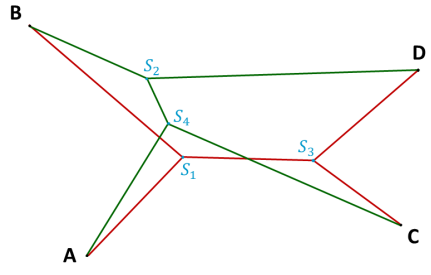

Figure 4.33: Example: $\beta_A < 0$.

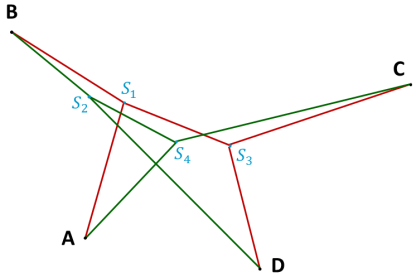

Figure 4.34: Example: $\alpha_3 < 0$.

In the following we will abbreviate

$$\psi_A = \psi_A^1 + \psi_A^4, \qquad \psi_B = \psi_B^1 + \psi_B^2, \qquad \psi_C = \psi_C^3 + \psi_C^4, \qquad \psi_D = \psi_D^2 + \psi_D^3.$$

Moreover, we will frequently use that by (4.5.12) for any two adjacent vertices $(X, Y) \in \{(A, B), (B, D), (D, C), (C, A)\}$ one has

$$[\beta_X + \beta_Y < 0] \quad \Leftrightarrow \quad [\psi_X + \psi_Y < \pi]. \tag{4.5.19}$$

Theorem 4.5.16. *Assume $\psi_X + \psi_Y < \pi$ for two adjacent vertices X and Y and let $\xi \in (-\pi, \pi]$ such that Assumption 4.5.13 holds. Then, there cannot exist two non-degenerate same cost graphs.*

The proof involves several case distinctions and is therefore split into numerous smaller Lemmas and Propositions. A schematic of the proof outline is given in Figure 4.35. The proof itself is stated further down, after all auxiliary results are assembled. Afterwards, some corollaries are given. A central object in the various proofs are the *cost quadrilaterals* that are introduced next.

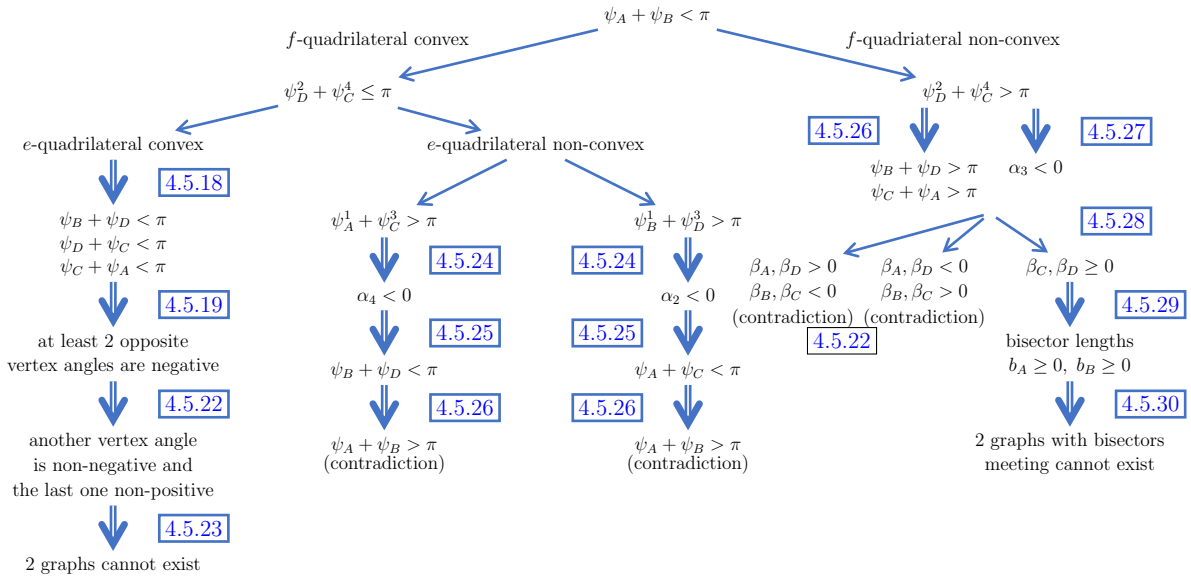

Figure 4.35: Outline of the proof of Theorem 4.5.16. Single-line arrows denote case distinction, double-line arrows denote implication. The numbers refer to the respective lemmas and propositions. The cost quadrilaterals are introduced below, see Figure 4.36.

Definition 4.5.17 (Cost quadrilaterals)**.** The quadrilateral obtained by joining the two cost triangles for the free vertices S_1 and S_3 at their common e-edge will be referred to as *e-quadrilateral*. Likewise, the f-quadrilateral is obtained by joining the cost triangles for free vertices S_2 and S_4 at their common f-edge. Both are sketched in Figure 4.36.

In each quadrilateral, the angles adjacent to the joining edge are obtained by combining two angles from the respective cost triangles. They will be referred to as *compound* angles and their value can range between $(0, 2\pi)$, potentially making the respective quadrilateral non-convex. The two angles opposite the joining edge will be referred to as *non-compound* and may take values in $(0, \pi)$.

In the e-quadrilateral we denote the length of the diagonal corresponding to the f-edge in the cost triangles by f', and likewise with e' in the f-quadrilateral.

We use the notation $(\widehat{x,y})_z$ to refer to the angle between edges of lengths x and y from the triangle with edges of lengths x, y, z. For example (see Figure 4.36), $(\widehat{a,e})_b = \psi_A^1$.

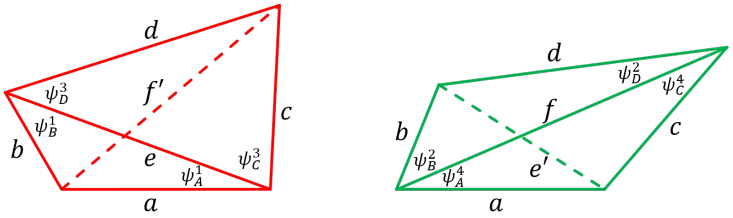

Figure 4.36: The two cost quadrilaterals of Definition 4.5.17.

Proposition 4.5.18. *Let $\psi_A + \psi_B < \pi$ and all the compound angles in both cost quadrilaterals be smaller or equal than π, that is $\psi_C^4 + \psi_D^2 \leq \pi$, $\psi_A^1 + \psi_C^3 \leq \pi$, and $\psi_B^1 + \psi_D^3 \leq \pi$. Then $\psi_A + \psi_C < \pi$, $\psi_B + \psi_D < \pi$, and $\psi_C + \psi_D < \pi$.*

Proof. The first condition implies that the angle between sides a and b in the e-quadrilateral is bigger than in the f-quadrilateral:

$$\psi_A + \psi_B < \pi \Leftrightarrow \psi_A^4 + \psi_B^2 < \pi - (\psi_A^1 + \psi_B^1) \Leftrightarrow (\widehat{a,b})_{e'} < (\widehat{a,b})_e.$$

By the law of cosines this implies that $e' < e$. The other conditions imply that both quadrilaterals are convex. The angles in a convex quadrilateral with four fixed edge lengths is fully determined by specifying either of the two diagonal lengths. Thus, by gradually decreasing e to e', the e-quadrilateral can be morphed into the f-quadrilateral.

Again, by the law of cosines, while e is decreasing, so will the angles $(\widehat{a,b})_e$ and $(\widehat{c,d})_e$ and therefore

$$(\widehat{c,d})_e > (\widehat{c,d})_{e'} \Leftrightarrow \pi - (\psi_C^3 + \psi_D^3) < \psi_C^4 + \psi_D^2 \Leftrightarrow \psi_C + \psi_D < \pi.$$

Since the sum of inner angles must remain fixed, this implies that the sum of the two compound angles must increase during the morphing. Each of these angles is tied to the length of the diagonal f' by the law of cosines. Therefore, since the quadrilateral remains convex during the morphing, both compound angles individually must strictly increase:

$$(\widehat{a,c})_{f'} < (\widehat{a,c})_f \Leftrightarrow \psi_A^1 + \psi_C^3 < \pi - (\psi_A^4 + \psi_C^4) \Leftrightarrow \psi_A + \psi_C < \pi,$$
$$(\widehat{b,d})_{f'} < (\widehat{b,d})_f \Leftrightarrow \psi_B^1 + \psi_D^3 < \pi - (\psi_B^2 + \psi_D^2) \Leftrightarrow \psi_B + \psi_D < \pi. \qquad \square$$

Proposition 4.5.19. *Let all pairs of adjacent vertex angles sum to a negative value: $\beta_A + \beta_B < 0$, $\beta_B + \beta_D < 0$, $\beta_D + \beta_C < 0$, $\beta_C + \beta_A < 0$. Then at least 2 opposite vertex angles are negative.*

Proof. This is proved by analyzing all relevant combinations of signs of the vertex angles.

- If all vertex angles are non-negative, all pairwise sums are non-negative.
- If only one vertex angle is negative, then pairwise sums that do not include this angle are non-negative. E.g. if only $\beta_A < 0$, then $\beta_B + \beta_D \geq 0$.

- If only two adjacent vertex angles are negative, the sum of the "opposite two" remains non-negative. E.g. if $\beta_A, \beta_B < 0$, then $\beta_C + \beta_D \geq 0$.

Therefore, at least two opposite vertex angles must be negative. □

Several of the following propositions rely on a corollary of the Farkas' Lemma, which we therefore state here.

Lemma 4.5.20 (Farkas' Lemma)**.** *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, with $b \neq 0$. Then exactly one of the following two assertions is true:*

- *there exists $x \in \mathbb{R}^n$ such that $Ax = b$ and $x \geq 0$,*
- *there exists $y \in \mathbb{R}^m$ such that $A^\intercal y \geq 0$ and $b^\intercal y < 0$.*

Corollary 4.5.21. *Let $A \in \mathbb{R}^{m\times n}$. Either there exists $x \in \mathbb{R}^n$ such that [$Ax = 0$, $x \geq 0$, $x \neq 0$], or there exists $y \in \mathbb{R}^m$ such that $A^\intercal y > 0$ (with the strict inequality holding for each component).*

Proof. Let $\tilde{A} = [A, 1_n] \in \mathbb{R}^{(m+1)\times n}$ be the matrix obtained by stacking A with the row vector $1_n \in \mathbb{R}^n$ with all entries one. Let $\tilde{b} = [0, \dots, 0, 1] \in \mathbb{R}^{m+1}$ be the vector with m zeros and a single 1. Existence of solutions $x \in \mathbb{R}^n$ to $Ax = 0$, $x \geq 0$, $x \neq 0$ is then equivalent to existence of solutions $x \in \mathbb{R}^n$, $\tilde{A}x = \tilde{b}$, $x \geq 0$. By Farkas' Lemma, either such a solution exists, or there exists some $\tilde{y} = [y, y^*] \in \mathbb{R}^{m+1}$ with $y \in \mathbb{R}^m$, $y^* \in \mathbb{R}$ such that $\tilde{A}^\intercal \tilde{y} = A^\intercal y + y^* \geq 0$ and $\tilde{b}^\intercal \tilde{y} = y^* < 0$. Clearly, such $\tilde{y}$ exists if and only if there exists some $y \in \mathbb{R}^m$ such that $A^\intercal y > 0$, with the inequality holding strictly for each component. □

Lemma 4.5.22. *Let $\xi \in (-\pi, \pi]$. Assume there exist two non-degenerate graphs and that $\beta_A, \beta_D < 0$ (i.e., two opposite vertex angles are negative). Then, either*

$$(\psi_A^1 < \psi_D^3,\ \psi_A^4 > \psi_D^2,\ \beta_C \leq 0) \quad \text{or} \quad (\psi_A^1 > \psi_D^3,\ \psi_A^4 < \psi_D^2,\ \beta_B \leq 0).$$

Additionally, in the respective cases, we have

$$\begin{aligned} &\text{if} \quad \psi_B^1 + \psi_D^3 \leq \pi && \text{if} \quad \psi_C^4 + \psi_D^2 \leq \pi \\ &\quad \text{then} \quad \beta_B \geq 0, && \quad \text{then} \quad \beta_C \geq 0. \end{aligned}$$

Proof.
Step 1. Find possible configurations.
From $\beta_A < 0$ we can find $\xi > \psi_A^1 + \psi_A^4$ and similarly from $\beta_D < 0$ we find $\xi > \psi_D^2 + \psi_D^3$. In particular, we have $\xi > 0$. Since there exist two non-degenerate graphs and ξ is positive, another boundary for ξ is given by Proposition 4.5.4: $\xi \leq \max\{\psi_A^1, \psi_D^3\} + \max\{\psi_A^4, \psi_D^2\}$. Now, if the right side of this inequality becomes $\psi_A^1 + \psi_A^4$ or $\psi_D^2 + \psi_D^3$, we get a contradiction with the inequalities we obtained from the givens. Therefore, the only configurations possible are $(\psi_A^1 < \psi_D^3,\ \psi_A^4 > \psi_D^2)$ or $(\psi_A^1 > \psi_D^3,\ \psi_A^4 < \psi_D^2)$. We are left to prove that in the first case existence of two graphs implies $\beta_C \leq 0$, while in the second case we must have $\beta_B \leq 0$.

Step 2. $(\psi_A^1 < \psi_D^3,\ \psi_A^4 > \psi_D^2) \implies \beta_C \leq 0$
Without loss of generality, consider the first configuration, $(\psi_A^1 < \psi_D^3,\ \psi_A^4 > \psi_D^2)$. For the other configuration a symmetric argument applies. Assume by contradiction to the statement of the lemma that $\beta_C > 0$.

We first consider the case $\beta_C \geq \pi/2$. We can write this down in terms of angle ξ as

$$\xi \geq 2\pi - \psi_C^3 - \psi_C^4 > \psi_D^3 + \psi_A^4,$$

where the last inequality follows from the cost triangles. However, the inequality above contradicts the upper bound for positive angle ξ from Proposition 4.5.4, which can be written as

$$\xi \leq \max\{\psi_A^1, \psi_D^3\} + \max\{\psi_A^4, \psi_D^2\} = \psi_D^3 + \psi_A^4$$

using that $\psi_A^1 < \psi_D^3$ and $\psi_A^4 > \psi_D^2$. We can therefore only focus on the case $\beta_C \in (0, \pi/2)$.

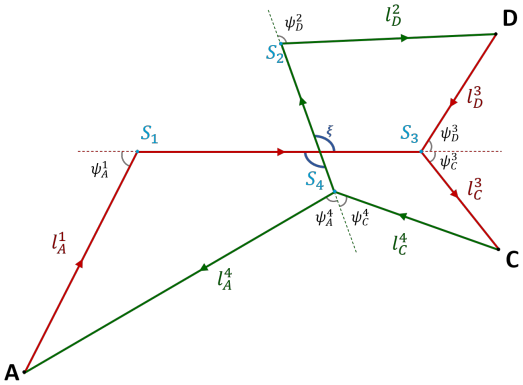

Figure 4.37: Loops (A, S_1, S_3, C, S_4, A) and (C, S_4, S_2, D, S_3, C).

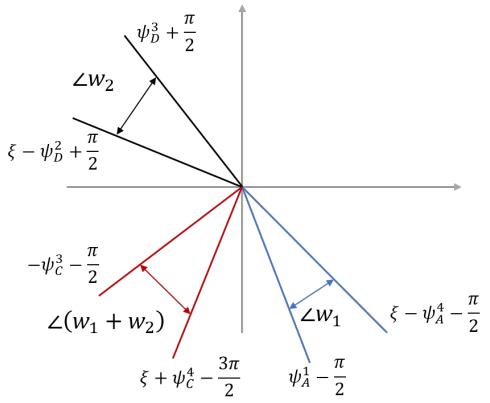

Figure 4.38: Cones of the dual vector components.

Now assume $\beta_C \in (0, \pi/2)$. Consider the bottom and the right loop, cf. (4.5.6), along the vertices (A, S_1, S_3, C, S_4, A) and (C, S_4, S_2, D, S_3, C). We will show that non-negative, non-trivial solutions to both loops simultaneously cannot exist if $\beta_C \in (0, \pi/2)$. Simultaneously considering solvability of two loops requires a somewhat elaborate invocation of Farkas' lemma (and its corollary). Similar as above, the loop equations can be written as

$$\begin{aligned} l_A^A \cdot v_A^4 + l_A^1 \cdot v_A^1 + l_E \cdot v_E + l_C^3 \cdot v_C^3 + l_C^4 \cdot v_C^4 = 0, \\ l_F \cdot v_F + l_D^2 \cdot v_D^2 + l_D^3 \cdot v_D^3 + l_C^3 \cdot v_C^3 + l_C^4 \cdot v_C^4 = 0, \end{aligned} \tag{4.5.20}$$

where we choose as orientations for the unit vectors along the segments

$$\begin{aligned} &\angle v_A^4 = \xi - \pi - \psi_A^4, && \angle v_A^1 = \psi_A^1, && \angle v_E = 0, \\ &\angle v_F = \xi, && \angle v_D^2 = \xi - \psi_D^2, && \angle v_D^3 = \psi_D^3 - \pi, \\ &\angle v_C^3 = -\psi_C^3, && \angle v_C^4 = \xi + \psi_C^4. \end{aligned}$$

Of course, these equations hold up to adding multiples of 2π and in the following we will need to be vigilant when working with intervals of angles. Both loops with chosen segment orientations are shown in Figure 4.37.

By Corollary 4.5.21, a non-negative, non-zero solution to (4.5.20) does not exist if there exists some $y \in \mathbb{R}^4$ such that $A^\intercal y > 0$ (element-wise) where

$$A = \begin{pmatrix} v_A^4 & v_A^1 & v_E & & & v_C^3 & v_C^4 \\ & & & v_F & v_D^2 & v_D^3 & v_C^3 & v_C^4 \end{pmatrix} \in \mathbb{R}^{4\times 8}.$$

This corresponds to finding two vectors $w_1, w_2 \in \mathbb{R}^2$ (and then setting $y = (w_1, w_2)$ to be the concatenation) such that

$$v^\intercal w_1 > 0 \text{ for } v \in \{v_A^4, v_A^1, v_E\}, \quad v^\intercal w_2 > 0 \text{ for } v \in \{v_F, v_D^2, v_D^3\}, \quad v^\intercal(w_1 + w_2) > 0 \text{ for } v \in \{v_C^3, v_C^4\}.$$

We start by considering the first condition for w_1. By the condition $\beta_A < 0$ we have that $\psi_A^1 - (\xi - \pi - \psi_A^4) \in (0, \pi)$. Moreover, $0 \in (\xi - \pi - \psi_A^4, \psi_A^1)$. So v_A^1 and v_A^4 span a cone with angle strictly less than π, with v_E contained in this cone. Therefore, for w_1 to have strictly positive inner product with v_A^4, v_A^1 and v_E, it must have strictly positive inner product with the former two, i.e. it must lie in the interior of the dual cone of $\{v_A^4, v_A^1\}$. So its orientation must satisfy

$$\angle w_1 \in (\psi_A^1 - \pi/2, \xi - \pi/2 - \psi_A^4).$$

See Figure 4.38 for an illustration.

By analogous reasoning for the second condition, for w_2 to have strictly positive inner product with v_D^2, v_D^3, v_F, it must lie in the interior of the dual cone of $\{v_D^2, v_D^3, v_F\}$, which equals (the interior of) the dual cone of $\{v_D^2, v_D^3\}$. This imposes the condition

$$\angle w_2 \in (\psi_D^3 + \pi/2, \xi - \psi_D^2 + \pi/2).$$

Next, again using the same reasoning, we find that $w_1 + w_2$ must lie in the interior of the dual cone of $\{v_C^3, v_C^4\}$, and using $\beta_C \in (0, \pi/2)$, this can be shown to mean

$$\angle(w_1 + w_2) \in (-\psi_C^3 - \pi/2, \xi + \psi_C^4 - 3\pi/2).$$

Such w_1, w_2 exist if and only if the sum of the interiors of the dual cones of $\{v_A^1, v_A^4\}$ and $\{v_D^2, v_D^3\}$ intersect with the interior of the dual cone of $\{v_C^3, v_C^4\}$.

Using that $\psi_A^1 < \psi_D^3$ one obtains that

$$(\psi_A^1 - \pi/2 + 2\pi) - (\psi_D^3 + \pi/2) = \pi + \psi_A^1 - \psi_D^3 \in (0, \pi).$$

Note that we added 2π to the orientation of v_A^1 here to get a meaningful interval. This means that the sum of the interior of the two dual cones contains at least the interval $(\psi_D^3 + \pi/2, \xi - \psi_A^4 + 3\pi/2)$. Using that

$$\beta_D < 0 \Rightarrow \xi > \psi_D^3 \Rightarrow \xi > \psi_D^3 - \psi_C^4 \Leftrightarrow \xi + \psi_C^4 + \pi/2 > \psi_D^3 + \pi/2,$$
$$\beta_A < 0 \Rightarrow \xi > \psi_A^4 \Rightarrow \xi > \psi_A^4 - \psi_C^3 \Leftrightarrow \xi - \psi_A^4 + 3\pi/2 > -\psi_C^3 + 3\pi/2,$$

we find that the interior of the latter cone is contained in the interior of the sum of the former two and thus by Farkas' lemma, a non-trivial solution to (4.5.20) cannot exist.

Step 3. $(\psi_A^1 < \psi_D^3,\ \psi_A^4 > \psi_D^2,\ \psi_B^1 + \psi_D^3 \leq \pi) \implies \beta_B \geq 0$

This is essentially a repetition of the previous argument, using now the top and right loops instead, and showing that for $\beta_B < 0$ a non-existence certificate for the solvability of these two loops can be obtained via Farkas' Lemma. □

Proposition 4.5.23. *There exists no choice of costs $a, b, c, d, e, f > 0$ such that the corresponding cost triangles are non-degenerate and $\psi_A + \psi_B < \pi$, $\psi_B + \psi_D < \pi$, $\psi_D + \psi_C < \pi$, $\psi_C + \psi_A < \pi$, $\psi_C \leq \psi_B$, $\psi_A^1 < \psi_D^3$ and $\psi_A^4 > \psi_D^2$ (see Lemma 4.5.22 for these conditions).*

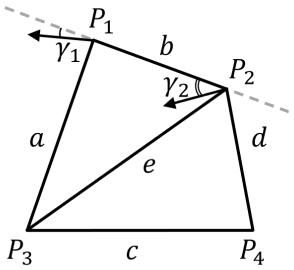

Figure 4.39: Cost quadrilateral with introduced rotation nomenclature.

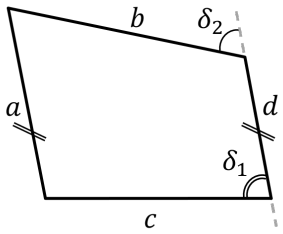

Figure 4.40: Cost quadrilateral with a and d-segments parallel.

Proof. The proof works by contradiction, showing that the assumed inequalities for the ψ-angles cannot all be satisfied simultaneously.

By the assumptions both quadrilaterals are convex, since $\psi_A^1 + \psi_C^3 < \psi_A + \psi_C < \pi$ and likewise for all other compound angles. Since $\pi - \psi_D^3 - \psi_C^3 > \psi_C^4 + \psi_D^2$ one has $e > e'$. We can therefore, as earlier, morph the e-quadrilateral into the f-quadrilateral by gradually decreasing e to e'. Let this be parametrized by some $[0,1] \ni t \mapsto e(t)$ that continuously and strictly monotonously interpolates between $e(0) = e$ and $e(1) = e'$. (By abuse of notation we will in the following usually drop the explicit time-dependency from the notation for simplicity and merely write e.) Figure 4.39 provides some notation and intuition for the following arguments. During the morphing the angle $(\widehat{c,d})_e$ strictly decreases, the angle $(\widehat{a,c})_f$ strictly increases. Hence, points P_1 and P_2 rotate counter-clockwise around P_3 and P_4 on circles of radii a and d respectively. Denote by v_1 and v_2 the velocities of points P_1 and P_2. They must be perpendicular to the faces a and d respectively, pointing in the counter-clockwise direction. Denote by γ_1 and γ_2 the signed angles between v_1, v_2 and $P_1 - P_2$. By convexity of the quadrilateral (which is preserved during the morphing), one must have $\gamma_1, \gamma_2 \in (-\pi/2, \pi/2)$ and we adopt the sign convention that $\gamma_i > 0$ if v_i is rotated counter-clockwise relative to $P_1 - P_2$, cf. Figure 4.39. Since the length b of the quadrilateral must be preserved during the morphing, the components of v_1 and v_2 that are parallel to $P_1 - P_2$ must coincide, i.e.

$$\|v_1\| \cos(\gamma_1) = \|v_2\| \cos(\gamma_2).$$

So one obtains for the perpendicular component of $v_1 - v_2$, denoted by v_{perp},

$$v_{\text{perp}} = (v_1 - v_2)^\intercal (P_1 - P_2)/b = \|v_1\| \sin(\gamma_1) - \|v_2\| \sin(\gamma_2) = \|v_1\| \cos(\gamma_1) \cdot (\tan(\gamma_1) - \tan(\gamma_2)).$$

By monotonicity of tan one has that v_{perp} has the same sign as $\gamma_1 - \gamma_2$. If $v_{\text{perp}} < 0$, then the orientation of the b-segment is rotating clockwise, which corresponds to the sum of angles $(\widehat{c,d})_e + (\widehat{d,b})_f$ increasing.

At $t = 0$ one finds $\gamma_1 = \pi/2 - \psi_A^4 - \psi_B^2$ and $\gamma_2 = \pi/2 - \psi_B^2 - \psi_D^2$ and therefore $\gamma_1 - \gamma_2 = \psi_D^2 - \psi_A^4 < 0$ by assumptions. Likewise, for $t = 1$ one obtains $\gamma_1 - \gamma_2 = \psi_A^1 - \psi_D^3 < 0$. We will show below, that indeed $\gamma_1 - \gamma_2 \leq 0$ for all $t \in [0,1]$, and thus the b-edge rotates clockwise and $(\widehat{c,d})_e + (\widehat{d,b})_f$ is increasing when morphing from the e to the f-quadrilateral. However, this would imply $\psi_B^1 - \psi_C^3 < \psi_C^4 - \psi_B^2$, which contradicts the assumption $\psi_C \leq \psi_B$. Therefore, such cost coefficients cannot exist, as the corresponding quadrilaterals would have contradicting properties.

We now rule out that the sign of $\gamma_1 - \gamma_2$ changes in the interval $t \in (0,1)$. As $\gamma_1 - \gamma_2$ is continuous in t, for a sign change one must have an instant where $\gamma_1 = \gamma_2$, which corresponds to

the a and d edges being parallel. We show that unless $a = d \wedge b = c$ (i.e. the quadrilateral is a parallelogram, and stays so during morphing), this can happen at most for one single value of t, i.e. in our case $\gamma_1 - \gamma_2$ cannot change it sign from being negative to positive and back within the interval $t \in [0, 1]$. The case of a parallelogram can be ruled out since $\gamma_1 - \gamma_2 \neq 0$ for $t \in \{0, 1\}$. Introduce now the auxiliary angles δ_1 and δ_2 as in Figure 4.40. For the quadrilateral to be closed, and a and d-segments to be parallel the two following conditions must be satisfied:

$$c \sin(\delta_1) = b \sin(\delta_2), \qquad d - a = c \cos(\delta_1) - b \cos(\delta_2).$$

By convexity (and non-degeneracy of the cost triangles) we can restrict to potential solutions with $(\cos(\delta_1), \cos(\delta_2)) \in (-1, 1)^2$. By elementary curve sketching arguments it is then possible to show that unless the quadrilateral is a parallelogram, there can be at most a single solution (δ_1, δ_2), which corresponds to a single time t. □

Lemma 4.5.24. *Let $\psi_A + \psi_B < \pi$ and $\psi_D^2 + \psi_C^4 \leq \pi$. In this setting, if $\psi_A^1 + \psi_C^3 > \pi$ then $\alpha_4 < 0$. If $\psi_B^1 + \psi_D^3 > \pi$ then $\alpha_2 < 0$.*

Proof. First note that at most one of the two conditions $\psi_A^1 + \psi_C^3 > \pi$ or $\psi_B^1 + \psi_D^3 > \pi$ can be satisfied at the same time: If both are satisfied, then $\psi_A^1 + \psi_B^1 + \psi_C^3 + \psi_D^3 > 2\pi$, which is excluded by the cost triangles.

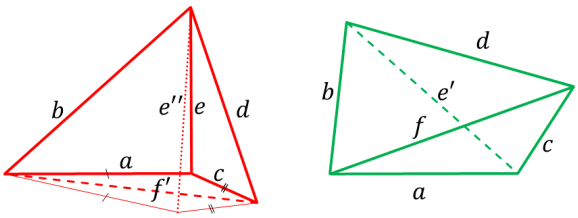

Figure 4.41: Cost quadrilaterals: $\psi_A^1 + \psi_C^3 > \pi$.

We now assume $\psi_A^1 + \psi_C^3 > \pi$. This means that the e-quadrilateral is non-convex due to the vertex where the a and c edge meet (while the other three angles are each strictly less than π). We add an auxiliary edge f' to the e-quadrilateral and reflect edges a and c to construct a new convex quadrilateral with edges a, b, c, d (see Figure 4.41). We then add edge e'' to the new convex quadrilateral and edge e' to the f-quadrilateral, as shown in the figure.

The angle formed by the reflected edge a and edge b in the e-quadrilateral is larger than the angle between the original edge a and the same edge b: $(\widehat{a, b})_{e''} > (\widehat{a, b})_e$, while the latter is larger than the corresponding angle in the f-quadrilateral, due to the assumption $\psi_A + \psi_B < \pi$:

$$(\widehat{a, b})_e = \pi - (\psi_A^1 + \psi_B^1) > \psi_A^4 + \psi_B^2 = (\widehat{a, b})_{e'}$$

The new convex quadrilateral can be morphed into the f-quadrilateral by decreasing e'' to e, or equivalently, by increasing f' to f, in the process of which the angle between (the reflected) a and c increases. Therefore,

$$(\widehat{a, c})_{f'} < (\widehat{a, c})_f \Leftrightarrow 2\pi - \psi_A^1 - \psi_C^3 < \pi - \psi_A^4 - \psi_C^4 \Leftrightarrow$$
$$\psi_C^4 + \psi_A^4 - \psi_C^3 - \psi_A^3 + \pi < 0 \Leftrightarrow \alpha_4 < 0.$$

The case $\psi_B^1 + \psi_D^3 > \pi$ can be treated with a symmetric argument. □

Lemma 4.5.25. *Assume $\alpha_1 < 0$. Then the vertex angles in the opposite sector sum to a negative value: $\beta_C + \beta_D < 0$.*

Note that the opposite statement (from negative sum to negative alpha) is true only with an additional condition on the angles (see Lemma 4.5.27).

Proof. By definition, (4.5.13), one has

$$\alpha_1 = \frac{1}{2}\left(\psi_A^1 + \psi_B^1 - \psi_A^4 - \psi_B^2 + \pi\right),$$

and from the non-degenerate cost triangles we conclude $\psi_A^1 + \psi_B^1 > 0$. Therefore, if $\alpha_1 < 0$, then $\pi - \psi_A^4 - \psi_B^2 < 0$. Figure 4.42 shows the two cost quadrilaterals with $\psi_A^4 + \psi_B^2 > \pi$ with an additional edge e' in the f-quadrilateral.

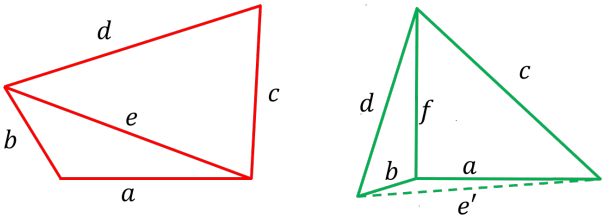

Figure 4.42: Cost quadrilaterals, $\psi_A^4 + \psi_B^2 > \pi$.

From condition $\alpha_1 < 0$ we now infer that

$$0 > \alpha_1 = \frac{1}{2}\left(\psi_A^1 + \psi_B^1 - \psi_A^4 - \psi_B^2 + \pi\right) \Leftrightarrow 2\pi - \psi_A^4 - \psi_B^2 < \pi - \psi_A^1 - \psi_B^1 \Leftrightarrow$$
$$(\widehat{a,b})_{e'} < (\widehat{a,b})_e \Leftrightarrow e' < e \Leftrightarrow (\widehat{c,d})_{e'} < (\widehat{c,d})_e \Leftrightarrow$$
$$\psi_C^4 + \psi_D^2 < \pi - \psi_C^3 - \psi_D^3 \Leftrightarrow \psi_C^3 + \psi_D^3 + \psi_C^4 + \psi_D^2 < \pi < 0 \Leftrightarrow \beta_C + \beta_D < 0. \qquad \square$$

Proposition 4.5.26. *Let $\psi_A + \psi_B < \pi$ and $\psi_C^4 + \psi_D^2 > \pi$. Then $\psi_A + \psi_C > \pi$, $\psi_B + \psi_D > \pi$, and $\psi_C + \psi_D > \pi$.*

Proof. The second condition $\psi_C^4 + \psi_D^2 > \pi$ immediately implies $\psi_C + \psi_D > \pi$. It therefore remains to prove that $\psi_A + \psi_C > \pi$ and $\psi_B + \psi_D > \pi$ hold. We do this by showing that the two cost quadrilaterals have contradicting properties if at least one of the two inequalities fails.

Case 1: assume $\psi_A + \psi_C \leq \pi$ and $\psi_B + \psi_D > \pi$
These two inequalities are equivalent respectively to

$$\psi_A^1 + \psi_C^3 \leq \pi - (\psi_A^4 + \psi_C^4), \tag{4.5.21}$$
$$\psi_B^1 + \psi_B^3 > \pi - (\psi_B^2 + \psi_D^2). \tag{4.5.22}$$

By non-degeneracy of the cost triangles, we get from (4.5.21) that $\psi_A^1 + \psi_C^3 < \pi$. Hence, the e-quadrilateral can be either convex or not depending on the value of $\psi_B^1 + \psi_D^3$. The f-quadrilateral is non-convex by assumption, due to the corner between the c and d segments. The cost quadrilaterals in their potential configurations are shown in Figure 4.43, where we added edges e' and f' as before.

By (4.5.21) one has

$$(\widehat{a,c})_f = \pi - \psi_A^4 - \psi_C^4 \geq \psi_A^1 + \psi_C^3 = (\widehat{a,c})_{f'}$$

and therefore $f \geq f'$. Consider now the two possible configurations of the e-quadrilateral.

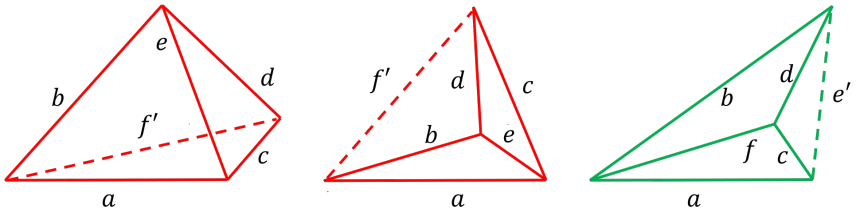

Figure 4.43: Case 1: Possible configurations of the quadrilaterals.

- If the e-quadrilateral is convex, i.e. $\psi_B^1+\psi_D^3 \leq \pi$ (see Figure 4.43, left), then $f \geq f'$ implies

$$\pi - (\psi_B^2 + \psi_D^2) = (\widehat{b,d})_f \geq (\widehat{b,d})_{f'} = \psi_B^1 + \psi_B^3,$$

  which contradicts (4.5.22).

- If the e-quadrilateral is non-convex, i.e. $\psi_B^1 + \psi_D^3 > \pi$ (see Figure 4.43, middle), then $f \geq f'$ implies

$$\pi - (\psi_B^2 + \psi_D^2) = (\widehat{b,d})_f \geq (\widehat{b,d})_{f'} = 2\pi - \psi_B^1 - \psi_D^3,$$

  which implies $\alpha_2 \leq 0$. Similarly, recalling that $\psi_A + \psi_B < \pi$ and considering the triangles bae' and bae, one proves $e' < e$. Looking now at triangles dce' and dce, the estimate $e' < e$ implies $\alpha_4 < 0$. That means we have two non-positive splitting angles α_i, which contradicts Proposition 4.5.9.

So the combination $\psi_A + \psi_C \leq \pi$ and $\psi_B + \psi_D > \pi$ is excluded and by a symmetric argument one can rule out the reverse situation, which is $\psi_A + \psi_C > \pi$ and $\psi_B + \psi_D \leq \pi$.

Case 2: assume $\psi_A + \psi_C \leq \pi$ and $\psi_B + \psi_D \leq \pi$
In this case the e-quadrilateral is convex, and as before let f' be the second diagonal in the e-quadrilateral. Since the f-quadrilateral is concave, to morph it into the e-quadrilateral one must increase f, i.e. $f < f'$. But at the same time

$$(\widehat{a,c})_f = \pi - \psi_A^4 - \psi_C^4 \geq \psi_A^1 + \psi_C^3 = (\widehat{a,c})_{f'},$$

i.e. $f \geq f'$ by the law of cosines, which is a contradiction. □

Lemma 4.5.27. *Let $\psi_A + \psi_B < \pi$ and $\psi_C^4 + \psi_D^2 > \pi$. Then $\alpha_3 < 0$.*

Proof. The assumption $\psi_C^4 + \psi_D^2 > \pi$ implies that the f-quadrilateral is non-convex, due to the corner between the c and d edge. As before, we add the e' edge, which lies outside of the quadrilateral. By assumption we have

$$\psi_A + \psi_B < \pi \Leftrightarrow \psi_A^4 + \psi_B^2 < \pi - (\psi_A^1 + \psi_B^1) \Leftrightarrow (\widehat{a,b})_{e'} < (\widehat{a,b})_e$$

and therefore $e' < e$, and finally $(\widehat{c,d})_{e'} < (\widehat{c,d})_e$. By non-convexity of the f-quadrilateral this means that

$$2\pi - \psi_D^2 - \psi_C^4 = (\widehat{c,d})_{e'} < (\widehat{c,d})_e = \pi - \psi_D^3 - \psi_C^3,$$

which implies that $\alpha_3 < 0$. □

Lemma 4.5.28. *Assume $\beta_A + \beta_B < 0$ and $\beta_B + \beta_D > 0$, $\beta_D + \beta_C > 0$, $\beta_C + \beta_A > 0$, and $\beta_X \neq 0$, $X \in \{A, B, C, D\}$, then one of the following must be true:*

1.	$\beta_A < 0$,	$\beta_B < 0$,	$\beta_C > 0$,	$\beta_D > 0$,
2.	$\beta_A < 0$,	$\beta_B > 0$,	$\beta_C > 0$,	$\beta_D < 0$,
3.	$\beta_A < 0$,	$\beta_B > 0$,	$\beta_C > 0$,	$\beta_D > 0$,
4.	$\beta_A > 0$,	$\beta_B < 0$,	$\beta_C < 0$,	$\beta_D > 0$,
5.	$\beta_A > 0$,	$\beta_B < 0$,	$\beta_C > 0$,	$\beta_D > 0$.

Proof. This is easily verified by checking all 16 possible combinations of signs. □

Proposition 4.5.29. *Assume there exist two non-degenerate same cost graphs with $\beta_A + \beta_B < 0$, $\beta_C, \beta_D > 0$ and $\alpha_3 < 0$. Assume that Assumption 4.5.13 holds. Then bisector lengths $b_A \geq 0$ and $b_B \geq 0$.*

Proof. Since there exist two non-degenerate same cost graphs, by Theorem 4.5.15 all bisectors meet in a single point and thus the bisector lengths for any vertex induced by the two adjacent sectors coincide, see Definition 4.5.10, and can be denoted by $b_A, b_B, b_C, d_D \in \mathbb{R}$. Since $\alpha_3 < 0$, Proposition 4.5.9 provides that $\alpha_2 \in (0, \pi)$. Hence, by Definition 4.5.10 we have

$$\begin{bmatrix} b_B \\ b_D \end{bmatrix} = \frac{1}{\sin \alpha_2} \begin{bmatrix} \sin(\beta_B + \alpha_2) & \sin(\beta_D) \\ \sin(\beta_B) & \sin(\beta_D + \alpha_2) \end{bmatrix} \begin{bmatrix} l_B^2 \\ l_D^2 \end{bmatrix}, \tag{4.5.23}$$

for some lengths $l_B^2, l_D^2 \geq 0$. We are interested in showing that $b_B \geq 0$. Since $\alpha_2 \in (0, \pi)$, we need to check that

$$\sin(\beta_B + \alpha_2)\, l_B^2 + \sin(\beta_D)\, l_D^2 \geq 0.$$

Clearly $\sin(\beta_D) \geq 0$, and we claim that $\beta_B + \alpha_2 \in [0, \pi]$. Indeed, from $\beta_D \geq 0$ we deduce $\xi \leq \psi_D^2 + \psi_D^3$ and estimate

$$\beta_B + \alpha_2 = \frac{\xi + 2\psi_B^2 + \psi_D^2 - \psi_D^3}{2} \leq \psi_B^2 + \psi_D^2 < \pi,$$

where the last inequality follows from non-degeneracy of cost triangles. On the other hand, we know that $\psi_D^2 - \psi_D^3 > \psi_C^3 - \psi_C^4 + \pi$ because $\alpha_3 < 0$ and also $\xi \geq \pi - \psi_C$ because $\beta_C \geq 0$, so that we can estimate

$$\beta_B + \alpha_2 = \frac{\xi + 2\psi_B^2 + \psi_D^2 - \psi_D^3}{2} > \frac{\pi - \psi_C^3 - \psi_C^4 + 2\psi_B^2 + \psi_C^3 - \psi_C^4 + \pi}{2} = \pi + \psi_B^2 - \psi_C^4 > 0$$

by non-degeneracy of the cost triangles. This proves our claim, and so $b_B \geq 0$. The analogue argument for b_A via the loop in sector 4 yields $b_A \geq 0$. □

Proposition 4.5.30. *Let $\beta_A + \beta_B < 0$, $\beta_C, \beta_D > 0$ and $\alpha_3 < 0$, $\psi_A + \psi_B < \pi$, assume that two non-degenerate same cost graphs exist and that Assumption 4.5.13 holds. Then the bisector lengths b_A, b_B must be negative.*

Proof. By Assumption 4.5.13, $\psi_A + \psi_B < \pi$, and the existence of two non-degenerate same cost graphs, by Theorem 4.5.15 the four bisectors must meet in a single point, and thus the two bisector lengths for each fixed vertex induced by any of the two adjacent sectors must coincide.

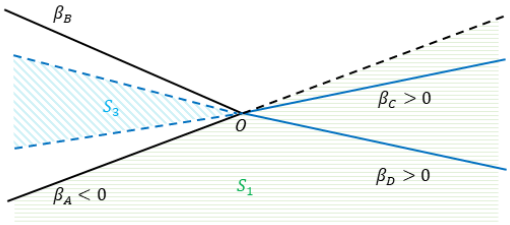

Figure 4.44: Possible locations of S_1 and S_3.

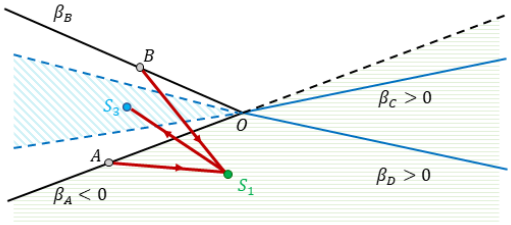

Figure 4.45: Invalid triple junction at S_1.

By $\beta_A + \beta_B < 0$ at least one of the two vertex angles must be negative. By symmetry we can assume without loss of generality that $\beta_A < 0$.

Given that $\alpha_3 < 0$ and $0 < \alpha_1, \alpha_2, \alpha_4 < \pi$ (Proposition 4.5.9), the general configuration of the bisectors lines is described in Figure 4.44 (note that the order of the C and D bisector lines is swapped). We first look at the location of point S_3, the meeting point of C- and D-branches of the e-graph. Using that $\beta_C, \beta_D > 0$ (and therefore in $(0, \pi)$, see Proposition 4.5.8), we observe that point S_3 has to lie above the C-bisector line and below the D-bisector line. Taking into account that $\alpha_2 < \pi$ and $\alpha_4 < \pi$, the open cone defined by these two conditions is entirely contained in the cone spanned by the A- and B-bisector lines. Hence, point S_3 is contained in the interior of the cone spanned by the A- and B-bisector lines.

As for the point S_1, because angle $\beta_A < 0$ (and therefore in $(-\pi/2, 0)$ by Proposition 4.5.8), point S_1 must be below or on the A-bisector line.

If b_A, b_B were non-negative, points A and B would lie on the non-negative part of the A- and B-bisector lines. Due to $\alpha_1 < \pi$ the latter is above the former, see Figure 4.45. Since S_3 must lie in the interior of the AB-cone, and by the non-degeneracy of Definition 4.5.3 A and B must be distinct, the three points A, B and S_3 are distinct.

In the e-graph, line segments from A, B and S_3 will meet at S_1 to form a triple junction. We have just shown that under the given assumptions, A, B and S_3 are distinct and will lie on or above the A-bisector line, whereas S_1 will lie below it. But this contradicts the momentum preservation condition (4.2.20) that needs to be satisfied at free vertices. Hence, the assumed situation cannot occur. □

We can now finally state the proof for the main result of this section.

Proof of Theorem 4.5.16. By symmetry (or by relabeling the cost coefficients) we may without loss of generality assume that $(X, Y) = (A, B)$ in the statement of the theorem. The proof follows the schematic presented in Figure 4.35. We first distinguish the case where the f-quadrilateral is convex and the case when it is not convex. As $\psi_A + \psi_B < \pi$ by assumption, the only way convexity can be violated is if $\psi_D^2 + \psi_C^4 > \pi$.

Let us first assume $\psi_D^2 + \psi_C^4 \leq \pi$, i.e. that the f-quadrilateral is convex. If in addition the e-quadrilateral is also convex, we can use Proposition 4.5.18 to show that all pairs $\psi_B + \psi_D < \pi$, $\psi_D + \psi_C < \pi$, and $\psi_C + \psi_A < \pi$. Then, by Proposition 4.5.19, we conclude that at least 2 opposite vertex angles are strictly negative. Assume these angles are β_A and β_D (the argument for β_B and β_C is symmetric). Then, by Lemma 4.5.22, we have either ($\psi_A^1 < \psi_D^3$, $\psi_A^4 > \psi_D^2$, $\beta_C \leq 0$, $\beta_B \geq 0$) or ($\psi_A^1 > \psi_D^3$, $\psi_A^4 < \psi_D^2$, $\beta_C \geq 0$, $\beta_B \leq 0$). Assume the former (again, the latter can be treated by symmetric arguments). Then $\beta_C \leq \beta_B$ implies $\psi_C \leq \psi_B$ and therefore Proposition 4.5.23 implies a contradiction.

Now consider the case when the f-quadrilateral is convex, but the e-quadrilateral is not. The latter can be due to two pairs of angles: $\psi_A^1 + \psi_C^3 > \pi$ or $\psi_B^1 + \psi_D^3 > \pi$. As above, we assume that the former is true, the proof for the latter case is symmetric. First, by Lemma 4.5.24, splitting angle α_4 is negative. Then, by Lemma 4.5.25, we have $\psi_B + \psi_D < \pi$. Finally, we can apply Proposition 4.5.26 (but applied to a different sector than stated), which implies that $\psi_A + \psi_B > \pi$, which contradicts the assumptions. This concludes the case when the f-quadrilateral is convex.

Assume now that the f-quadrilateral is not convex since $\psi_D^2 + \psi_C^4 > \pi$. Then by Lemma 4.5.26 one has $\psi_B + \psi_D > \pi$, $\psi_D + \psi_C > \pi$ and $\psi_C + \psi_A > \pi$. This means that $\beta_B + \beta_D > 0$, $\beta_D + \beta_C > 0$, $\beta_C + \beta_A > 0$, while $\beta_A + \beta_B < 0$. This leaves only five possible combinations of vertex angles signs (see Proposition 4.5.28). Two of these combinations ($\beta_A, \beta_D > 0$, $\beta_B, \beta_C < 0$ and $\beta_A, \beta_D < 0$, $\beta_B, \beta_C > 0$) are impossible, because the conditions of these cases imply a contradicting vertex angles signs combination by Proposition 4.5.22. We therefore only need to consider the other 3 combinations, where $\beta_C, \beta_D > 0$ and $\beta_A + \beta_B < 0$. We now note that from the conditions of the case, by Lemma 4.5.27, splitting angle α_3 is negative. We can then apply Proposition 4.5.29 to get information about prescribed signs of bisector lengths. Finally, we can apply Proposition 4.5.30, to see that the construction of two graphs with prescribed geometry and meeting bisectors is impossible. □

Theorem 4.5.16 allows to obtain additional insight into the properties of non-degenerate equal cost graphs that are given in the following corollaries.

Corollary 4.5.31. *Let $\xi \in (-\pi, \pi]$, assume there exists two non-degenerate same cost graphs and that Assumption 4.5.13 holds. Then, the following configurations of angles are impossible:*

1. *Two adjacent vertex angles are both negative.*
2. *Two opposite vertex angles are both negative.*
3. *The sum of two adjacent vertex angles is negative.*
4. *A splitting angle α_i is negative.*
5. *The sum of all ψ-angles satisfies $\sum \psi_X^i < 2\pi$, where the sum runs over*

$$(X, i) \in \{(A, 1), (A, 4), (B, 1), (B, 2), (D, 2), (D, 3), (C, 3), (C, 4)\}.$$

Proof.

1. For two adjacent vertices X, Y one has $[\beta_X + \beta_Y < 0] \Leftrightarrow [\psi_X + \psi_Y < \pi]$ and thus if the first sum is negative, then by Theorem 4.5.16 no solution exists.
2. If two opposite vertex angles are negative, by Lemma 4.5.22, a third vertex angle is also negative. This leads to two negative adjacent vertex angles and hence to the first case.
3. In the first case, in fact not both vertex angles need to be negative, but their sum being negative is already enough to apply the argument.
4. If some α_i were negative, by Lemma 4.5.25 the vertex angles of the two opposite vertices would sum to a negative value, which is the previous case.
5. If the sum of all ψ-angles is less than 2π, then there must be two adjacent vertices X, Y such that $\psi_X + \psi_Y < \pi$, which allows to invoke Theorem 4.5.16. □

Corollary 4.5.32. *Assume there exist two non-degenerate same cost graphs and that Assumption 4.5.13 holds. Then, all bisector lengths b_X have to be non-negative.*

Proof. Since the two graphs are non-degenerate and have equal cost, their bisectors meet in a single point (Theorem 4.5.15) and thus the bisector lengths for each vertex induced by the two adjacent sectors must coincide.

By Corollary 4.5.31 and Assumption 4.5.13, all α_i must be strictly positive, by Proposition 4.5.9 they must lie in $(0, \pi)$. According to Corollary 4.5.31, at most one vertex angle can be strictly negative, while by Proposition 4.5.8 all vertex angles lie in $(-\pi/2, \pi)$. We split the proof in two steps: first we assume all vertex angles to be positive, then we treat the case where one of them is strictly negative.

Case 1: all vertex angles positive

Without loss of generality, consider b_A, as defined via the first sector via Definition 4.5.10. The other lengths can be handled in the same fashion. One has

$$b_A = \frac{1}{\sin(\alpha_1)} \left(\sin\left(\beta_A + \alpha_1\right) \cdot l_A^1 + \sin\left(\beta_B\right) \cdot l_B^1 \right). \tag{4.5.24}$$

Hence, since $\beta_A, \beta_B \in (0, \pi)$, non-negativity of b_A follows as soon as $\sin\left(\beta_A + \alpha_1\right) \geq 0$, which is equivalent to $\beta_A + \alpha_1 \in [0, \pi]$. This is indeed satisfied because

$$\beta_A + \beta_B + \alpha_1 = \frac{\psi_A^1 + \psi_A^4 - \xi + \xi + \psi_B^1 + \psi_B^2 - \pi + \psi_A^1 + \psi_B^1 - \psi_A^4 - \psi_B^2 + \pi}{2} = \psi_A^1 + \psi_B^1 < \pi.$$

Case 2: one vertex angle is negative

Assume without loss of generality that the negative angle is β_C, so $\beta_C < 0$. By Corollary 4.5.31, Assumption 4.5.13, and Proposition 4.5.8 the other three must be strictly positive, contained in $(0, \pi)$. Therefore, we can apply the argument from the first case to the bisector lengths in sectors 1 and 2, getting that $b_A, b_B, b_D \geq 0$. For b_C we get two equations, one each from sectors 3 and 4:

$$b_C = \frac{1}{\sin(\alpha_3)} \left(\sin\left(\beta_C + \alpha_3\right) \cdot l_C^3 + \sin\left(\beta_D\right) \cdot l_D^3 \right)$$
$$b_C = \frac{1}{\sin(\alpha_4)} \left(\sin\left(\beta_A\right) \cdot l_A^4 + \sin\left(\beta_C + \alpha_4\right) \cdot l_C^4 \right)$$

Since all graph lengths, $\sin(\alpha_3)$, $\sin(\alpha_4)$, $\sin(\beta_A)$, and $\sin(\beta_D)$ are all positive b_C can only be negative if $\sin(\beta_C + \alpha_3) < 0$ and $\sin(\beta_C + \alpha_4) < 0$, which (since $\alpha_i \in (0, \pi)$ and $\beta_C \in (-\pi/2, 0)$) is equivalent to $\beta_C + \alpha_3 < 0$ and $\beta_C + \alpha_4 < 0$, meaning that $2\beta_C + \alpha_3 + \alpha_4 < 0$. On the other hand, by direct computations, we have

$$2\beta_C + \alpha_3 + \alpha_4 = \xi + \frac{\psi_C + \psi_A}{2} + \frac{\psi_C + \psi_D}{2} - \psi_D^2 - \psi_A^1.$$

We observe now that $\beta_B \geq 0$ provides $\xi \geq \pi - \psi_B^1 - \psi_B^2$ and Theorem 4.5.16 provides $\psi_C + \psi_A \geq \pi$ and $\psi_C + \psi_D \geq \pi$ (otherwise two non-degenerate same cost graphs would not exist). Hence, we can estimate

$$2\beta_C + \alpha_3 + \alpha_4 \geq \pi - \psi_B^1 - \psi_B^2 + \frac{\pi}{2} + \frac{\pi}{2} - \psi_D^2 - \psi_A^1 = \left(\pi - \psi_A^1 - \psi_B^1\right) + \left(\pi - \psi_B^2 - \psi_D^2\right) > 0$$

with the last step guaranteed by the cost triangles. Therefore, b_C cannot be negative either, which concludes the second case and thus the proof. □

4.6 Conclusion

In this chapter, we have studied the branched and multimaterial transport problems, with a particular focus on some special cases of the latter. After recalling the branched transport problem, we focused on its convex relaxation in terms of multimaterial transport. We have collected and adapted the known formulations and results related to the primal-dual optimality, and then studied the multimaterial problem in a special setting when only a single topology of the solution is admissible. We discovered and showed the relation of the single topology problem to the free vertex optimization problem and observed that the dual solutions are globally linear in this setting. We also studied an explicit example of a problem of the single topology type with 2 sources, 1 sink, and 1 free vertex and completely characterized its solution set based on the initial data (i.e. the cost coefficients of the material vectors and the locations of the fixed vertices). We then discussed numerical methods for the multimaterial problem, presenting two different discretization approaches, to show that even with advanced numerical schemes, this problem is difficult to solve, and concluded that it is necessary to better understand the behavior of the solutions before attempting to develop more efficient numerical methods. We observed also that the dual solutions are seemingly piecewise affine in the general case (i.e. when not only a single topology of solutions is admissible).

Based on these findings, we decided to study the special problem on 4 vertices (i.e. a problem with 3 sources, 1 sink and 2 free vertices) when 2 networks with different topologies give the same transportation cost. Investigating the problem from geometric and algebraic perspectives, we discovered and proved that the non-degenerate graphs have the same cost if and only if the bisectors (described in Section 4.5.3) meet in one point. We then attempted an investigation of the solution set (for non-degenerate solutions) and discovered a necessary condition for the existence of the 2-graph same cost solution based on just the initial data, i.e. the cost coefficients of the material vectors, which we then used to study and identify the types of solutions that are not possible in this problem. It should be noted that the conditions that have been found are only necessary, but not sufficient, and finding sufficient conditions for the existence of non-degenerate solutions of the described type could be an interesting future research question. Another open question would be the explicit construction of dual candidates: As we discovered that the bisectors meet for same cost graphs, the mesh induced by the four bisector lines would be a natural basis for piecewise affine dual solutions. Unfortunately, preliminary numerical experiments indicated, that it is not always possible to build feasible dual solution on this simple mesh. Exploring this issue in more detail would be another potential direction of the future work.

5 Conclusion

Summary. In this thesis, we studied the optimal and branched transport problems in some specific settings.

In Chapter 3, we studied the properties of the barycenters in the unbalanced optimal transport setting, in particular in the Hellinger–Kantorovich distance. We first investigated the problem between an uncountable number of general input measures, derived a corresponding dual problem, and showed the existence and stability of the solutions with respect to input data and the length scale parameter. We then focused on a special case, when all the input measures are single Dirac masses. In this setting, we showed the existence of continuous dual maximizers, their uniqueness (almost everywhere with respect to the distribution of the input measures) and primal-dual optimality conditions. We also studied the behavior of the solutions when the scaling parameter $\kappa \to 0$, including the limit solution and asymptotic mass and density estimates. We discovered and showed that in some cases no discrete minimizer can exist. We then performed and presented some numerical experiments using the developed numerical scheme, to demonstrate the intricate behavior of the barycenters with the change of the scaling parameter and to support the theoretical findings presented before, such as the role of the dual solution and the robustness of the solution under empirical approximation.

Chapter 4 was focused on the branched transport problem and the multimaterial transport problem. After briefly recalling the branched transport problem and the challenges of solving it because of the concave cost function, we concentrated on the multimaterial problem, which has been shown under some conditions to be a convex relaxation of the branched transport problem. We studied the problem in the setting when only a single topology of the solution is admissible and showed its relation to the free vertex optimization problem. As an explicit example, we studied the problem with 2 vertices, 1 sink and 1 free vertex, for which, using the vertex optimization problem, we completely characterized the solution set based on the initial data. We then studied the multimaterial problem from the numerical perspective. We presented two numerical schemes based on different discretization approaches and confirmed that the problem presents difficulties even in some seemingly simple settings. We also observed the simple structure of the dual solution, which we already described theoretically in the single topology setting. Based on these findings, we proceeded to study the special problem on 4 vertices that allows for 2 solutions of different topologies which have the same transportation cost. We showed the equivalence between the same cost condition and the fact that the bisectors of the 2-graph solution meet in one point and then investigated the solution set of the problem. We discovered and proved a necessary condition for the existence of a non-degenerate solution of the prescribed type and used it to further characterize the possible solution set.

Future work. In the course of this study, we came across some interesting related questions that we could not fully investigate. Based on our findings and the context of research, we present some of them here.

- Given the discovered intricate structure presented by the solutions to the Hellinger–Kantorovich barycenter problem and the preliminary comparison with hierarchical clustering methods, it would be interesting to use the barycenter problem for structure analysis of real data point clouds, as it could potentially help to further interpret the behavior of the barycenters.
- While investigating the special 4-vertex problem from Section 4.5, we discovered the special role of the bisectors of the 2-graph solution. It would be interesting to investigate meshes based on the partition of the domain by such bisectors and consider not only optimality but also feasibility of the solutions obtained on such meshes.
- The described 4-vertex problem is a very special case of the general multimaterial transport problem, and even of the multimaterial problem on 4 vertices. Our study revealed the special geometric structure imposed by this problem, and therefore, it would be interesting to investigate the possibility of applying the findings of this research to the solution of the general multimaterial transport problem, starting of course from the solution of the general 4-vertex problem in two-dimensional space.

Bibliography

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] R. Ahuja, T. Magnanti, and J. Orlin. *Network flows*. Prentice Hall, 1988.

[3] L. Ambrosio. Lecture notes on optimal transport problems. (Funchal, 2000). In *Mathematical Aspects of Evolving Interfaces*, volume 1812), pages 1–52. Springer, Berlin, 2003.

[4] L. Ambrosio, E. Brué, and D. Semola. *Lectures on optimal transport*. Springer, 2021.

[5] M. Audin. *Remembering Sofya Kovalevskaya*. Springer Science & Business Media, 2011.

[6] E. Aurell, C. Mejía-Monasterio, and P. Muratore-Ginanneschi. Optimal protocols and optimal transport in stochastic thermodynamics. *Physical review letters*, 106(25):250601, 2011.

[7] F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20:61–76, 1998.

[8] Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

[9] H. Bauschke and P. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer New York, 2011.

[10] J.-D. Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5):851–868, 2003.

[11] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[12] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[13] J.-D. Benamou, G. Carlier, and R. Hatchi. A numerical solution to Monge's problem with a Finsler distance as cost. *ESAIM: Mathematical Modelling and Numerical Analysis*, 52(6):2133–2148, 2018.

[14] J.-D. Benamou, B. Froese, and A. Oberman. Numerical solution of the optimal transportation problem using the Monge–Ampère equation. *Journal of Computational Physics*, 260:107–126, 2014.

[15] M. Bernot, V. Caselles, and J.-M. Morel. *Optimal transportation networks: models and theory*. Springer, 2008.

[16] D. Bertsekas. A distributed algorithm for the assignment problem. *Lab. for Information and Decision Systems Working Paper, MIT*, 1979.

[17] D. Bertsekas. The auction algorithm: A distributed relaxation method for the assignment problem. *Annals of operations research*, 14(1):105–123, 1988.

[18] D. Bertsekas. *Network optimization: continuous and discrete models*, volume 8. Athena Scientific, 1998.

[19] D. Bertsekas and D. Castanon. The auction algorithm for the transportation problem. *Annals of Operations Research*, 20(1):67–96, 1989.

[20] S. Bhaskaran and J. Franz. Optimal design of gas pipeline networks. *Journal of the Operational Research Society*, 30:1047–1060, 1979.

[21] J. Bigot and T. Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

[22] M. Bonafini. Convex relaxation and variational approximation of the Steiner problem: theory and numerics. *Geometric Flows*, 3(1):19–27, 2018.

[23] M. Bonafini, O. Minevich, and B. Schmitzer. Hellinger–Kantorovich barycenter between Dirac measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 29:19, 2023.

[24] M. Bonafini, O. Minevich, B. Schmitzer, and B. Wirth. Some explicit solutions for the multimaterial transport problem. In preparation, 2023.

[25] M. Bonafini and E. Oudet. A convex approach to the Gilbert–Steiner problem. *Interfaces and Free Boundaries*, 22(2):131–155, 2020.

[26] N. Bonneel and J. Digne. A survey of optimal transport for computer graphics and computer vision. *Computer Graphics forum*, 42(2):439–460, 2023.

[27] M. Bonnivard, E. Bretin, and A. Lemenant. Numerical approximation of the Steiner problem in dimension 2 and 3. *Mathematics of Computation*, 89(321):1–43, 2020.

[28] J. Borwein and A. Lewis. *Convex Analysis*. Springer, 2006.

[29] R. Boţ, E. Csetnek, and C. Hendrich. Recent developments on primal–dual splitting methods with applications to convex minimization. *Mathematics Without Boundaries: Surveys in Interdisciplinary Research*, pages 57–99, 2014.

[30] D. Bourne and S. Roper. Centroidal power diagrams, Lloyd's algorithm, and applications to optimal location problems. *SIAM Journal on Numerical Analysis*, 53(6):2545–2569, 2015.

[31] A. Brancolini and G. Buttazzo. Optimal networks for mass transportation problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 11(1):88–101, 2005.

[32] 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*, 57:1–39, 2018.

[33] L. Brasco. A survey on dynamical transport distances. *Journal of Mathematical Sciences*, 181(6):755–781, 2012.

[34] M. Brazil, R. Graham, D. Thomas, and M. Zachariasen. On the history of the Euclidean Steiner tree problem. *Archive for history of exact sciences*, 68:327–354, 2014.

[35] L. Caffarelli, M. Feldman, and R. McCann. Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. *Journal of the American Mathematical Society*, 15(1):1–26, 2002.

[36] T. Cai, J. Cheng, B. Schmitzer, and M. Thorpe. The linearized Hellinger–Kantorovich distance. *SIAM Journal on Imaging Sciences*, 15(1):45–83, 2022.

[37] M. Carioni, A. Marchese, A. Massaccesi, A. Pluda, and R. Tione. The oriented mailing problem and its convex relaxation. *Nonlinear Analysis*, 199:112035, 2020.

[38] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 2011.

[39] K. Chaudhuri, S. Dasgupta, S. Kpotufe, and U. von Luxburg. Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory*, 60(12):7900–7912, 2014.

[40] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18:1–44, 2018.

[41] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[42] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[43] N.-P. Chung and M.-N. Phung. Barycenters in the Hellinger–Kantorovich space. *Applied Mathematics and Optimization*, 84(2):1791–1820, 2021.

[44] C. Clason, C. Tameling, and B. Wirth. Convex relaxation of discrete vector-valued optimization problems. *SIAM Review*, 63(4):783–821, 2021.

[45] P. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer New York, 2011.

[46] R. Courant and H. Robbins. *What is Mathematics?* Oxford University Press, 1941.

[47] N. Courty, R. Flamary, A. Rakotomamonjy, and D. Tuia. Optimal transport for domain adaptation. In *NIPS 2014, Workshop on Optimal Transport and Machine Learning*, 2014.

[48] W. H. Cunningham. Theoretical properties of the network simplex method. *Mathematics of Operations research*, 4(2):196–208, 1979.

[49] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[50] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International conference on Machine Learning*, pages 685–693. PMLR, 2014.

[51] G. Dantzig. Programming of interdependent activities: II mathematical model. *Econometrica, Journal of the Econometric Society*, pages 200–211, 1949.

[52] G. Dantzig. Application of the simplex method to a transportation problem. In T. Koopmans, editor, *Activity analysus of production and allocation*, chapter 23, pages 359–373. John Wiley and Sons, New York, 1951.

[53] G. Dantzig. Origins of the simplex method. In S. Nash, editor, *A history of scientific computing*, pages 141–151. ACM, 1990.

[54] G. Dantzig, L. Ford, and D. Fulkerson. A primal-dual algorithm for linear programs. In H. Kuhn and A. Tucker, editors, *Linear Inequalities and Related Systems*, pages 171–182. Princeton University Press, 1956.

[55] G. Dantzig, A. Orden, and P. Wolfe. The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics*, 5(2):183–195, 1955.

[56] I. Dikin. Iterativnoe reshenie zadach lineynogo i kvadratichnogo programmirovaniya [Russian; Iterative solution of problems of linear and quadratic programming]. *Doklady Akademii Nauk SSSR [Russian; Proceedings of the USSR Academy of Science]*, 174(4):747–748, 1967.

[57] R. Duan and S. Pettie. Linear-time approximation for maximum weight matching. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

[58] M. Dyer, B. Gärtner, N. Megiddo, and E. Welzl. Linear programming. In *Handbook of discrete and computational geometry*, pages 1291–1309. Chapman and Hall/CRC, 2017.

[59] L. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, second edition, 2010.

[60] L. Evans and W. Gangbo. *Differential equations methods for the Monge-Kantorovich mass transfer problem*. American Mathematical Society, 1999.

[61] J. Feydy. *Geometric data analysis, beyond convolutions*. Doctoral thesis, École Normale Supérieure de Cachan, 2020.

[62] J. Feydy, B. Charlier, F.-X. Vialard, and G. Peyré. Optimal transport for diffeomorphic registration. In *MICCAI 2017: Medical Image Computing and Computer Assisted Intervention*, pages 291–299. Springer, 2017.

[63] A. Figalli and N. Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions. *Journal de mathématiques pures et appliquées*, 94(2):107–130, 2010.

[64] J. Fish and T. Belytschko. *A First Course in Finite Elements*. John Wiley and Sons, 2007.

[65] G. Friesecke, D. Matthes, and B. Schmitzer. Barycenters for the Hellinger–Kantorovich distance over $\mathbb{R}^d$. *SIAM Journal on Mathematical Analysis*, 53(1):62–110, 2021.

[66] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[67] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

[68] E. Gilbert. Minimum cost communication networks. *Bell System Technical Journal*, 46(9):2209–2227, 1967.

[69] D. Goldfarb. Efficient dual simplex algorithms for the assignment problem. *Mathematical Programming*, 33(2):187–203, 1985.

[70] A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *IPMI 2015: Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

[71] W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM Journal on optimization*, 16(1):170–192, 2005.

[72] F. Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of mathematics and physics*, 20(1-4):224–230, 1941.

[73] G.-J. Huizing, G. Peyré, and L. Cantini. Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics*, 38(8):2169–2177, 2022.

[74] D. Iudin. Lightning as an asymmetric branching network. *Atmospheric Research*, 256:105560, 2021.

[75] L. Kantorovich. O peremestchenii mass [Russian; On the displacement of masses]. *Doklady Akademii Nauk SSSR [Russian; Proceedings of the USSR Academy of Science]*, 37:227–230, 1942.

[76] N. Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, 1984.

[77] L. Khachiyan. Polynomial'nyi algoritm v lineynom programmirovanii [Russian; A polynomial algorithm in linear programming]. *Doklady Akademii Nauk SSSR [Russian; Proceedings of the USSR Academy of Science]*, 244(5):1093–1096, 1979.

[78] M. Klatt, C. Tameling, and A. Munk. Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2):419–443, 2020.

[79] T. Koopmans. Optimum utilization of the transportation system. *Econometrica: Journal of the Econometric Society*, pages 136–146, 1949.

[80] J. Kosowsky and A. Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural networks*, 7(3):477–490, 1994.

[81] H. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[82] V. Laschos and A. Mielke. Geometric properties of cones with applications on the Hellinger-Kantorovich space, and a new distance on the space of probability measures. *Journal of Functional Analysis*, 276(11):3529–3576, 2019.

[83] C. Lemke. The dual method of solving the linear programming problem. *Naval Research Logistics Quarterly*, 1(1):36–47, 1954.

[84] C. Léonard. From the Schrödinger problem to the Monge–Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[85] B. Lévy and E. Schwindt. Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, 72:135–148, 2018.

[86] M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[87] J. Lohmann. *On the branched transport problem: Reformulation as geometry optimization and calibration using convex duality.* Doctoral thesis, Westfälische Wilhelms-Universität Münster, 2023.

[88] J. Lohmann, B. Schmitzer, and B. Wirth. Duality in branched transport and urban planning. *Applied Mathematics & Optimization*, 86(3):45, 2022.

[89] J. Lohmann, B. Schmitzer, and B. Wirth. Formulation of branched transport as geometry optimization. *Journal de Mathématiques Pures et Appliquées*, 163:739–779, 2022.

[90] F. Maddalena, S. Solimini, and J.-M. Morel. A variational model of irrigation patterns. *Interfaces and Free Boundaries*, 5(4):391–415, 2003.

[91] A. Marchese, A. Massaccesi, S. Stuvard, and R. Tione. A multi-material transport problem with arbitrary marginals. *Calculus of Variations and Partial Differential Equations*, 60(3):1–49, 2021.

[92] A. Marchese, A. Massaccesi, and R. Tione. A multimaterial transport problem and its convex relaxation via rectifiable G-currents. *SIAM Journal on Mathematical Analysis*, 51(3):1965–1998, 2019.

[93] A. Massaccesi, E. Oudet, and B. Velichkov. Numerical calibration of Steiner trees. *Applied Mathematics & Optimization*, 79:69–86, 2019.

[94] W. Matusik, M. Zwicker, and F. Durand. Texture design using a simplicial complex of morphable textures. *ACM Transactions on Graphics (TOG)*, 24(3):787–794, 2005.

[95] K. McCulloh, J. Sperry, and F. Adler. Water transport in plants obeys Murray's law. *Nature*, 421(6926):939–942, 2003.

[96] Q. Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30 (5), pages 1583–1592. Wiley Online Library, 2011.

[97] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Mémoires de Mathématique et de Physique, Présentés à l'Académie Royale des Sciences, par divers Savans, et lûs dans ses Assemblées*, pages 666–704, 1781.

[98] A. Natale and G. Todeschi. Computation of optimal transport with finite volumes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(5):1847–1871, 2021.

[99] NobelPrize.org. Nobel Prize Outreach: Press release, 14 October 1975. Available at: https://www.nobelprize.org/prizes/economic-sciences/1975/press-release/.

[100] V. Oliker and L. Prussner. On the numerical solution of the equation $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 = f$ and its discretizations, I. *Numerische Mathematik*, 54(3):271–293, 1989.

[101] C. Olson. Parallel algorithms for hierarchical clustering. *Parallel computing*, 21(8):1313–1325, 1995.

[102] N. Papadakis, G. Peyré, and E. Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[103] B. Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264(4):947–963, 2013.

[104] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

[105] P. Prusinkiewicz. Modeling plant growth and development. *Current opinion in plant biology*, 7(1):79–83, 2004.

[106] N. Retiere, G. Muratore, G. Kariniotakis, A. Michiorri, P. Frankhauser, J.-G. Caputo, Y. Sidqi, R. Girard, and A. Poirson. Fractal grid – towards the future smart grid. In *CIRED 2017-24th International Conference on Electricity Distribution*, page 1236, 2017.

[107] R. Rockafellar. *Convex analysis*, volume 18. Princeton University Press, 1970.

[108] P. Roitman and H. Le Ferrand. The strange case of Paul Appell's last memoir on Monge's problem:"sur les déblais et remblais". *Historia Mathematica*, 43(3):288–309, 2016.

[109] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[110] B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

[111] B. Schmitzer and B. Wirth. Dynamic models of Wasserstein-1-type unbalanced transport. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:23, 2019.

[112] R. Seidel. Linear programming and convex hulls made easy. In *Proceedings of the sixth annual symposium on Computational geometry*, pages 211–215, 1990.

[113] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

[114] Z. Su, Y. Wang, R. Shi, W. Zeng, J. Sun, F. Luo, and X. Gu. Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2246–2259, 2015.

[115] V. Sudakov. Geometricheskiye problemy teorii beskonechnomernyh veroyatnostnyh raspredeleniy [Russian; Geometric problems of the theory of infinite-dimensional probability distributions]. *Proceedings of Steklov mathematical institute*, 141(0):3–191, 1976.

[116] A. Tolstoy. Metody nakhozhdeniya naimen'shego summovogo kilometrazha pri planirovanii perevozok v prostranstve [Russian; Methods of finding the minimal total kilometrage in cargotransportation planning in space]. *Planirovanie Perevozok, Sbornik pervyı [Russian; Transportation Planning, Volume I], Transpechat'NKPS [TransPress of the National Commissariat of Transportation], Moscow*, pages 23–55, 1930.

[117] N. Trudinger and X.-J. Wang. On the Monge mass transfer problem. *Calculus of Variations and Partial Differential Equations*, 13:19–31, 2001.

[118] A. Uwitonze, J. Huang, Y. Ye, W. Cheng, and Z. Li. Exact and heuristic algorithms for space information flow. *Plos one*, 13(3):e0193350, 2018.

[119] A. Vershik. Some remarks on the infinite-dimensional problems of linear programming. *Uspekhi Matematicheskikh Nauk*, 25(5):117–124, 1970.

[120] A. Vershik. Long history of the Monge–Kantorovich transportation problem. *Math Intelligencer*, 35:1–9, 2013.

[121] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

[122] Q. Xia. Optimal paths related to transport problems. *Communications in Contemporary Mathematics*, 5(02):251–279, 2003.

[123] Q. Xia, C. Salafia, and S. Morgan. Optimal transport and placental function. In *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science*, pages 509–515. Springer, 2015.