A Quadratic Regularization of Optimal Transport Problems and Its Application to Bilevel Optimization

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

der Fakultät für Mathematik
der Technischen Universität Dortmund
vorgelegt von

Sebastian Hillbrecht

am 28. November 2023.

Dissertation

A Quadratic Regularization of Optimal Transport Problems and Its Application to Bilevel Optimization

Fakultät für Mathematik
Technische Universität Dortmund

Erstgutachter: Prof. Dr. Christian Meyer
Zweitgutachter: Prof. Dr. Dirk Lorenz

Tag der mündlichen Prüfung: 14. März 2024

Abstract

This thesis consists of two parts, in each of which a quadratic regularization is applied to an optimal transport problem and its effect on a prototypical bilevel optimization problem is investigated.

In the first part, we use the mentioned quadratic regularization in combination with a smoothing of the marginals to improve certain properties of the well-known Kantorovich problem, which is a linear optimization problem defined on infinite-dimensional spaces. In this way we obtain, for example, the uniqueness of the optimal solution and an associated optimality system containing (non-unique) dual variables. We then use these improved properties of the problem to regularize a bilevel optimization problem whose constraints require to solve the Kantorovich problem. We then show that the regularized bilevel problem has a solution and that we can, under certain conditions, approximate solutions of the non-regularized bilevel problem by solutions of the regularized one. We conclude the first part with a brief overview of possible applications of this regularization approach.

In the second part, we apply the same regularization approach to the also well-known Hitchcock problem, which we introduce as a finite-dimensional special case of the Kantorovich problem. Due to the structure of this problem, however, we can dispense with the additional smoothing of the boundary conditions. Similar to the first part, we regularize a bilevel problem whose constraints require the solution of the Hitchcock problem. We again show the existence of solutions to the regularized bilevel problem and that we can use this to approximate solutions to the non-regularized bilevel problem, in certain cases. By introducing a further regularization of the Lagrangian dual problem, we enforce the uniqueness of the dual variables from the optimality system. This enables us to calculate derivatives of the marginal-to-transport-plan mapping and, in turn, to establish an implicit programming approach for the solution of the regularized bilevel problem. To conclude the second part, we test our findings numerically by means of an transportation identification problem.

To my son, Paul Henri.

I dedicate this thesis to you in the hope that it demonstrates the importance of pursuing one's goals, even in challenging moments. May it serve as a reminder that, with perseverance, anything is possible.

With all my love
Papa

Acknowledgment

I would like to dedicate this page to the people who have supported me in completing this thesis.

First and foremost, I owe a great debt to my supervisor, Christian. His expertise, tireless commitment, and infinite patience formed the bedrock of this thesis and its associated research. I am grateful for the countless discussions, his invaluable feedback, and the constant encouragement that have carefully orchestrated my journey toward the successful completion of this thesis.

Beyond that, I am also deeply grateful to Paul, who has dedicated numerous hours to technical and personal discussions with me. His boundless willingness to share his knowledge and his engagement have significantly propelled the progress of this thesis more than once.

To my wife, Anna, I would like to say first of all that I am incredibly lucky that she has been my constant companion from the beginning of my doctorate to this very moment. In difficult moments, I found comfort in knowing that I could always rely on her. In moments of success, sharing this success has made it even sweeter.

Last but not least, I would like to thank my family, especially my parents Renate and Volker, as well as all my friends and colleagues who have accompanied and encouraged me along the way and who have provided a little distraction at the right times. This work would not have been possible without your contributions and your belief in me.

Thankfully
Sebastian

Contents

1 Introduction **1**

2 Notation **7**

I The Infinite-Dimensional Case **11**

3 The Bilevel Kantorovich Problem **13**
3.1 Problem Statement 13
3.2 Quadratic Regularization 17
3.3 Existence of Regularized Bilevel Solutions 22
3.4 Approximation of Non-Regularized Solutions 37
3.4.1 Relaxation of Assumption 3.29 47
3.5 Existence of a Recovery Sequence 51

II The Finite-Dimensional Case **57**

4 The Bilevel Hitchcock Problem **59**
4.1 Problem Statement 59
4.2 Quadratic Regularization 63
4.3 Existence of Regularized Bilevel Solutions 68
4.4 Approximation of Non-Regularized Solutions 69
4.5 Existence of a Recovery Sequence 71

5 Towards Implicit Programming **85**
5.1 Regularization of the Dual Problem 86
5.2 The Regularized Marginal-to-Transport-Plan Mapping 99
5.3 Application of the Implicit Programming Approach 104

6 Implementation of the Implicit Programming Approach **109**
6.1 A Constrained Nonsmooth Trust Region Method 109
6.2 Construction of a Model Function 119
6.3 A Transportation Identification Problem 122
6.3.1 Details on the Implementation 124
6.3.2 Presentation & Discussion of the Numerical Results . . . 127

Appendix xi

A On the Convolution of Marginals With Mollifiers xiii

B On the Theory of Measure & Integration xix

C On the Theory of Optimal Transport xxix

D On Functional Analysis xxxi

Bibliography xxxv

Index xli

Chapter 1

Introduction

Optimal transport, also known as *transportation theory*, is a mathematical theory that models the transportation of masses (e.g. goods and resources, but also any abstract objects) from one place to another and seeks to organize this transportation in such a way that it minimizes the resulting transport costs. On account of this general formulation of the problem, the theory finds widespread application in various different fields, including economics (e.g. [47, 15, 39]), computer graphics (e.g. [66, 78, 11]), statistics and in particular machine learning (e.g. [41, 65, 26]), or even fluid dynamics (e.g. [7, 40]).

For a detailed overview of possible applications and an in-depth discussions of optimal transportation, we refer the interested reader to the books by Villani [75, 76] and Santambrogio [68] as well as to the extensive review articles [3, 61]. In all of these, the authors show that optimal transportation can also be linked to other mathematical disciplines such as (differential) geometry, partial differential equations, and several others.

It is also due to the general formulation of the optimal transportation problem that there are several (seemingly) independent formulations of the transportation problem. The oldest of them can be traced back to the French mathematician Monge, see [57], who in the late 18th century tried to find a transport map, which is an injective mapping from the source domain to the target domain and determines from where to where mass is transported. Another popular (and more general) formulation is that of the Soviet mathematician and economist Kantorovich from the early 1940s, see [49], who sought to find a transport plan which is a joint distribution between (mass) distributions on the source domain and the target domain and, unlike Monge's transport map, allows for splitting and merging of masses during transportation. Ten years later, the German economist Beckmann presented a formulation that is based on the minimization of gradient flows between sources and sinks, see [6]. It is worth noting that under certain circumstances the formulations of Monge and Kantorovich coincide and, moreover, that the solution of the Beckmann problem can be related to the other problems via the well-known Monge-Kantorovich equations, see e.g. [68].

In Part I of this thesis, we focus on Kantorovich's formulation of optimal transportation. As already indicated, the Kantorovich optimal transportation problem tries to find a joint probability distribution for given mass distributions (represented by regular Borel probability measures) on both a source and a

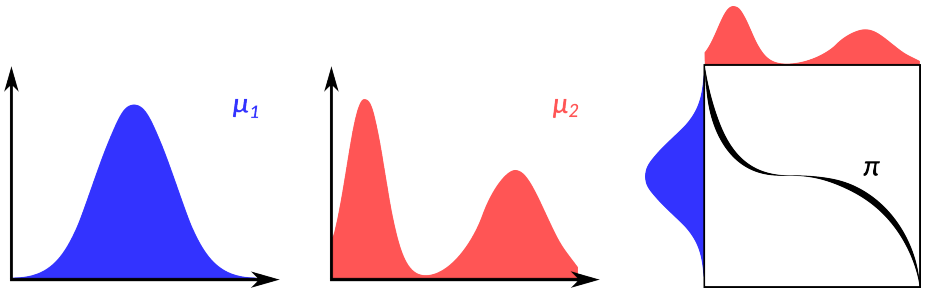


Figure 1.1: Kantorovich optimal transportation. The distribution μ_1 from the left picture is transported to the distribution μ_2 from the middle picture by means of the transport plan π from the right picture. If the cost of transportation between the source domain and the target domain is strictly convex and if the mass distributions have sufficient regularity, then the optimal transport plan is concentrated on the graph of a strictly increasing function.

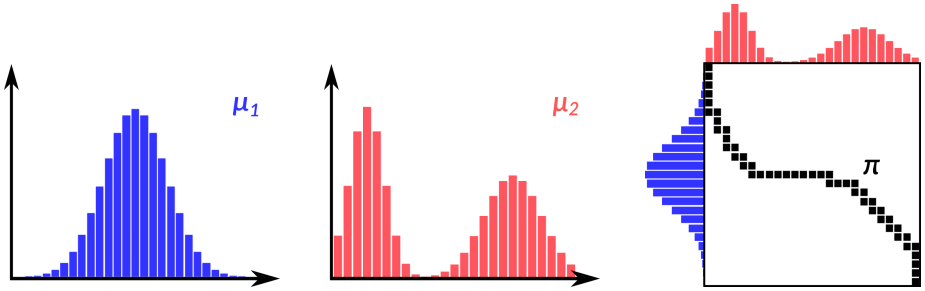


Figure 1.2: Hitchcock optimal transportation. The vector μ_1 from the left picture is transported to the vector μ_2 from the middle picture by means of the transportation matrix π whose sparsity pattern is shown in the right picture. Again, strictly convex transportation costs result in the optimal transport matrix being concentrated on the graph of a strictly increasing function.

target domain that has the given distributions as its first and second marginals. The fairly intuitive concept behind this rather unwieldy description is illustrated in Figure 1.1.

In Part II of this thesis, we consider the finite dimensional equivalent of the Kantorovich problem, the Hitchcock problem of optimal transport. It can be seen as a discretization of the Kantorovich problem in which the mass distributions are replaced by vectors and the joint distribution by a matrix. Figure 1.2 shows the undeniable similarities between the two formulations.

We will investigate both of this problems, the Kantorovich problem and the Hitchcock problem, in a bilevel context, i.e., we are considering the *prototypical bilevel problem*

$$\begin{aligned} \inf_{\pi,\mu_1} \quad & \mathcal{J}(\pi,\mu_1) \\ \text{s.t.} \quad & \mu_1 \text{ is a mass distribution on a source domain,} \\ & \pi \text{ is an optimal transport plan between } \mu_1 \text{ and } \mu_2^{\mathrm{d}} \text{ w.r.t. } c_{\mathrm{d}}, \end{aligned}$$

where μ_2^{d} is a (fixed) mass distribution on a target domain, c_{d} is a fixed description of the transportation cost between the source domain and the target domain, and $\mathcal{J}$ is a suitably chosen target function. Note that this is in fact a bilevel problem, since its feasible set depends on the optimization variable μ_1

and one needs to solve an optimal transport problem (which we sometimes call *subordinate* or *lower-level* problem) in order to obtain a feasible transport plan π.

We call this bilevel optimization problem "prototypical" for the following reason: depending on the formulation chosen for the subordinate optimal transport problem and depending on the choice of the objective function $\mathcal{J}$ and the assumptions made about the given data, the above bilevel problem can be used to solve a variety of problems. For example, if we consider the Kantorovich formulation as the subordinate problem and if we make certain assumptions on the domains and the data, we can show that a special case of the prototypical bilevel problem is given by the *Wasserstein inverse problem*

$$\begin{aligned} \inf_{\mu} \quad & \tfrac{1}{2}\|G\mu - y_{\mathrm{d}}\|_Y^2 + \nu W_\rho(\mu, \mu_{\mathrm{d}})^\rho \\ \text{s.t.} \quad & \mu \in \mathfrak{P}(\Omega_*), \end{aligned}$$

see Subchapter 3.5. In the above, G can be an arbitrary compact operator which maps the space of probability measures $\mathfrak{P}(\Omega_*)$ onto some Banach space Y and W_ρ denotes the Wasserstein ρ-distance on $\mathfrak{P}(\Omega_*)$. A popular choice for this operator is, for example, the solution operator of an (elliptic) differential equation. With this choice, the Wasserstein space problem turns into an optimal control problem on measure spaces. In [17, 18, 19, 23, 62], the authors consider the same kind of optimal control problem, but measure the distance between the control μ and the data μ_{d} by means of the total variation norm instead of the Wasserstein ρ-distance.

Another example would be to consider the *tracking-type* target functional

$$\mathcal{J} = \|\pi - \pi_{\mathrm{d}}\| + \left\|\mu_1 - \mu_1^{\mathrm{d}}\right\|.$$

With this objective functional, the prototypical bilevel problem becomes the problem of reconstructing the source distribution and transport plan based on (possibly incomplete and noisy) observations π_{d} and μ_1^{d}, which is an inverse problem on measure spaces. Problems of this form belong to a field of research that allows for a wide range of different approaches, both with and without connections to optimal transport. For the latter case, we only mention [13, 30] and the references therein. For the former case, we refer to [33, 32, 56], where optimal transport (directly or indirectly) enters the formulation of the inverse problem in form of a metric to measure the misfit of data, and [73], where the authors assume that the forward operator is given as the solution operator of the optimal transport problem and apply a Bayesian approach in order to reconstruct the cost of transportation through (noisy) observations of the transport plan.

Moreover, we want to mention the work of Mahler [55], which is closely related to the topic of this thesis, where the author considers a similar bilevel problem with Beckmann's optimal transport problem taking the role of the subordinate problem.

By their very nature, the Kantorovich problem and the Hitchcock problem are linear problems. On the one hand, this has the advantage that (after discretization) efficient linear solvers can be applied to calculate their solutions. On the other hand, this has the disadvantage that their solutions, depending on the cost function, are generally nonunique and, since the transport plans live on

the Cartesian product of the source and target distribution's domains, subject to a high dimensionality.

For this and other reasons, many authors prefer to apply a regularization to these optimal transport problems. Probably the best known and most commonly used approach is the so-called *entropic regularization*. It is broadly applied in different fields like imaging and machine learning but is also of theoretical interest, see e.g. [24, 72, 38, 16]. This can at least to some extent be attributed to the seminal work of Cuturi, who showed in [28] that entropic regularization allows for the use of Sinkhorn's algorithm [71] to (efficiently) solve the Hitchcock problem and even the challenging Wasserstein barycenter problem, see [29].

As the title of this thesis indicates, we will, however, follow a different regularization approach and regularize both the Kantorovich problem and the Hitchcock problem by means of a quadratic regularization. In [52], the authors propose an L^2-regularization of the Kantorovich problem, which serves not only to improve the regularity of its solutions (compared to the Kantorovich problem, the solutions of the regularized version are L^2-functions instead of measures) but also to guarantee their uniqueness and provide an optimality system including (Lagrangian) dual variables. Similar to entropic regularization, solving for the dual variables significantly reduces the dimension of the problem. The L^2-regularization (often referred to as *Tikhonov regularization*) has a rich tradition of being successfully used throughout different applications, see e.g. [79, 77, 8, 63] and the references therein. In [51], the authors show that the quadratic L^2-regularization, as a special case of a more general regularization approach, Γ-converges to the non-regularized problem, as the regularization parameter tends towards 0.

In direct comparison to the entropic regularization, the L^2-regularization preserves the sparsity of the transport plans (which is a unique feature of the solutions of the Kantorovich problems) much better: solutions of the entropically regularized Kantorovich problem are strictly positive on their domains, whereas solutions of the L^2-regularized Kantorovich problem have a representation including the $(\cdot)_+$-operator which promotes sparsity of the transport plan. However, the sparsity of the regularized transport plan comes with a price: the optimality system of the L^2-regularized Kantorovich problem includes the $(\cdot)_+$-operator and is therefore, in contrast to the optimality system of then entropically regularized Kantorovich problem, nonsmooth and nonlinear, ruling out the application of the Sinkhorn algorithm.

However, we may still apply standard nonsmooth optimization methods to compute solutions of the L^2-regularized Kantorovich problems, see e.g. [52, Section 4]. Applying a further regularization of the corresponding dual problem, we expect nonsmooth optimization methods in the spirit of [46, 43, 21] to be applicable to the twice regularized bilevel problem.

The rest of this thesis is organized as follows:

In Chapter 2, we introduce the most important notation for our purposes and state a number of basic properties of the spaces that are used in this thesis.

Chapter 3, which has in parts already been published in [45, 44], is the only chapter of Part I. Therein, we first carefully define the Kantorovich problem

and the prototypical bilevel problem and prove existence of solutions. We then introduce the quadratic regularization of both the Kantorovich problem and the prototypical bilevel problem, again prove the existence of solutions to the regularized problems, and subsequently address the approximability of solutions of the non-regularized bilevel problem. We conclude the chapter by giving two examples of possible applications.

Chapter 4, the first chapter of Part II, reproduces the results of the previous chapter for the case of the Hitchcock problem and its associated bilevel formulation. It does, however, provide added value in that we explicitly construct a nontrivial recovery sequence for a slightly more general case than was discussed at the end of Chapter 3.

In Chapter 5, we introduce an additional regularization to the dual problem of the regularized Hitchcock problem. This allows us to define a marginal-to-dual-variables mapping and to investigate its differentiability properties. Considering its concatenation with a mapping from the dual variables to a transport plan allows us to adopt an implicit programming approach in the context of the bilevel formulation of the Hitchcock problem.

Chapter 6 concludes the second part and also the main part of the thesis. We propose a trust region algorithm for the solution of nonsmooth optimization problems with convex constraints and implement the implicit programming approach we derived in the previous chapter. Finally, we test our implementation on an example that fits exactly into the setting of the second part and discuss the results.

The main part of the thesis is followed by a rather detailed appendix, which takes a closer look at individual aspects from the areas of convolutions of marginals with mollifiers (Appendix A), measure and integration theory (Appendix B), optimal transport (Appendix C), and functional analysis (Appendix D), which would have been distracting in the main part of this thesis.

Chapter 2

Notation

Finite Dimensional Spaces

On some finite-dimensional vector space X where each element $x \in X$ (think of matrices or vectors) takes the form $x = (x_i)_{i \in I}$ with I being some finite set, $\|x\|_1$ and $\|x\|_\infty$ denote the *1-norm* and the *∞-norm*, which sum the absolute values of all entries of x and return the largest absolute value of all entries of x, respectively, i.e.,

$$\|x\|_1 \coloneqq \sum_{i \in I} |x_i| \quad \text{and} \quad \|x\|_\infty \coloneqq \max_{i \in I} |x_i|.$$

Given $m, n \in \mathbb{N}$, we denote the *Euclidean norm* of a vector $v = (v_1, \dots, v_n)^\top \in \mathbb{R}^n$ by

$$\|v\|_{\mathbb{R}^n} \coloneqq \sqrt{|v_1|^2 + \cdots + |v_n|^2}$$

and the *spectral norm* of some matrix $M \in \mathbb{R}^{m \times n}$ by

$$\|M\|_{\mathbb{R}^{m \times n}} \coloneqq \max_{\|v\|_{\mathbb{R}^n} = 1} \frac{\|Mv\|_{\mathbb{R}^m}}{\|v\|_{\mathbb{R}^n}} = \sqrt{\sigma_{\max}},$$

where $\sigma_{\max}$ denotes the largest singular value of the matrix M. The *Frobenius norm* of the matrix M is defined by

$$\|M\|_F \coloneqq \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} M_{i,j}^2}$$

and it is induced by the *Frobenius scalar product*

$$(M, N)_F \coloneqq \sum_{i=1}^{m} \sum_{j=1}^{n} M_{i,j} N_{i,j}, \quad \text{for } M, N \in \mathbb{R}^{m \times n}.$$

Consequently, if we equip the space of real valued matrices with the Frobenius scalar product and its induced norm, this space becomes a Hilbert space.

Spaces of Continuous Functions

By $C(X)$, $C_b(X)$, and $C_c(X)$, we denote the function spaces of continuous,

continuous & bounded, continuous & compactly supported real valued functions $f\colon X \to \mathbb{R}$ on a locally compact Hausdorff space X, respectively. While $C(X)$ and $C_b(X)$ are Banach spaces w.r.t. the *uniform norm*

$$\|f\|_\infty \coloneqq \sup_{x\in X} |f(x)|,$$

the linear space $\big(C_c(X), \|\cdot\|_\infty\big)$ is in general not complete. We therefore consider its norm closure

$$C_0(X) \coloneqq \overline{C_c(X)}^{\|\cdot\|_\infty}$$

which is the Banach space of functions that are vanishing towards the boundary of X.

If X happens to be compact, then $C(X)$ is also a Banach space w.r.t. the uniform norm and coincides with all of the above Banach spaces of continuous functions. Occasionally and in particular if we need to distinguish between different domains, we denote the uniform norm by $\|f\|_{C(X)}$.

Borel Sets & Spaces of Measures

By $\mathfrak{B}(X)$, we denote the *Borel σ-algebra* on some arbitrary topological space (X,τ). It is the smallest σ-algebra that contains all open sets of X, i.e., all elements of τ. We call the elements of $\mathfrak{B}(X)$ *(Borel) measurable sets*.

For $d \in \mathbb{N}$, let $X \subset \mathbb{R}^d$ be a subset that we equip with the subspace topology of $\mathbb{R}^d$. We denote the Banach *space of regular Borel measures* on the measurable space $\big(X, \mathfrak{B}(X)\big)$ by $\mathfrak{M}(X)$. It consists of all signed Borel measures $\mu\colon \mathfrak{B}(X) \to \mathbb{R}$ whose variation measures $|\mu|$ are (inner and outer) regular. Its norm is the *total variation norm* $\|\mu\|_{\mathfrak{M}(X)} \coloneqq |\mu|(X)$. We write $\mu \geq 0$ short for "$\mu(B) \geq 0$ for all measurable sets B.

If X happens to be compact, then the Riesz-Radon theorem (see e.g. [2, Theorem 6.23]) ensures that $\mathfrak{M}(X) \cong C(X)^*$, i.e., the topological dual space of the Banach space of continuous functions can be identified with the Banach space of regular Borel measures. We refer the interested reader to Appendix B for further information on signed measures.

We denote the set of *regular Borel probability measures* on $\big(X, \mathfrak{B}(X)\big)$, by $\mathfrak{P}(X)$. This is the subset of regular Borel measures $\mu \in \mathfrak{M}(X)$ that satisfy $\mu \geq 0$ and $\|\mu\|_{\mathfrak{M}(X)} = 1$.

Lebesgue- & Sobolev Spaces

For $d \in \mathbb{N}$, let $X \subset \mathbb{R}^d$ be a *domain* in [1]'s sense, i.e., a non-empty open subset of the d-dimensional real Euclidean space. Moreover, let $\lambda\colon \mathfrak{B}(X) \to \mathbb{R}_+ \cup \{+\infty\}$ denote the well-known Lebesgue measure on X. We say that a measurable set B is a *Lebesgue null set*, if $\lambda(B) = 0$. We abbreviate the Lebesgue measure of some measurable subset $B \in \mathfrak{B}(X)$ by $|B| \coloneqq \lambda(B)$.

For $p \in [1,\infty)$, we denote by $L^p(X)$, which is short for $L^p(X, \mathfrak{B}(X), \lambda)$, the Banach space of equivalence classes of Lebesgue-Borel measurable and to the p-th power absolutely Lebesgue integrable functions $u\colon X \to \mathbb{R}$. Its norm is the *L^p norm*, which is defined by

$$\|[u]\|_{L^p(X)} \coloneqq \left(\int_X |u|^p \, d\lambda \right)^{\frac{1}{p}} \quad \text{for any } u \in [u].$$

For $p = \infty$, we denote by $L^\infty(X)$, which is short for $L^\infty(X, \mathfrak{B}(X), \lambda)$, the Banach space of equivalence classes of $\mathfrak{B}(X)$-$\mathfrak{B}(\mathbb{R})$-measurable functions $u\colon X \to \mathbb{R}$, whose absolute value is essentially bounded, i.e.,

$$\|[u]\|_{L^\infty(X)} \coloneqq \inf_{\substack{N \subset X \text{ is a} \\ \text{Lebesgue null set}}} \sup_{x \in X \setminus N} |u(x)| \; < \; \infty \quad \text{for any } u \in [u].$$

Two functions u_1 and u_2 belong to the same equivalence class $[u]$ (and are thus considered equal), if they differ only on a Lebesgue null set. We follow the usual convention of omitting the brackets of equivalence classes, i.e., we write $u \in L^p(X)$ instead of $[u] \in L^p(X)$.

For $p \in (1, \infty)$, the topological dual space of $L^p(X)$ is given by $L^{p'}(X)$ where $p' = {}^{p}\!/\!_{(p-1)} \in (1, \infty)$. If the domain X is bounded, then $L^p(X) \subset L^q(X)$ for all $q \in [1, p]$ and $L^\infty(X) \cong \big(L^1(X)\big)^*$.

By $W^{1,p}(X)$, where $1 \leq p \leq \infty$, we denote the *Sobolev space* of functions $u \in L^p(X)$ on X whose first-order weak partial derivatives are elements of $L^p(X)$ again. It is a Banach space w.r.t. the *Sobolev norm*

$$\|u\|_{W^{1,p}(X)} \coloneqq \begin{cases} \Big(\|u\|^p_{L^p(X)} + \sum_{i=1}^d \|D^{x_i} u\|^p_{L^p(X)}\Big)^{\frac{1}{p}}, & \text{if } 1 \leq p < \infty, \\ \max\Big\{\|u\|_{L^\infty(X)}, \max_{i \in \{1,\dots,d\}} \|D^{x_i} u\|_{L^\infty(X)}\Big\}, & \text{if } p = \infty. \end{cases}$$

If $p < \infty$, then $W^{1,p}(X)$ is separable. If additionally $p > 1$, then $W^{1,p}(X)$ is even uniform convex and reflexive, see e.g. [1, Theorem 3.6].

Moreover, if we close $C_0^\infty(X)$ w.r.t. the Sobolev norm, i.e., if we define

$$W_0^{1,p}(X) \coloneqq \overline{C_0^\infty(X)}^{\|\cdot\|_{W^{1,p}(X)}},$$

then $W_0^{1,p}(X)$ is a Banach space (w.r.t. the Sobolev norm). For $1 < p < \infty$, we denote its topological dual space by $W^{-1,p'}(X)$, see [1, Theorem 3.12 & Theorem 3.13].

In the case that X is closed, we write (slightly abusing the notation) $W^{1,p}(X)$ and $W^{-1,p'}(X)$ instead of $W^{1,p}(\operatorname{int} X)$ and $W^{-1,p'}(\operatorname{int} X)$, respectively, for the Sobolev spaces defined on its interior.

Miscellaneous

On some metric space (X, d), we denote the *open ball* with radius $r > 0$ around some point $x_0 \in X$ by

$$B_X(x_0; r) \coloneqq \{x \in X \colon d(x_0, x) < r\}.$$

Analogously, we denote the *closed ball* with radius $r > 0$ around $x_0 \in X$ by

$$\overline{B_X(x_0; r)} \coloneqq \{x \in X \colon d(x_0, x) \leq r\}.$$

To simplify the notation, we frequently refrain from subscripting the space in the notation of the ball.

On some Hilbert space H, we denote the *scalar product* (sometime called "inner product") between two elements $h_1, h_2 \in H$ by $(h_1, h_2)_H$. Conversely, if X is a normed space and X^* its topological dual space, the *dual pairing* of $x \in X$ and $x^* \in X^*$ will be denoted by $\langle x^*, x \rangle_{X^*, X} \coloneqq x^*(x)$.

While the *support* of a function $f\colon X \to \mathbb{R}$ is defined by

$$\operatorname{supp}(f) \coloneqq \overline{\{x \in X\colon f(x) \neq 0\}},$$

the *support* of a (nonnegative) regular Borel measure $0 \leq \mu \in \mathfrak{M}(X)$ is defined by

$$\operatorname{supp}(\mu) \coloneqq \{x \in X\colon \mu(N) \geq 0 \text{ for all open neighborhoods } N \in \mathfrak{B}(X) \text{ of } x\}.$$

We note that the closedness of both of these supports follows directly from their definitions.

Part I

The Infinite-Dimensional Case

Chapter 3

Bilevel Optimization of the Kantorovich Optimal Transport Problem

We begin by deriving and investigating the bilevel optimal transport problem in the infinite-dimensional case. While there are various formulations of optimal transport problems, such as those of Monge, see [57], or Beckmann, see [6], we will concern ourselves with the commonly known formulation that originated from Kantorovich and is a generalization of Monge's formulation.

First of all, however, we feel obliged to mention that parts of the present chapter have, in slightly different form, already been published in [45, 44].

3.1 Problem Statement

The first step will be to carefully define the Kantorovich problem of optimal transport, which will then take the role of the subordinate problem in the prototypical bilevel optimization problem that we motivated in Chapter 1.

To this end, for $d_1, d_2 \in \mathbb{N}$, let $\Omega_1 \subset \mathbb{R}^{d_1}$ and $\Omega_2 \subset \mathbb{R}^{d_2}$ be *compact domains* (i.e., closures of bounded non-empty open sets, see Chapter 2) such that their Cartesian product $\Omega := \Omega_1 \times \Omega_2$ has a locally Lipschitz boundary[1] which is negligible with respect to the Lebesgue measure.

Moreover, for the approximation results of Subchapter 3.4 we assume that there exists some $\Delta > 0$, such that the *extension domain* $\Omega^\Delta := \Omega_1^\Delta \times \Omega_2^\Delta$, where $\Omega_i^\Delta := \Omega_i + \overline{B_{\mathbb{R}^{d_i}}(0;\Delta)}$ for $i = 1, 2$, also has a locally Lipschitz boundary which is negligible with respect to the Lebesgue measure. This is, for example, satisfied in (but not limited to) the case that Ω_1 and Ω_2 are closures of bounded non-empty open convex sets, see e.g. [42, Corollary 1.2.2.3]. Note that the Cartesian product Ω and its extension Ω^Δ themselves are compact domains.

We denote the Lebesgue measure on the Borel σ-algebras $\mathfrak{B}(\Omega_1)$, $\mathfrak{B}(\Omega_2)$, and $\mathfrak{B}(\Omega)$ by λ_1, λ_2, and λ, respectively. In the above setting where Ω is the

[1] A bounded set is said to have a *locally Lipschitz boundary*, if each point on its boundary has a neighborhood whose intersection with said boundary is the graph of a Lipschitz continuous function, see e.g. [1, p. 83].

Cartesian product of Ω_1 and Ω_2, we find that λ is the uniquely determined product measure of λ_1 and λ_2, i.e., $\lambda = \lambda_1 \otimes \lambda_2$. Because all of the above sets have non-empty interiors, we moreover find that $|\Omega_1|, |\Omega_2|, |\Omega| > 0$.

Essential to the theory of optimal transportation are the terms "marginal" and "transport plan", the meanings of which are clarified in the following definition.

Definition 3.1. 1. Let $\mu_1 \in \mathfrak{M}(\Omega_1)$ and $\mu_2 \in \mathfrak{M}(\Omega_2)$ with $\mu_1, \mu_2 \geq 0$ be arbitrary nonnegative (signed) regular Borel measures. Throughout this thesis, we will call μ_1 and μ_2 *marginals*. We say that the marginals are *compatible*, if

$$\|\mu_1\|_{\mathfrak{M}(\Omega_1)} = \mu_1(\Omega_1) = \mu_2(\Omega_2) = \|\mu_2\|_{\mathfrak{M}(\Omega_2)}.$$

2. A *transport plan* (sometimes also referred to as a *coupling*) between the marginals μ_1 and μ_2 is a nonnegative regular Borel measure $\pi \in \mathfrak{M}(\Omega)$ which satisfies

$$\pi(B_1 \times \Omega_2) = \mu_1(B_1) \quad \text{and} \quad \pi(\Omega_1 \times B_2) = \mu_2(B_2) \tag{3.1}$$

for all measurable sets $B_1 \in \mathfrak{B}(\Omega_1)$ and $B_2 \in \mathfrak{B}(\Omega_2)$. Using, for $i = 1, 2$, the i-th *projection map*, $P_i \colon \Omega \ni (x_1, x_2) \mapsto x_i \in \Omega_i$, and the *pushforward measure* of π via P_i,

$$P_{i\#}\pi := \pi \circ P_i^{-1} \colon \mathfrak{B}(\Omega_i) \to \mathbb{R}, \quad B_i \mapsto \pi\big(P_i^{-1}(B_i)\big),$$

we can write (3.1) equivalently as

$$P_{1\#}\pi = \mu_1 \quad \text{and} \quad P_{2\#}\pi = \mu_2. \tag{3.2}$$

We denote the *set of transport plans* (or the *set of couplings*) between the marginals μ_1 and μ_2 by

$$\Pi(\mu_1, \mu_2) := \{\pi \in \mathfrak{M}(\Omega) \colon P_{1\#}\pi = \mu_1 \text{ and } P_{2\#}\pi = \mu_2\}.$$

We immediately see that the compatibility of the marginals is both sufficient and necessary for the set of transportation plans to be non-empty:

Lemma 3.2. *$\Pi(\mu_1, \mu_2) \neq \emptyset$ if and only if μ_1 and μ_2 are compatible.*

Proof. For the forward implication, let $\pi \in \Pi(\mu_1, \mu_2)$ be arbitrary. We immediately receive from the definition and the equivalence of (3.1) and (3.2) that

$$\|\mu_1\|_{\mathfrak{M}(\Omega_1)} = \mu_1(\Omega_1) = \pi(\Omega_1 \times \Omega_2) = \mu_2(\Omega_2) = \|\mu_2\|_{\mathfrak{M}(\Omega_2)},$$

so that μ_1 and μ_2 are compatible.

For the backward implication, let μ_1 and μ_2 be compatible marginals and abbreviate $m := \mu_1(\Omega_1) = \mu_2(\Omega_2)$. The product measure $\mu_1 \otimes \mu_2$ is a measure on $\mathfrak{B}(\Omega)$ which satisfies $(\mu_1 \otimes \mu_2)(B_1 \times B_2) = \mu_1(B_1)\mu_2(B_2)$ for all measurable sets $B_1 \in \mathfrak{B}(\Omega_1)$ and $B_2 \in \mathfrak{B}(\Omega_2)$, see e.g. [31, Satz V.1.5]. Because μ_1 and μ_2 are nonnegative and finite measures, $\mu_1 \otimes \mu_2$ is nonnegative and finite, too. Also, because Ω is Polish, Ulam's theorem (see e.g. [31, Satz VIII.1.16]) ensures the regularity of $\mu_1 \otimes \mu_2$, i.e., $\mu_1 \otimes \mu_2 \in \mathfrak{M}(\Omega)$. If we set $\pi := m^{-1}(\mu_1 \otimes \mu_2) \in \mathfrak{M}(\Omega)$, then

$$\pi(B_1 \times \Omega_2) = \mu_1(B_1)\frac{\mu_2(\Omega_2)}{m} = \mu_1(B_1) \quad \text{for all } B_1 \in \mathfrak{B}(\Omega_1).$$

Analogously, $\pi(\Omega_1 \times B_2) = \mu_2(B_2)$ for all $B_2 \in \mathfrak{B}(\Omega_2)$ so that $\pi \in \Pi(\mu_1, \mu_2)$. □

Now, let μ_1 and μ_2 be compatible marginals and consider some measurable *cost function* $c\colon \Omega \to \mathbb{R}$ which is continuous and therefore bounded and measurable. Then, the *Kantorovich (optimal transport) problem* is given by

$$\begin{aligned} \inf_{\pi} \quad & \mathcal{K}_c(\pi) \coloneqq \int_\Omega c \, d\pi \\ \text{s.t.} \quad & \pi \in \Pi(\mu_1, \mu_2), \ \pi \geq 0. \end{aligned} \tag{K}$$

The first thing to note about the Kantorovich problem is that it admits a (possibly nonunique) solution for every pair of compatible marginals and every cost function as specified above.

Lemma 3.3 ([49]). *Given the above assumptions on the domains, the marginals, and the cost function, the Kantorovich problem* (K) *possesses at least one optimal solution.*

We will now let the Kantorovich problem (K) take the role of the subordinate problem of the prototypical bilevel problem from Chapter 1, i.e., we will consider a special instance of the prototypical bilevel problem.

To this end, let us fix a target marginal $\mu_2^{\mathrm{d}} \in \mathfrak{P}(\Omega_2)$ and choose, for some $p > d_1 + d_2$, a continuous representative[2] c_{d} of the equivalence class $[c_{\mathrm{d}}] \in W^{1,p}(\Omega)$ to be the cost function of the Kantorovich problem. Given this data, we define the *bilevel Kantorovich (optimal transport) problem* to be the optimization problem

$$\begin{aligned} \inf_{\pi, \mu_1} \quad & \mathcal{J}(\pi, \mu_1) \\ \text{s.t.} \quad & \mu_1 \in \mathfrak{P}(\Omega_1), \\ & \pi \in \arg\min \left\{ \int_\Omega c_{\mathrm{d}} \, d\theta \colon \theta \in \Pi(\mu_1, \mu_2^{\mathrm{d}}), \, \theta \geq 0 \right\}, \end{aligned} \tag{BK}$$

where $\mathcal{J}\colon \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1) \to \mathbb{R} \cup \{+\infty\}$ is a target functional with the following properties:

1. $\mathcal{J}$ is *weak* lower semicontinuous*, i.e., for all sequences $(\pi, \mu_{1,n})_{n\in\mathbb{N}} \subset \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1)$ with $(\pi_n, \mu_{1,n}) \rightharpoonup^* (\pi, \mu_1) \in \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1)$ as $n \to \infty$, it holds that
$$\mathcal{J}(\pi, \mu_1) \leq \liminf_{n\to\infty} \mathcal{J}(\pi_n, \mu_{1,n}). \tag{3.3}$$

2. $\mathcal{J}$ is *bounded on bounded sets*, i.e., for all $M > 0$ it holds that
$$\sup_{\|(\pi,\mu_1)\|_{\mathfrak{M}(\Omega)\times\mathfrak{M}(\Omega_1)} \leq M} |\mathcal{J}(\pi, \mu_1)| < \infty. \tag{3.4}$$

3. There exists an *extension of the target functional* $\mathcal{J}^\Delta\colon \mathfrak{M}(\Omega^\Delta) \times \mathfrak{M}(\Omega_1^\Delta) \to \mathbb{R} \cup \{+\infty\}$ which itself is weak* lower semicontinuous, bounded on bounded sets, and satisfies
$$\mathcal{J}^\Delta(\pi, \mu_1) = \mathcal{J}\big(\pi|_{\mathfrak{M}(\Omega)}, \mu_1|_{\mathfrak{M}(\Omega_1)}\big) \tag{3.5}$$
for all $(\pi, \mu_1) \in \mathfrak{M}(\Omega^\Delta) \times \mathfrak{M}(\Omega_1^\Delta)$ with $\operatorname{supp}(\pi) \subset \Omega$ and $\operatorname{supp}(\mu_1) \subset \Omega_1$. Here, $\pi|_{\mathfrak{M}(\Omega)}$ and $\mu_1|_{\mathfrak{M}(\Omega_1)}$ denote the restrictions of $\pi\colon \mathfrak{B}(\Omega^\Delta) \to \mathbb{R}$ to $\mathfrak{B}(\Omega) \subset \mathfrak{B}(\Omega^\Delta)$ and $\mu_1\colon \mathfrak{B}(\Omega_1^\Delta) \to \mathbb{R}$ to $\mathfrak{B}(\Omega_1) \subset \mathfrak{B}(\Omega_1^\Delta)$, respectively.

[2] This continuous representative exists due to the Rellich-Kondrachov theorem, see e.g. [1, Theorem 6.3].

Remark 3.4. While we require the target functional $\mathcal{J}$ to be (weak*) lower semicontinuous to be able to prove the existence of solutions to the bilevel Kantorovich problem (BK), see the proof of Theorem 3.5, we will need the other two properties, i.e., the boundedness of $\mathcal{J}$ on bounded sets and the existence of the extension $\mathcal{J}^\Delta$, for the approximation results of Subchapter 3.4. We will present examples of target functionals that satisfy all of these three properties in Subchapter 3.5. ○

A beneficial feature of the bilevel Kantorovich problem (BK) is that, similar to the Kantorovich problem (K), that it has a solution, guaranteeing its well-posedness.

Theorem 3.5. *Given the above assumptions on the domains, the target marginal, the cost function, and the target functional, the bilevel Kantorovich problem* (BK) *possesses at least one optimal solution.*

Proof. We prove the result with the direct method of the calculus of variations. For that purpose, we denote (BK)'s feasible set by $\mathcal{F}$. To see that $\mathcal{F}$ is non-empty, let $\hat{\mu}_1 = \delta_{\hat{x}} \in \mathfrak{P}(\Omega_1)$ be the Dirac measure on Ω_1 for some arbitrary point $\hat{x} \in \Omega_1$. By construction, $\hat{\mu}_1$ and μ_2^{d} are compatible. Following Lemma 3.3, there exists an optimal transport plan $\hat{\pi}$ between $\hat{\mu}_1$ and μ_2^{d} w.r.t. the cost function c_{d} so that $(\hat{\pi}, \hat{\mu}_1) \in \mathcal{F}$.

Because $\mathcal{F}$ is non-empty, there exists a minimizing sequence $(\pi_n, \mu_{1,n})_{n\in\mathbb{N}} \subset \mathcal{F}$ so that

$$\lim_{n\to\infty} \mathcal{J}(\pi_n, \mu_{1,n}) = \inf_{(\pi,\mu_1)\in\mathcal{F}} \mathcal{J}(\pi, \mu_1) \in [-\infty, \infty).$$

The feasibility of the minimizing sequence implies

$$\|\pi_n\|_{\mathfrak{M}(\Omega)} = \pi_n(\Omega) = \mu_{1,n}(\Omega_1) = \|\mu_{1,n}\|_{\mathfrak{M}(\Omega_1)} = 1 \quad \text{for all } n \in \mathbb{N},$$

so it is contained in the unit ball of the space $\mathfrak{M}(\Omega)\times\mathfrak{M}(\Omega_1)$ with the latter being isomorphic to the continuous dual space of $C(\Omega) \times C(\Omega_1)$, see [2, 6.23 Riesz-Radon theorem] and Lemma D.1. By virtue of [2, Theorem 8.5], a subsequence $(\pi_{n_k}, \mu_{1,n_k})_{k\in\mathbb{N}}$ of the minimizing sequence then converges weakly* to some point $(\bar{\pi}, \bar{\mu}_1) \in \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1)$.

The stability result from [68, Theorem 1.50] ensures that the cluster point $(\bar{\pi}, \bar{\mu}_1)$ is contained in the feasible set $\mathcal{F}$: it states that $\mu_{1,n_k} \rightharpoonup^* P_{1\#}\bar{\pi}$ as well as $\mu_2^{\mathrm{d}} \rightharpoonup^* P_{2\#}\bar{\pi}$ and that $\bar{\pi}$ must be an optimal transport plan between the marginals $P_{1\#}\bar{\pi}$ and $P_{2\#}\bar{\pi}$ with respect to the cost function c_{d}. Because of the uniqueness of the weak* limit and because the sequence $(\mu_2^{\mathrm{d}})_{k\in\mathbb{N}}$ is constant, we find that $\bar{\mu}_1 = P_{1\#}\bar{\pi}$ as well as $\mu_2^{\mathrm{d}} = P_{2\#}\bar{\pi}$ and that $\bar{\pi}$ is thus an optimal transport plan between $\bar{\mu}_1$ and μ_2^{d} with respect to c_{d}. Additionally, because of $\bar{\mu}_1 = P_{1\#}\bar{\pi}$ and $\bar{\pi}(\Omega) = \mu_2^{\mathrm{d}}(\Omega_2) = 1$, we observe that $\bar{\mu}_1 \in \mathfrak{P}(\Omega_1)$. To summarize, we have shown that $(\bar{\pi}, \bar{\mu}_1) \in \mathcal{F}$.

The optimality of $(\bar{\pi}, \bar{\mu}_1)$ for (BK) now follows directly from the weak* lower semicontinuity of the target functional:

$$-\infty < \mathcal{J}(\bar{\pi}, \bar{\mu}_1) \le \liminf_{k\to\infty} \mathcal{J}(\pi_{n_k}, \mu_{1,n_k}) = \lim_{n\to\infty} \mathcal{J}(\pi_n, \mu_{1,n}) = \inf_{(\pi,\mu_1)\in\mathcal{F}} \mathcal{J}(\pi, \mu_1),$$

see (3.3). Hence, the point $(\bar{\pi}, \bar{\mu}_1)$ is optimal for (BK). □

Remark 3.6. Note that Theorem 3.5 is an existence-only result and that in general we cannot assume that the solution to (BK) is unique. Furthermore, the presupposed $W^{1,p}$ regularity of the cost function c_{d} is not needed in the above proof but will be crucial for the existence proof of the regularized bilevel problem ($\mathrm{BK}_\gamma^\delta$) from Subchapter 3.2, see the proof of Theorem 3.26, and we have therefore already assumed it for the formulation of (BK). ○

Now that we have established the existence of minimizers, one could be tempted to compute a solution to (BK) directly by applying a discretization to the problem's variables. After all, the discrete formulation of (K) is equivalent to a linear program, see Subchapter 4.1, and could be solved with the simplex method. However, there are a number of difficulties, for example,

- the solution of the lower level Kantorovich problem may not be unique; this prevents us from using the so-called implicit programming approach, which we describe in more detail at the beginning of Chapter 5;

- the analytical derivation of a solution to the Kantorovich problem is possible only in certain special cases, requiring the application of (possibly error-prone) numerical methods to obtain solutions for the lower level problem;

- computing the solutions to the Kantorovich (K) is numerically hard due to the *curse of dimensionality*, i.e., if the marginals were each discretized by, say, n variables, the solution of the Kantorovich problem would be a n^2-dimensional object (remember that it lives on the Cartesian product $\Omega_1 \times \Omega_2$).

Note that the second difficulty is linked to the third one, when using optimization algorithms in order to solve the bilevel problem in (BK). In each iteration of that algorithm, one needs to solve a (discretized) linear problem on a possibly huge space and this can become, of course, very costly!

In the following subchapter, we present an approach to the regularization of the Kantorovich problem, with which we can regularize the bilevel Kantorovich problem and make it easier computable.

3.2 Quadratic Regularization of the Kantorovich Problem

To overcome at least some of the difficulties mentioned at the end of the previous subchapter, the authors of [52] suggest to formulate the Kantorovich optimal transport problem on L^2 spaces instead of measure spaces and to add a quadratic regularization term to its target functional. To be more precise, given arbitrary compact domains $X_1 \subset \mathbb{R}^{d_1}$ and $X_2 \subset \mathbb{R}^{d_2}$, their Cartesian product $X := X_1 \times X_2$, the marginals $\mu_1 \in L^2(X_1)$ and $\mu_2 \in L^2(X_2)$, as well as a cost function $c \in L^2(X)$ and some *regularization parameter* $\gamma > 0$, they consider the

(quadratically L^2) regularized Kantorovich (optimal transport) problem

$$\begin{aligned} \inf_{\pi} \quad & \mathcal{K}_c^\gamma(\pi) \coloneqq (c,\pi)_{L^2(X)} + \frac{\gamma}{2}\|\pi\|^2_{L^2(X)} \\ \text{s.t.} \quad & \pi \in L^2(X), \quad \pi \geq 0 \quad \lambda\text{-a.e. in } X, \\ & \int_{X_2} \pi \, d\lambda_2 = \mu_1 \quad \lambda_1\text{-a.e. in } X_1, \\ & \int_{X_1} \pi \, d\lambda_1 = \mu_2 \quad \lambda_2\text{-a.e. in } X_2. \end{aligned} \tag{K_γ}$$

Remark 3.7. 1. Even though we have reserved the terms "marginal" and "transport plan" for elements of measure spaces, see Definition 3.1, we use it in this case too, since every absolutely integrable function can be interpreted as the density function of some measure, which becomes clear when considering the embedding $L^1(X) \hookrightarrow \mathfrak{M}(X)$ which is realized by means of the operator

$$\iota\colon L^1(X) \to \mathfrak{M}(X), \quad \iota(f)(B) \coloneqq \int_B f \, d\lambda, \quad f \in L^1(X), \ B \in \mathfrak{B}(X),$$

see Theorem B.16.

2. Figuratively speaking, in the case of (K_γ), the improved regularity of the marginals results in improved regularity of the optimal transport plan, and the quadratic regularization term in the objective function ensures the uniqueness of the solution.

3. The linear integral constraints defining the feasible set of (K_γ) are nothing else than the linear constraints of the Kantorovich problem (K), if we interpret μ_1, μ_2, and π as measures, see the first point of this remark. Using Fubini's theorem, we see that

$$\begin{aligned} \iota(\pi)(B_1 \times X_2) &= \int_{B_1 \times X_2} \pi \, d\lambda \\ &= \int_{B_1} \int_{X_2} \pi \, d\lambda_2 \, d\lambda_1 = \int_{B_1} \mu_1 \, d\lambda_1 = \iota(\mu_1)(B_1) \end{aligned}$$

for all measurable $B_1 \in \mathfrak{B}(X_1)$ and analogously

$$\iota(\pi)(X_1 \times B_2) = \iota(\mu_2)(B_2)$$

for all measurable $B_2 \in \mathfrak{B}(X_2)$.

○

We now collect some known results on the regularized Kantorovich problem (K_γ) which will be essential for this thesis.

Lemma 3.8 ([52, Lemma 2.1])**.** *Given the above assumptions on the domains, the marginals, and the cost function, the regularized Kantorovich problem* (K_γ) *admits a unique solution if and only if*

$$\mu_i \geq 0 \ \lambda_i\text{-a.e.}, \ i = 1,2, \quad \text{and} \quad \int_{X_1} \mu_1 \, d\lambda_1 = \int_{X_2} \mu_2 \, d\lambda_2 \,.$$

Theorem 3.9 ([52, Theorem 2.11]). *Let $c \in L^2(X)$ be bounded from below by some constant $\underline{c} > -\infty$ and $\mu_1 \in L^2(X_1)$ and $\mu_2 \in L^2(X_2)$ with $\mu_1, \mu_2 \geq \delta > 0$. Then $\pi \in L^2(X)$ is a solution of (K_γ) if and only if there exist* dual variables *$\alpha_1 \in L^2(X_1)$ and $\alpha_2 \in L^2(X_2)$ satisfying*

$$\pi = \frac{1}{\gamma}(\alpha_1 \oplus \alpha_2 - c)_+ \quad \lambda\text{-a.e. in } X, \tag{3.6a}$$

$$\int_{X_2} (\alpha_1 \oplus \alpha_2 - c)_+ \, d\lambda_1 = \gamma\mu_1 \quad \lambda_1\text{-a.e. in } X_1, \tag{3.6b}$$

$$\int_{X_1} (\alpha_1 \oplus \alpha_2 - c)_+ \, d\lambda_1 = \gamma\mu_2 \quad \lambda_2\text{-a.e. in } X_2. \tag{3.6c}$$

Remark 3.10. In Theorem 3.9,

$$(\alpha_1 \oplus \alpha_2)(x_1, x_2) \coloneqq \alpha_1(x_1) + \alpha_2(x_2) \quad \lambda\text{-a.e. in } X$$

refers to the *outer sum* of the functions α_1 and α_2, whereas,

$$u_+(x) \coloneqq \max\{u(x), 0\} \quad \text{and} \quad u_-(x) \coloneqq -\min\{u(x), 0\}$$

λ-a.e. in X, denote the *nonnegative part* and *nonpositive part* of u, respectively. We know from Corollary B.3 that $\alpha_1 \oplus \alpha_2$ is an element of $L^2(X)$. Therefore,

$$(\alpha_1 \oplus \alpha_2 - c)_+ = \chi_{\{\alpha_1 \oplus \alpha_2 - c \geq 0\}}(\alpha_1 \oplus \alpha_2 - c) \in L^2(X),$$

so that the equations of system (3.6) make sense. ○

Lemma 3.11 ([52, Section 2]). *The* (Lagrangian) dual problem *to (K_γ) is given by*

$$\sup_{\substack{\alpha_1 \in L^2(X_1),\\ \alpha_2 \in L^2(X_2)}} \mathcal{D}_c^\gamma(\alpha_1, \alpha_2) \coloneqq (\alpha_1, \mu_1)_{L^2(X_1)} + (\alpha_2, \mu_2)_{L^2(X_2)} - \tfrac{1}{2\gamma}\|(\alpha_1 \oplus \alpha_2 - c)_+\|^2_{L^2(X)}. \tag{KD_γ}$$

Moreover,

1. *the equations (3.6b) and (3.6c) are the first-order sufficient and necessary optimality condition of (KD_γ).*
2. *there is no duality gap, i.e., if π solves (K_γ) and (α_1, α_2) solves (KD_γ) (w.r.t. the same marginals μ_1 and μ_2), then $\mathcal{K}_c^\gamma(\pi) = \mathcal{D}_c^\gamma(\alpha_1, \alpha_2)$.*
3. *if (α_1, α_2) is a solution to (KD_γ), then $\mathcal{D}_c^\gamma(\alpha_1 + a, \alpha_2 - a) = \mathcal{D}_c^\gamma(\alpha_1, \alpha_2)$ for any $a \in \mathbb{R}$, i.e., the solution to (KD_γ) is not unique.*

The above results directly tackle two of the aforementioned difficulties and replacing Kantorovich problem by its regularized counterpart opens up several opportunities:

– In contrast to (K), the optimal solution to (K_γ) is unique, see Lemma 3.8. This implicitly defines a solution operator which maps the given data (the marginals and the cost function) to the unique solution (the transport plan) of the problem, allowing us to replace the Kantorovich problem by this solution operator, see the formulation of ($\mathrm{BK}_\gamma^\delta$) below.

– Theorem 3.9 allows us, to some extent, to avoid the curse of dimensionality. Instead of solving a linear optimization problem, it suffices to solve the nonlinear equations of (3.6) with respect to the dual variables α_1 and α_2. After discretizing the problem, this significantly reduces the number of variables required ($m + n$ instead of $m \cdot n$ variables!).

– Additionally, the structure of (3.6) allows for the application of a number of standard algorithms for the solution of a discretized version of the nonlinear equations, e.g. nonlinear Gauß-Seidel algorithm or semismooth Newton method, see e.g. [52, Section 3].

However, (K_γ) requires the marginals to no longer be measures, but to be elements of the corresponding L^2-spaces, which in general corresponds to an increase in regularity. But rather than restricting our choice of marginals to elements of $L^2(X_1)$ and $L^2(X_2)$, we preserve some generality and fit our data to the above situation by means of convolution. To this end, we need two definitions.

Definition 3.12 ([1, Definition 2.28])**.** For $d \in \mathbb{N}$, define a nonnegative, compactly supported smooth function $\varphi \in C_c^\infty(\mathbb{R}^d)$ with $\operatorname{supp}(\varphi) \subset \overline{B_{\mathbb{R}^d}(0;1)}$ by

$$\varphi(x) \coloneqq \begin{cases} k \exp\bigl(-1/(1-\|x\|^2)\bigr), & \text{if } \|x\| < 1, \\ 0, & \text{if } \|x\| \geq 1, \end{cases} \quad \text{for all } x \in \mathbb{R}^d.$$

In the above, let the scaling $k > 0$ be chosen in a way that $\int_{\mathbb{R}^d} \varphi(x)\, dx = 1$.

For $\delta > 0$, we receive a *mollifier* $\varphi_\delta \in C_c^\infty(\mathbb{R}^d)$ by defining

$$\varphi_\delta(x) \coloneqq \frac{\varphi(x/\delta)}{\delta^d} \quad \text{for all } x \in \mathbb{R}^d.$$

By construction,

$$\operatorname{supp}(\varphi_\delta) = \overline{B(0;\delta)}, \quad \varphi_\delta \geq 0, \quad \text{and} \quad \int_{\mathbb{R}^d} \varphi_\delta(x)\, dx = \int_{\overline{B(0;\delta)}} \varphi_\delta(x)\, dx = 1,$$

where the value of the integral follows from a substitution of variables.

Definition 3.13. Given some compact subset $X \subset \mathbb{R}^d$, with $d \in \mathbb{N}$, let $\mu \in \mathfrak{M}(X)$ be a nonnegative regular Borel measure and $\varphi_\delta \in C_c^\infty(\mathbb{R}^d)$, for $\delta > 0$, be a mollifier. Then, the *convolution of the measure μ with the mollifier φ* is defined by

$$(\varphi_\delta * \mu)(x) \coloneqq \int_X \varphi_\delta(x - y)\, d\mu(y) \quad \text{for all } x \in \mathbb{R}^d.$$

We use the above definitions to fit the marginals $\mu_1 \in \mathfrak{M}(\Omega_1)$ and $\mu_2 \in \mathfrak{M}(\Omega_2)$ to the setting of the regularized Kantorovich problem (K_γ). To this end, we choose a (the same for both marginals) *smoothing parameter* $\delta > 0$ as well as mollifiers $\varphi_1^\delta \in C_c^\infty(\mathbb{R}^{d_1})$ and $\varphi_2^\delta \in C_c^\infty(\mathbb{R}^{d_2})$. Then, for $i = 1, 2$, the convoluted marginals

$$\bigl(\varphi_i^\delta * \mu_i\bigr)(x_i) = \int_{\Omega_i} \varphi_i^\delta(x_i - y)\, d\mu_i(y) \quad \text{for all } x_i \in \mathbb{R}^{d_i},$$

are no longer measures, but smooth and compactly supported and thus (quadratically) integrable nonnegative functions.

As is known, this approach will enlarge the domains of the marginals. In order to avoid loss of information in the proximity of the boundary of the domains, we therefore define the *smoothed domains* $\Omega_i^\delta := \Omega_i + \overline{B(0;\delta)}$, $i = 1, 2$, to be the domains of the convoluted measures. As usual, we abbreviate their Cartesian product by $\Omega^\delta := \Omega_1^\delta \times \Omega_2^\delta$. Moreover, we set $B_i^\delta := \overline{B(0;\delta)} \subset \mathbb{R}^{d_i}$, $i = 1, 2$, and note that the compactness of Ω_i and B_i^δ is carried over to their Minkowski sum $\Omega_i + B_i^\delta$. For further information on the convolution of the marginals and mollifiers we refer the interested reader to Appendix A.

Also, since we want to be able to use the dual representation of the optimal transport plan from Theorem 3.9, we raise the convoluted marginals by a bit, i.e., for $i = 1, 2$ we define the *convolution (& raising operator)* $\mathcal{T}_i^\delta \colon \mathfrak{M}(\Omega_i) \to L^2(\Omega_i^\delta)$ where

$$\mathcal{T}_i^\delta(\mu_i) := (\varphi_i^\delta * \mu_i + \delta|\Omega_{3-i}^\delta|)|_{\Omega_i^\delta},$$

which turns a nonnegative measure into a smooth and strictly positive function, see the following remark.

Remark 3.14.

1. By the above construction, $\mathcal{T}_i^\delta(\mu_i) \in L^\infty(\Omega_i^\delta) \subset L^2(\Omega_i^\delta)$ and $\mathcal{T}_i^\delta(\mu_i) \geq \delta \min\{|\Omega_1^\delta|, |\Omega_2^\delta|\} > 0$ for $i = 1, 2$. Theorem A.3 and Lemma A.4 combined with the assumption on the mass of the marginals and mollifiers yield that

$$\|\mathcal{T}_i^\delta(\mu_i)\|_{L^1(\Omega_i^\delta)} = \|\varphi_i^\delta\|_{L^1(B_i^\delta)}\|\mu_i\|_{\mathfrak{M}(\Omega_i)} + \delta|\Omega_1^\delta||\Omega_2^\delta| = 1 + \delta|\Omega_1^\delta||\Omega_2^\delta|$$

   for $i = 1, 2$. Lemma 3.8 then implies the existence of a unique solution to (K_γ) for each $\gamma > 0$ and $\delta > 0$ and Theorem 3.9 yields the representation of said solution by means of the dual variables $\alpha_1 \in L^2(\Omega_1^\delta)$ and $\alpha_2 \in L^2(\Omega_2^\delta)$.

2. Not only does the convolution of the marginals serve to fit our data to the setting of the regularized Kantorovich problem, it is also essential for existence proof of the regularized bilevel problem ($\mathrm{BK}_\gamma^\delta$) defined below. This is because the solution operator of the regularized Kantorovich problem (K_γ) is not weak* continuous (see Example 3.18) but only Hölder continuous. Hence, the compactness of the convolution operator is needed to guarantee the admissibility of the limiting transport plan, see the proof of Theorem 3.26.

○

Another point we have to address before we can actually formulate the regularized version of the bilevel Kantorovich problem are the regularities of the cost function and the solution of regularized Kantorovich problem. Just like the marginals, the cost function of the regularized problem needs to be an element of the corresponding L^2 space. At the same time, its solution is an element of the same L^2 space, but the target functional of the bilevel Kantorovich problem only operates on regular Borel measures. To solve this discrepancy, we have the following definition:

Definition 3.15. Let $\mathcal{E}_\delta \colon C(\Omega) \to L^2(\Omega^\delta)$ be the *extension (by zero) operator* that extends a continuous function $f \colon \Omega \to \mathbb{R}$ to a not necessarily continuous but square integrable function $\mathcal{E}_\delta(f) \colon \Omega^\delta \to \mathbb{R}$ in a way that $\mathcal{E}_\delta(f) \equiv f$ on Ω and $\mathcal{E}_\delta(f) \equiv 0$ on $\Omega_\delta \setminus \Omega$.

We denote its adjoint by $\mathcal{E}_\delta^*$, which is the unique operator $\mathcal{E}_\delta^*\colon L^2(\Omega^\delta) \to \mathfrak{M}(\Omega)$ given by $\mathcal{E}_\delta^*(u)(B) \coloneqq \int_B u \, d\lambda$ for all $B \in \mathfrak{B}(\Omega)$ and all $u \in L^2(\Omega^\delta)$.

Remark 3.16. Even though the operator $\mathcal{E}_\delta$ only allows for continuous functions as input, we (ab-)use its symbol for elements of $W^{1,p}(\Omega)$ too. According to the Rellich-Kondrachov theorem each of these Sobolev functions has a continuous representative which we then plug into the operator $\mathcal{E}_\delta$, which justifies the ambiguous use of its symbol. ○

After our preparatory considerations, we are now in a position to define for fixed $\gamma > 0$ and $\delta > 0$, the *(quadratically L^2-) regularized (& $W^{1,p}$-penalized) bilevel Kantorovich problem.* Given the domains Ω_1, Ω_2, and Ω from Subchapter 3.1, $\mu_2^{\mathrm{d}} \in \mathfrak{P}(\Omega_2)$, and the continuous representative c_{d} of $[c_{\mathrm{d}}] \in W^{1,p}(\Omega)$ (remember that $p > d_1 + d_2$), we consider the problem

$$\begin{aligned} \inf_{\pi,\mu_1,c} \quad & \mathcal{J}_\gamma(\pi,\mu_1,c) \coloneqq \mathcal{J}(\pi,\mu_1) + \tfrac{1}{\gamma}\|c - c_{\mathrm{d}}\|^p_{W^{1,p}(\Omega)} \\ \text{s.t.} \quad & c \in W^{1,p}(\Omega),\ \mu_1 \in \mathfrak{P}(\Omega_1), \\ & \pi = (\mathcal{E}_\delta^* \circ \mathcal{S}_\gamma)\big(\mathcal{E}_\delta(c), \mathcal{T}_1^\delta(\mu_1), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big). \end{aligned} \tag{$\mathrm{BK}_\gamma^\delta$}$$

Remark 3.17. In the context of the regularized bilevel Kantorovich problem ($\mathrm{BK}_\gamma^\delta$) we want to mention the following:

- As announced earlier, we replaced the lower-level Kantorovich problem from the formulation of the non-regularized bilevel Kantorovich problem (BK) by its solution operator $\mathcal{S}_\gamma$, which is rigorously defined in Subchapter 3.3 below.
- In comparison to (BK), in ($\mathrm{BK}_\gamma^\delta$) we use the cost function of the Kantorovich problem as an optimization variable as well. This is motivated by the fact that we expect that the set of optimization variables is not rich enough to obtain non-trivial recovery sequences for the approximation of solutions of (BK). This is particularly evident in Chapter 4.5, where we only succeeded in constructing a nontrivial recovery sequence with the aid of the cost function being an optimization variable.

○

Of course, the first question that arises is whether ($\mathrm{BK}_\gamma^\delta$) is well-posed and possesses an optimal solution. Furthermore, we wish to know whether solutions to the non-regularized problem (BK) can be approximated by a sequence of solutions to the regularized problems ($\mathrm{BK}_\gamma^\delta$). We explore the answers to those questions in the next two subchapters, beginning with the former.

3.3 Existence of Solutions to the Regularized Bilevel Kantorovich Problem

Using this subchapter we show that the regularized bilevel Kantorovich problem ($\mathrm{BK}_\gamma^\delta$) possesses at least one optimal solution.

As was the case in the formulation of (K_γ), we again consider arbitrary compact domains $X_i \subset \mathbb{R}^{d_i}$, $i = 1, 2$, and their Cartesian product $X \coloneqq X_1 \times X_2$.

Given the scalar lower bounds $\underline{c} > -\infty$ and $\underline{\mu} > 0$ as well as the mass $m > 0$, we define the *set of cost functions (bounded from below)*,

$$\mathcal{C}_{\underline{c}}(X) := \{c \in L^2(X) \colon c \geq \underline{c} \ \lambda\text{-a.e. in } X\},$$

and the set of *(strictly positive) compatible marginals*,

$$\mathcal{M}^m_{\underline{\mu}}(X_1, X_2) := \Big\{(\mu_1, \mu_2) \in L^2(X_1) \times L^2(X_2) \colon \int_{X_i} \mu_i \, d\lambda_i = m, \ \mu_i \geq \underline{\mu} \ \lambda_i\text{-a.e. in } X_i, \ i = 1, 2, \Big\}.$$

A major difficulty in the existence proof of ($\mathrm{BK}^{\delta}_{\gamma}$) is the fact that the *solution operator* of the regularized Kantorovich problem,

$$\mathcal{S}_\gamma \colon \mathcal{C}_{\underline{c}}(X) \times \mathcal{M}^m_{\underline{\mu}}(X_1, X_2) \to L^2(X), \quad (c, \mu_1, \mu_2) \mapsto \pi,$$

with $\gamma > 0$ and π being the unique solution to (K_γ) with respect to c, μ_1, and μ_2, is not continuous w.r.t. the weak* convergence as the following example shows:

Example 3.18 ($\mathcal{S}_\gamma$ Not Weak Continuous)**.** Consider the compact domains $X_1 = X_2 = [0, 1]$, the regularization parameter $\gamma = 1$ (this choice is only for convenience), and the cost function

$$c(x_1, x_2) = \frac{1}{4}|x_1 - x_2|^2.$$

Moreover, for $n \in \mathbb{N}$, define

$$f_n(x) := \operatorname{sgn}\big(\sin(2\pi n x)\big) \quad \text{for all } x \in [0, 1],$$

as well as

$$\alpha_{1,n} := f_n + \frac{9}{4}\chi_{[0,\frac{1}{2}]} + \frac{5}{4}\chi_{(\frac{1}{2},1]} \quad \text{and} \quad \alpha_{2,n} := -\frac{1}{2}\chi_{(\frac{1}{2},1]}.$$

Based on this definitions, one can construct a sequence of transport plans $(\pi_n)_{n\in\mathbb{N}} \subset L^2([0, 1]^2)$ via

$$\begin{aligned}\pi_n(x_1, x_2) &:= \big(\alpha_{1,n}(x_1) + \alpha_{2,n}(x_2) - c(x_1, x_2)\big)_+ \\ &= \begin{cases} f_n(x_1) + \frac{9}{4} - \frac{1}{4}|x_1 - x_2|^2, & \text{if } x_1, x_2 \in [0, \frac{1}{2}], \\ f_n(x_1) + \frac{5}{4} - \frac{1}{4}|x_1 - x_2|^2, & \text{if } x_1 \in (\frac{1}{2}, 1], \ x_2 \in [0, \frac{1}{2}], \\ f_n(x_1) + \frac{7}{4} - \frac{1}{4}|x_1 - x_2|^2, & \text{if } x_1 \in [0, \frac{1}{2}], \ x_2 \in (\frac{1}{2}, 1], \\ \big(f_n(x_1) + \frac{3}{4} - \frac{1}{4}|x_1 - x_2|^2\big)_+, & \text{if } x_1, x_2 \in (\frac{1}{2}, 1]. \end{cases}\end{aligned}$$

If we set

$$\mu_{1,n} := \int_0^1 \pi_n \, d\lambda_2 \quad \text{and} \quad \mu_{2,n} := \int_0^1 \pi_n \, d\lambda_1,$$

then we find that, for $i = 1, 2$, $\mu_i \geq \frac{1}{16}$ λ_i-almost everywhere. We can therefore apply Theorem 3.9, to obtain that $\pi_n = \mathcal{S}_1(c, \mu_{1,n}, \mu_{2,n})$ for all $n \in \mathbb{N}$. Let us take a close look at the last case in the definition of π_n and abbreviate

$$F_n(x_1, x_2) := \Big(f_n(x_1) + \frac{3}{4} - \frac{1}{4}|x_1 - x_2|^2\Big)_+$$

$$= \begin{cases} \frac{7}{4} - \frac{1}{4}|x_1 - x_2|^2, & \text{if } x_1 \in \left(\frac{2k-2}{2n}, \frac{2k-1}{2n}\right),\ k = 1, \dots, n, \\ 0, & \text{if } x_1 \in \left(\frac{2k-1}{2n}, \frac{2k}{2n}\right),\ k = 1, \dots, n. \end{cases}$$

If we set $F(x_1, x_2) \coloneqq \frac{1}{2} \cdot \left(\frac{7}{4} - \frac{1}{4}|x_1 - x_2|^2\right) + \frac{1}{2} \cdot 0 = \frac{7}{8} - \frac{1}{8}|x_1 - x_2|^2$, then for every $\phi \in C_c^\infty([0,1]^2)$ we get that

$$\begin{aligned} &\left| \int_{[0,1]^2} \phi(F_n - F)\, dx \right| \\ &= \left| \int_0^1 \sum_{k=1}^n \left(\int_{\frac{2k-2}{2n}}^{\frac{2k-1}{2n}} (\phi F)(x_1, x_2)\, dx_1 - \int_{\frac{2k-1}{2n}}^{\frac{2k}{2n}} (\phi F)(x_1, x_2)\, dx_1 \right) dx_2 \right| \\ &\leq \int_0^1 \sum_{k=1}^n \int_{\frac{2k-2}{2n}}^{\frac{2k-1}{2n}} \left| (\phi F)(x_1, x_2) - (\phi F)\left(x_1 + \frac{1}{2n}, x_2\right) \right| dx_1\, dx_2 \\ &\leq \int_0^1 \sum_{k=1}^n \int_{\frac{2k-2}{2n}}^{\frac{2k-1}{2n}} L_{\phi F} \frac{1}{2n}\, dx_1\, dx_2 \;=\; L_{\phi F} \frac{1}{4n} \;\xrightarrow[n\to\infty]{}\; 0, \end{aligned}$$

where $L_{\phi F} > 0$ denotes the Lipschitz constant of ϕF. Because $C_c^\infty([0,1]^2)$ is dense in $L^2([0,1]^2)$, see e.g. [1, Corollary 2.30], for any $\phi \in L^2([0,1]^2)$ and every $\varepsilon > 0$ there exists some $\phi_\varepsilon \in C_c^\infty([0,1]^2)$ with $\|\phi - \phi_\varepsilon\|_{L^2([0,1]^2)} < \varepsilon$. Hence,

$$\begin{aligned} &\left| \int_{[0,1]^2} \phi(F_n - F)\, dx \right| \\ &\quad \leq \int_{[0,1]^2} |\phi - \phi_\varepsilon||F_n - F|\, dx + \left| \int_{[0,1]^2} \phi_\varepsilon(F_n - F)\, dx \right| \\ &\quad \leq \|F_n - F\|_{L^2([0,1]^2)} \|\phi - \phi_\varepsilon\|_{L^2([0,1]^2)} + \left| \int_{[0,1]^2} \phi_\varepsilon(F_n - F)\, dx \right| < 2\varepsilon \end{aligned}$$

for all n sufficiently large. Therefore, $F_n \rightharpoonup F$ in $L^2([0,1]^2)$. Together with $f_n \rightharpoonup 0$, this shows that

$$\pi_n \;\rightharpoonup\; \bar{\pi} = \begin{cases} \frac{9}{4} - \frac{1}{4}|x_1 - x_2|^2, & \text{if } x_1, x_2 \in [0, \frac{1}{2}], \\ \frac{5}{4} - \frac{1}{4}|x_1 - x_2|^2, & \text{if } x_1 \in (\frac{1}{2}, 1],\ x_2 \in [0, \frac{1}{2}], \\ \frac{7}{4} - \frac{1}{4}|x_1 - x_2|^2, & \text{if } x_1 \in [0, \frac{1}{2}],\ x_2 \in (\frac{1}{2}, 1], \\ \frac{7}{8} - \frac{1}{8}|x_1 - x_2|^2, & \text{if } x_1, x_2 \in (\frac{1}{2}, 1], \end{cases} \quad \text{as } n \to \infty.$$

The weak convergence of the transport plans implies the convergences $\mu_{1,n} \rightharpoonup \bar{\mu}_1 \coloneqq \int_0^1 \bar{\pi}\, d\lambda_2$ and $\mu_{2,n} \rightharpoonup \bar{\mu}_2 \coloneqq \int_0^1 \bar{\pi}\, d\lambda_1$, because $\mu_{1,n}$ and $\mu_{2,n}$ are linear and continuous images of the transport plan π_n and this is preserved by weak convergence. Now, assume that

$$\bar{\pi} = (\bar{\alpha}_1 \oplus \bar{\alpha}_2 - c)_+ \tag{3.7}$$

for some $\bar{\alpha}_1, \bar{\alpha}_2 \in L^2([0,1])$. Because of $\bar{\pi} > 0$ a.e. in $[0,1]^2$, it must hold that $\bar{\pi} = \bar{\alpha}_1 \oplus \bar{\alpha}_2 - c$ or equivalently

$$\bar{\alpha}_1 \oplus \bar{\alpha}_2 = \begin{cases} \frac{9}{4}, & \text{in } \left[0, \frac{1}{2}\right]^2, \\ \frac{5}{4}, & \text{in } \left(\frac{1}{2}, 1\right] \times \left[0, \frac{1}{2}\right], \\ \frac{7}{4}, & \text{in } \left[0, \frac{1}{2}\right] \times \left(\frac{1}{2}, 1\right], \\ \frac{7}{8} + \frac{1}{8}|x_1 - x_2|^2, & \text{in } \left(\frac{1}{2}, 1\right]^2. \end{cases}$$

Choosing an arbitrary representative of the equivalence class $[\bar{\alpha}_2]$ and fixing an Lebesgue point $\tilde{x}_2 \in (0, 1/2)$, we obtain

$$\bar{\alpha}_1(x_1) = \begin{cases} \frac{9}{4} - \bar{\alpha}_2(\tilde{x}_2), & \text{if } 0 \le x_1 \le \frac{1}{2}, \\ \frac{5}{4} - \bar{\alpha}_2(\tilde{x}_2), & \text{if } \frac{1}{2} < x_1 \le 1. \end{cases} \tag{3.8}$$

Similarly, fix $\hat{x}_2 \in (1/2, 1)$ to obtain

$$\bar{\alpha}_1(x_1) = \begin{cases} \frac{7}{4} - \bar{\alpha}_2(\hat{x}_2), & \text{if } 0 \le x_1 \le \frac{1}{2}, \\ \frac{7}{8} + \frac{1}{8}|x_1 - \hat{x}_2|^2 - \bar{\alpha}_2(\hat{x}_2), & \text{if } \frac{1}{2} < x_1 \le 1. \end{cases} \tag{3.9}$$

Obviously, the functions in (3.8) and (3.9) cannot be the same, regardless the choice of $\bar{\alpha}_2$. Therefore, (3.7) cannot be true and by Theorem 3.9, $\bar{\pi}$ is not the optimal transport plan between $\bar{\mu}_1$ and $\bar{\mu}_2$. To summarize our findings,

$$(\mu_{1,n}, \mu_{2,n}) \rightharpoonup (\bar{\mu}_1, \bar{\mu}_2) \quad \not\Longrightarrow \quad \mathcal{S}_1(c, \mu_{1,n}, \mu_{2,n}) \rightharpoonup \mathcal{S}_1(c, \bar{\mu}_1, \bar{\mu}_2),$$

i.e., the solution operator of the regularized Kantorovich problem (K_γ) is not continuous w.r.t. the weak* convergence. Again, the choice $\gamma = 1$ in the above example was only for convenience. The same example holds (up to scaling) for arbitrary choices of $\gamma > 0$ so that none of the solution operators $\mathcal{S}_\gamma$ for $\gamma > 0$ are continuous w.r.t. the weak* convergence. ◊

Nevertheless, we can show that the solution operators are Hölder continuous, by deriving L^2-bounds for both the solution of the corresponding regularized Kantorovich problem and the dual variables evolving from Theorem 3.9.

Lemma 3.19. *Let $\gamma > 0$ and $(c, \mu_1, \mu_2) \in \mathcal{C}_{\underline{c}}(X) \times \mathcal{M}^m_{\underline{\mu}}(X_1, X_2)$ be arbitrary. Then, the solution of the regularized Kantorovich problem (K_γ), $\pi = \mathcal{S}_\gamma(c, \mu_1, \mu_2)$, is bounded by*

$$\|\pi\|_{L^2(X)} \le C \cdot \big(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)}\big),$$

where $C = C(\gamma, m) > 0$ is a constant solely depending of γ and m.

Proof. Because the cost function is bounded from below and the marginals are strictly positive, Theorem 3.9 implies the existence of dual variables $\alpha_1 \in L^2(X_1)$ and $\alpha_2 \in L^2(X_2)$ such that the optimality in (3.6) is fulfilled. Multiplying (3.6b) and (3.6c) with α_1 and α_2, respectively, integrating and adding the resulting equations leads to

$$\begin{aligned} \gamma \|\pi\|^2_{L^2(X)} &= \int_X \pi(\alpha_1 \oplus \alpha_2 - c) \, d\lambda \\ &= \int_{X_1} \mu_1 \alpha_1 \, d\lambda_1 + \int_{X_2} \mu_2 \alpha_2 \, d\lambda_2 - \int_X \pi c \, d\lambda, \end{aligned} \tag{3.10}$$

where we used (3.6a) and that $x_+ x = (x_+)^2$ for all $x \in \mathbb{R}$. Exploiting the equality of the mass of the marginals, i.e.,

$$\|\mu_1\|_{L^1(X_1)} = \int_{X_1} \mu_1 \, d\lambda_1 = m = \int_{X_2} \mu_2 \, d\lambda_2 = \|\mu_2\|_{L^1(X_2)},$$

we obtain, for $i = 1, 2$, using Fubini's theorem

$$\int_{X_i} \mu_i \alpha_i \, d\lambda_i = \frac{1}{m} \int_{X_i} \mu_i \alpha_i \int_{X_{3-i}} \mu_{3-i} \, d\lambda_{3-i} \, d\lambda_i = \frac{1}{m} \int_X (\mu_1 \otimes \mu_2) \, \alpha_i \, d\lambda \,.$$

In the above, $(\mu_1 \otimes \mu_2)(x_1, x_2) \coloneqq \mu_1(x_1)\mu_2(x_2)$ λ-a.e. in X refers to the *tensor product* of the functions μ_1 and μ_2. Corollary B.3 guarantees that $\mu_1 \otimes \mu_2 \in L^2(X)$. This allows us to estimate (3.10) by

$$\begin{aligned}
&\gamma \|\pi\|^2_{L^2(X)} \\
&\quad = \frac{1}{m} \int_X (\mu_1 \otimes \mu_2)(\alpha_1 \oplus \alpha_2 - c) \, d\lambda + \frac{1}{m} \int_X (\mu_1 \otimes \mu_2) c \, d\lambda - \int_X \pi c \, d\lambda \\
&\quad \leq \frac{1}{m} \int_X (\mu_1 \otimes \mu_2)(\alpha_1 \oplus \alpha_2 - c)_+ \, d\lambda + \frac{1}{m} \int_X |(\mu_1 \otimes \mu_2) c| \, d\lambda + \int_X |\pi c| \, d\lambda \\
&\quad \leq \frac{\gamma}{m} \|\pi\|_{L^2} \|\mu_1 \otimes \mu_2\|_{L^2} + \frac{1}{m} \|\mu_1 \otimes \mu_2\|_{L^2} \|c\|_{L^2} + \|\pi\|_{L^2} \|c\|_{L^2}.
\end{aligned} \tag{3.11}$$

Next, we apply the scaled version of Young's inequality from Lemma D.3 to obtain

$$\frac{\gamma}{m} \|\pi\|_{L^2(X)} \|\mu_1 \otimes \mu_2\|_{L^2(X)} \leq \frac{\gamma}{3} \|\pi\|^2_{L^2(X)} + \frac{3\gamma}{m^2} \|\mu_1 \otimes \mu_2\|^2_{L^2(X)} \tag{3.12}$$

and

$$\|\pi\|_{L^2(X)} \|c\|_{L^2(X)} \leq \frac{\gamma}{3} \|\pi\|^2_{L^2(X)} + \frac{3}{\gamma} \|c\|^2_{L^2(X)}. \tag{3.13}$$

Substituting (3.12) and (3.13) into (3.11), we receive

$$\begin{aligned}
\|\pi\|^2_{L^2(X)} &\leq \frac{9}{m^2} \|\mu_1 \otimes \mu_2\|^2_{L^2(X)} + \frac{3}{m\gamma} \|\mu_1 \otimes \mu_2\|_{L^2(X)} \|c\|_{L^2(X)} + \frac{9}{\gamma^2} \|c\|^2_{L^2(X)} \\
&\leq \Big(\frac{3}{m} \|\mu_1 \otimes \mu_2\|_{L^2(X)} + \frac{3}{\gamma} \|c\|_{L^2(X)} \Big)^2.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\|\pi\|_{L^2(X)} &\leq \frac{3}{m} \|\mu_1 \otimes \mu_2\|_{L^2(X)} + \frac{3}{\gamma} \|c\|_{L^2(X)} \\
&\leq \max\Big\{ \frac{3}{m}, \frac{3}{\gamma} \Big\} \cdot \big(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)} \big),
\end{aligned}$$

which yields the claim. □

Remark 3.20. Although C and all of the following constants may or may not additionally depend on the domains X_1, X_2, and X, we are content with emphasizing only the dependence on the parameters γ, $\underline{c}$, $\underline{\mu}$, and m, since the domains were fixed from the very beginning of this subchapter.

This means that whenever a constant named C appears, we know that it depends directly or indirectly only on the given parameters and that there is no further dependence on entities other than those mentioned above. ○

To ease the subsequent argumentation, we make the following technical assumption:

Assumption 3.21. If $(\alpha_1, \alpha_2) \in L^2(X_1) \times L^2(X_2)$ is a given pair of dual variables to a given optimal solution of (K_γ) as given in Theorem 3.9, then $\int_{X_2} \alpha_2 \, d\lambda_2 = 0$, i.e., α_2 is a *zero-mean dual variable*.

Remark 3.22. Although this assumption may seem restrictive at first glance, it is actually not a limitation. If a given α_2 were not a zero-mean dual variable, we would abbreviate $a := |X_2|^{-1} \int_{X_2} \alpha_2 \, d\lambda_2$ and observe that

$$\begin{aligned} |a| &\leq |X_2|^{-1} \int_{X_2} |\alpha_2| \, d\lambda_2 \\ &\leq |X_2|^{-1} \|\mathbb{1}\|_{L^2(X_2)} \|\alpha_2\|_{L^2(X_2)} = |X_2|^{-\frac{1}{2}} \|\alpha_2\|_{L^2(X_2)} < \infty. \end{aligned}$$

We could then consider the pair

$$(\tilde{\alpha}_1, \tilde{\alpha}_2) := (\alpha_1 + a, \alpha_2 - a) \in L^2(X_1) \times L^2(X_2),$$

which is according to Lemma 3.11 also a solution to (KD_γ) and therefore satisfies the conditions in (3.6). A quick calculation then shows that $\tilde{\alpha}_2$'s mean value indeed vanishes:

$$\int_{X_2} \tilde{\alpha}_2 \, d\lambda_2 = \int_{X_2} \alpha_2 \, d\lambda_2 - a \int_{X_2} \mathbb{1} \, d\lambda_2 = \int_{X_2} \alpha_2 \, d\lambda_2 - \int_{X_2} \alpha_2 \, d\lambda_2 = 0.$$

Also, for $i = 1, 2$, $\|\tilde{\alpha}_i\|_{L^1(X_i)} \leq C$ implies $\|\alpha_i\|_{L^1(X_i)} \leq C + |a||X_i|$, i.e., the boundedness of the original solution. ○

Next, we wish to determine L^2-bounds for the dual variables. This is a crucial task to show the Hölder continuity of $\mathcal{S}_\gamma$. Prior to this, however, we need the following lemma to establish L^1-bounds for the dual variables.

Lemma 3.23. *Let Assumption 3.21 hold and let $\gamma > 0$ and $(c, \mu_1, \mu_2) \in \mathcal{C}_{\underline{c}}(X) \times \mathcal{M}^m_{\underline{\mu}}(X_1, X_2)$ be arbitrary and consider the corresponding optimal solution of the regularized Kantorovich problem (K_γ), namely $\pi = \mathcal{S}_\gamma(c, \mu_1, \mu_2)$. Then, the total masses of the dual variables $\alpha_1 \in L^2(X_1)$ and $\alpha_2 \in L^2(X_2)$ from Theorem 3.9 are bounded by some constant $C = C(\gamma, \underline{\mu}, m) > 0$, i.e.,*

$$\|\alpha_1\|_{L^1(X_1)}, \|\alpha_2\|_{L^1(X_2)} \leq C \cdot \left(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)} + 1 \right)^2.$$

Proof. Following our Assumption 3.21, we assume that (α_1, α_2) is a zero-mean dual solution. The target functional of the primal problem (K_γ) is bounded by

$$\mathcal{K}^\gamma_c(\pi) = (c, \pi)_{L^2(X)} + \frac{\gamma}{2} \|\pi\|^2_{L^2(X)} \geq -\|c\|_{L^2(X)} \|\pi\|_{L^2(X)}.$$

The strong duality of (K_γ) and (KD_γ), see Lemma 3.11, implies that

$$\mathcal{D}^\gamma_c(\alpha_1, \alpha_2) = \mathcal{K}^\gamma_c(\pi) \geq -\|c\|_{L^2(X)} \|\pi\|_{L^2(X)},$$

and therefore, similar to the proof of Lemma 3.19, we estimate that

$$\begin{aligned} &\|c\|_{L^2(X)} \|\pi\|_{L^2(X)} \\ &\quad \geq \frac{1}{2\gamma} \|(\alpha_1 \oplus \alpha_2 - c)_+\|^2_{L^2(X)} - \int_{X_1} \alpha_1 \mu_1 \, d\lambda_1 - \int_{X_2} \alpha_2 \mu_2 \, d\lambda_2 \end{aligned}$$

$$\geq -\frac{1}{m}\int_X (\mu_1 \otimes \mu_2)(\alpha_1 \oplus \alpha_2)\ d\lambda$$
$$\geq -\frac{1}{m}\int_X (\mu_1 \otimes \mu_2)(\alpha_1 \oplus \alpha_2 - c)\ d\lambda - \frac{1}{m}\|\mu_1 \otimes \mu_2\|_{L^2(X)}\|c\|_{L^2(X)}.$$

Using $u(x) = u_+(x) - u_-(x)$ with $u_+(x), u_-(x) \geq 0$ for almost all $x \in X$ for all $u \in L^2(X)$, see Remark 3.10, and the marginal's lower bound $\underline{\mu} > 0$, we find that

$$\begin{aligned}
&\|c\|_{L^2(X)}\Big(\|\pi\|_{L^2(X)} + \frac{1}{m}\|\mu_1 \otimes \mu_2\|_{L^2(X)}\Big) \\
&\quad \geq -\frac{1}{m}\int_X (\mu_1 \otimes \mu_2)(\alpha_1 \oplus \alpha_2 - c)_+\ d\lambda + \frac{1}{m}\int_X (\mu_1 \otimes \mu_2)(\alpha_1 \oplus \alpha_2 - c)_-\ d\lambda \\
&\quad \geq -\frac{\gamma}{m}\|\mu_1 \otimes \mu_2\|_{L^2(X)}\|\pi\|_{L^2(X)} + \frac{\underline{\mu}^2}{m}\int_X (\alpha_1 \oplus \alpha_2 - c)_-\ d\lambda\,.
\end{aligned}$$

Rearranging and using π's L^2-bound from Lemma 3.19, we obtain that

$$\begin{aligned}
&\|(\alpha_1 \oplus \alpha_2 - c)_-\|_{L^1(X)} \\
&\quad \leq \frac{m}{\underline{\mu}^2}\|c\|_{L^2(X)}\Big(\|\pi\|_{L^2(X)} + \frac{1}{m}\|\mu_1 \otimes \mu_2\|_{L^2(X)}\Big) + \frac{\gamma}{\underline{\mu}^2}\|\mu_1 \otimes \mu_2\|_{L^2(X)}\|\pi\|_{L^2(X)} \\
&\quad \leq C \cdot \big(\|\mu_1 \otimes \mu_2\|^2_{L^2(X)} + \|c\|^2_{L^2(X)}\big),
\end{aligned}$$

with some constant $C = C(\underline{\mu}, m, \gamma)$. Because of

$$\|(\alpha_1 \oplus \alpha_2 - c)_+\|_{L^1(X)} \leq \gamma|X|^{\frac{1}{2}}\|\pi\|_{L^2(X)} \leq C\big(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)}\big),$$

for some constant $C = C(\gamma, m) > 0$, we find that the outer sum of the dual variables is bounded by

$$\begin{aligned}
&\|\alpha_1 \oplus \alpha_2\|_{L^1(X)} \\
&\quad \leq \gamma\|\pi\|_{L^1(X)} + \|(\alpha_1 \oplus \alpha_2 - c)_-\|_{L^1(X)} + \|c\|_{L^1(X)} \\
&\quad \leq C \cdot \Big(\sum_{k=1}^{2}\|\mu_1 \otimes \mu_2\|^k_{L^2(X)} + \|c\|^k_{L^2(X)}\Big),
\end{aligned}$$

with some constant $C = C(\gamma, \underline{\mu}, m) > 0$.

With the help of the dual representation of the norm on $L^1(X)$, namely

$$\|u\|_{L^1(X)} = \sup_{\substack{\phi \in L^\infty(X), \\ \|\phi\|_{L^\infty(X)} \leq 1}} \int_X \phi u\ d\lambda \quad \text{for all } u \in L^1(X),$$

we find the following lower L^1-bound for the outer sum of the dual variables:

$$\begin{aligned}
\|\alpha_1 \oplus \alpha_2\|_{L^1(X)} &= \sup_{\substack{\phi \in L^\infty(X), \\ \|\phi\|_{L^\infty(X)} \leq 1}} \int_X \phi(\alpha_1 \oplus \alpha_2)\ d\lambda \\
&\geq \sup_{\substack{\phi_1 \in L^\infty(X_1), \\ \|\phi_1\|_{L^\infty(X_1)} \leq 1}} \int_X (\phi_1 \otimes \mathbb{1})(\alpha_1 \oplus \alpha_2)\ d\lambda
\end{aligned}$$

$$= \sup_{\substack{\phi_1 \in L^\infty(X_1),\\ \|\phi_1\|_{L^\infty(X_1)} \leq 1}} |X_2| \int_{X_1} \phi_1 \alpha_1 \, d\lambda_1 + \int_{X_1} \phi_1 \, d\lambda_1 \int_{X_2} \alpha_2 \, d\lambda_2$$
$$= |X_2| \|\alpha_1\|_{L^1(X_1)}.$$

Note that the last of above's equation holds because α_2 is a zero-mean dual variable. In a similar fashion, we receive that

$$\begin{aligned}\|\alpha_1 \oplus \alpha_2\|_{L^1(X)} &\geq \sup_{\substack{\phi_2 \in L^\infty(X_2),\\ \|\phi_2\|_{L^\infty(X_2)} \leq 1}} \int_{X_1} \alpha_1 \, d\lambda_1 \int_{X_2} \phi_2 \, d\lambda_2 + |X_1| \int_{X_2} \phi_2 \alpha_2 \, d\lambda_2 \\ &\geq -|X_2| \|\alpha_1\|_{L^1(X_1)} + |X_1| \|\alpha_2\|_{L^1(X_2)} \\ &\geq -\|\alpha_1 \oplus \alpha_2\|_{L^1(X)} + |X_1| \|\alpha_2\|_{L^1(X_2)}.\end{aligned}$$

Therefore, for $i = 1, 2$,

$$\begin{aligned}\|\alpha_i\|_{L^1(X_i)} &\leq \frac{2\|\alpha_1 \oplus \alpha_2\|_{L^1(X)}}{\min\{|X_1|, |X_2|\}} \\ &\leq C \cdot \left(\sum_{k=1}^{2} \|\mu_1 \otimes \mu_2\|_{L^2(X)}^k + \|c\|_{L^2(X)}^k \right) \\ &\leq C \cdot \left(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)} + 1 \right)^2,\end{aligned} \tag{3.14}$$

with $C = C(\gamma, \underline{\mu}, m) > 0$, as claimed.

Note that the additional "$+1$" on the right-hand side of (3.14) automatically accounts for the assumption of a zero-mean dual solution, see Remark 3.22. □

The last preparatory step before the final proof of the Hoelders continuity of $\mathcal{S}_\gamma$ is to construct L^2 bounds for the dual variables. This is done in the following lemma.

Lemma 3.24. *Let Assumption 3.21 hold and let $\gamma > 0$ and $(c, \mu_1, \mu_2) \in \mathcal{C}_{\underline{c}}(X) \times \mathcal{M}^m_{\underline{\mu}}(X_1, X_2)$ be arbitrary and consider the corresponding optimal solution of the regularized Kantorovich problem (K_γ), $\pi = \mathcal{S}_\gamma(c, \mu_1, \mu_2)$. Then, the dual variables $\alpha_1 \in L^2(X_1)$ and $\alpha_2 \in L^2(X_2)$ from Theorem 3.9 are bounded by*

$$\|\alpha_1\|_{L^2(X_1)}, \|\alpha_2\|_{L^2(X_2)} \leq C \cdot \left(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)} + 1 \right)^6,$$

with some constant $C = C(\gamma, \underline{c}, \underline{\mu}, m) > 0$.

Proof. We again assume that μ_2 is a zero-mean dual variable and proceed in two steps: in the first step, we show the boundedness of the positive parts of the dual solution (i); in the second step, we derive bounds for their negative parts (ii).

Ad (i): We abbreviate

$$M := C_1 \cdot \left(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)} + 1 \right)^2 > 0, \tag{3.15}$$

with $C_1 > 0$ being the constant from the formulation of Lemma 3.23, and define, up to sets of zero Lebesgue measure, the subset

$$\tilde{X}_2 := \left\{ x_2 \in X_2 \colon |\alpha_2(x_2)| \leq \frac{2M}{|X_2|} \right\} \subset X_2.$$

It follows by construction that

$$M \geq \int_{X_2} |\alpha_2(x_2)| \, dx_2 \geq \int_{X_2 \setminus \tilde{X}_2} |\alpha_2(x_2)| \, dx_2 \geq \frac{2M}{|X_2|} |X_2 \setminus \tilde{X}_2|.$$

This implies

$$\frac{|X_2|}{2} \geq |X_2 \setminus \tilde{X}_2| = |X_2| - |\tilde{X}_2|$$

and, in turn,

$$|\tilde{X}_2| \geq \frac{|X_2|}{2} > 0. \tag{3.16}$$

Now, we define, again up to sets of zero Lebesgue measure, the subsets

$$X_1^+ \coloneqq \{x_1 \in X_1 \colon \alpha_1(x_1) \geq 0\} \subset X_1$$

as well as

$$\tilde{X}^+ \coloneqq \big\{(x_1, x_2) \in X_1^+ \times \tilde{X}_2 \colon \alpha_1(x_1) + \alpha_2(x_2) \geq 0\big\} \subset X_1^+ \times \tilde{X}_2 \subset X.$$

In particular,

$$0 \leq \alpha_1(x_1) < -\alpha_2(x_2) \leq \frac{2M}{|X_2|} \quad \lambda\text{-a.e. in } (X_1^+ \times \tilde{X}_2) \setminus \tilde{X}^+.$$

Therefore,

$$\begin{aligned} |\tilde{X}_2| \int_{X_1} (\alpha_1)_+^2 \, d\lambda_1 &= \int_{\tilde{X}_2} \int_{X_1^+} |\alpha_1|^2 \, d\lambda_1 \, d\lambda_2 \\ &= \int_{\tilde{X}^+} |\alpha_1|^2 \, d\lambda + \int_{(X_1^+ \times \tilde{X}_2) \setminus \tilde{X}^+} |\alpha_1|^2 \, d\lambda \\ &\leq \int_{\tilde{X}^+} |\alpha_1|^2 \, d\lambda + \left(\frac{2M}{|X_2|}\right)^2 |(X_1^+ \times \tilde{X}_2) \setminus \tilde{X}^+|. \end{aligned}$$

Taking advantage of (3.16), this yields

$$\|(\alpha_1)_+\|_{L^2(X_1)}^2 \leq \frac{2}{|X_2|} \int_{\tilde{X}^+} |\alpha_1|^2 \, d\lambda + 8 \left(\frac{M}{|X_2|}\right)^2 |X_1|. \tag{3.17}$$

On the one hand, we observe that

$$\begin{aligned} \|(\alpha_1 \oplus \alpha_2)_+\|_{L^2(X)}^2 &= \int_X (\alpha_1 \oplus \alpha_2)_+^2 \, d\lambda \\ &\leq 2 \left(\int_X (\alpha_1 \oplus \alpha_2 - c)_+^2 \, d\lambda + \int_X c_+^2 \, d\lambda \right) \\ &\leq 2\big(\gamma^2 \|\pi\|_{L^2(X)}^2 + \|c\|_{L^2(X)}^2\big) \\ &\leq C_2\big(\|\mu_1 \otimes \mu_2\|_{L^2(X)}^2 + \|c\|_{L^2(X)}^2\big), \end{aligned} \tag{3.18}$$

for some $C_2 = C(\gamma, m)$, see Lemma 3.19. On the other hand,

$$\begin{aligned}\|(\alpha_1 \oplus \alpha_2)_+\|^2_{L^2(X)} &\geq \int_{\tilde{X}^+} (\alpha_1 \oplus \alpha_2)_+^2 \, d\lambda \\ &= \int_{\tilde{X}^+} (\alpha_1 \oplus \alpha_2)^2 \, d\lambda \\ &\geq \int_{\tilde{X}^+} |\alpha_1|^2 \, d\lambda - 2\int_{\tilde{X}^+} |\alpha_1||\alpha_2| \, d\lambda \\ &\geq \int_{\tilde{X}^+} |\alpha_1|^2 \, d\lambda - 2\,\|\alpha_1\|_{L^1(X_1)}\|\alpha_2\|_{L^1(X_2)}.\end{aligned} \tag{3.19}$$

Combining (3.15) as well as (3.17) – (3.19) with Lemma 3.23, then results in

$$\begin{aligned}&\|(\alpha_1)_+\|_{L^2(X_1)} \\ &\leq \left(\frac{2}{|X_2|}\left(\|(\alpha_1 \oplus \alpha_2)_+\|^2_{L^2(X)} + 2\,\|\alpha_1\|_{L^1(X_1)}\|\alpha_2\|_{L^1(X_2)}\right) + 8M^2\frac{|X_1|}{|X_2|^2}\right)^{\frac{1}{2}} \\ &\leq \left(\frac{2C_2}{|X_2|}\cdot\left(\|\mu_1 \otimes \mu_2\|^2_{L^2(X)} + \|c\|^2_{L^2(X)}\right) + \frac{4|X_2| + 8|X_1|}{|X_2|^2}\cdot M^2\right)^{\frac{1}{2}} \\ &\leq C_3 \cdot \left(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)} + 1\right)^2\end{aligned}$$

for some constant $C_3 = C(\gamma, \underline{\mu}, m) > 0$. A similar L^2-bound for $(\alpha_2)_+$ follows by means of reversed roles.

Ad (ii): Given $r \in \mathbb{R}$, we consider, up to sets of zero Lebesgue measure, the set

$$\hat{X}_2^r := \{x_2 \in X_2 \colon (\alpha_2)_+(x_2) > r + \underline{c}\} \subset X_2.$$

For any $r > -\underline{c}$, the mass of this subset can be estimated by

$$|\hat{X}_2^r| = \frac{1}{r + \underline{c}}\int_{\hat{X}_2^r} r + \underline{c} \, d\lambda_2 \leq \frac{1}{r + \underline{c}}\int_{\hat{X}_2^r} (\alpha_2)_+ \, d\lambda_2 \leq \frac{\|\alpha_2\|_{L^1(X_2)}}{r + \underline{c}} \leq \frac{M}{r + \underline{c}}.$$

Consequently, $|\hat{X}_2^r| \to 0$ as $r \to \infty$. For any $r > -\underline{c}$, we find that

$$\begin{aligned}&\int_{X_2} (-r + \alpha_2(x_2) - \underline{c})_+ \, dx_2 \\ &\quad\leq \int_{X_2} \left(-(r + \underline{c}) + (\alpha_2)_+(x_2)\right)_+ \, dx_2 \\ &\quad= \int_{\hat{X}_2^r} -(r + \underline{c}) + (\alpha_2)_+(x_2) \, dx_2 \\ &\quad\leq \int_{\hat{X}_2^r} (\alpha_2)_+ \, d\lambda_2 \leq |\hat{X}_2^r|^{\frac{1}{2}}\|(\alpha_2)_+\|_{L^2(X_2)} \leq \left(\frac{M}{r + \underline{c}}\right)^{\frac{1}{2}} K,\end{aligned}$$

where $K > 0$ is short for the bound for $\|(\alpha_1)_+\|_{L^2(X_1)}$ and $\|(\alpha_2)_+\|_{L^2(X_2)}$ from step (i). If we define

$$R := \frac{MK^2}{\gamma^2\underline{\mu}^2} + 1 - \underline{c} > -\underline{c},$$

then

$$\int_{X_2} (-R + \alpha_2(x_2) - \underline{c})_+ \, dx_2 < \gamma\underline{\mu}.$$

Now, let us assume that $\alpha_1 \leq -R$ λ_1-a.e. on a set $E \subset X_1$ with $\lambda_1(E) > 0$. Then,

$$\int_{X_2} (\alpha_1 \oplus \alpha_2 - c)_+ \, d\lambda_2 \leq \int_{X_2} (-R + \alpha_2(x_2) - \underline{c})_+ \, dx_2 < \gamma\underline{\mu} \leq \gamma\mu_1$$

λ_1-a.e. on E. This, however, contradicts (3.6b). Hence, $(\alpha_1)_-$ must be bounded essentially by R, i.e., $\|(\alpha_1)_-\|_{L^\infty(X_1)} < R$. Therefore,

$$\begin{aligned}
&\|(\alpha_1)_-\|_{L^2(X_1)} \\
&\quad \leq R|X_1|^{\frac{1}{2}} \\
&\quad = \frac{|X_1|^{\frac{1}{2}}}{\gamma^2\underline{\mu}^2} C_1 C_3 \cdot \left(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)} + 1\right)^6 + |X_1|^{\frac{1}{2}}(1 - \underline{c}) \\
&\quad \leq C_4 \cdot \left(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)} + 1\right)^6,
\end{aligned}$$

with some constant $C_4 = C(\gamma, \underline{c}, \underline{\mu}, m) > 0$ and, consequently,

$$\begin{aligned}
\|\alpha_1\|_{L^2(X_1)} &\leq \|(\alpha_1)_+\|_{L^2(X_1)} + \|(\alpha_1)_-\|_{L^2(X_1)} \\
&\leq C \cdot \left(\|\mu_1 \otimes \mu_2\|_{L^2(X)} + \|c\|_{L^2(X)} + 1\right)^6,
\end{aligned}$$

with some constant $C = C(\gamma, \underline{c}, \underline{\mu}, m) > 0$ as claimed. Again, the estimate for $(\alpha_2)_-$ and thus for α_2 follows by means of reversed roles. □

We are now in a position to establish the Hölder continuity of the solution operator associated with the regularized Kantorovich problem (K_γ). We shall see in Theorem 3.26 that this is essential for proving the existence of solutions to ($\mathrm{BK}_\gamma^\delta$).

But first, for the sake of readability, we define the Hilbert space

$$\mathfrak{H} := L^2(X) \times L^2(X_1) \times L^2(X_2)$$

which carries the norm

$$\|(u, v, w)\|_{\mathfrak{H}} := \left(\|u\|^2_{L^2(X)} + \|v\|^2_{L^2(X_1)} + \|w\|^2_{L^2(X_2)}\right)^{\frac{1}{2}}.$$

Proposition 3.25. *Let the parameters $\gamma, \underline{\mu}, m > 0$ and $\underline{c} > -\infty$ be given. Then, the solution operator of the regularized Kantorovich problem,*

$$\mathcal{S}_\gamma \colon \mathcal{C}_{\underline{c}}(X) \times \mathcal{M}^m_{\underline{\mu}}(X_1, X_2) \to L^2(X), \quad (c, \mu_1, \mu_2) \mapsto \pi,$$

with π being the solution to (K_γ) w.r.t. the cost function c and the marginals μ_1 and μ_2, is Hölder continuous (on bounded sets) with exponent ½, i.e., for each radius $\rho > 0$, we can find a constant $C = C(\underline{c}, \underline{\mu}, m, \gamma, \rho) > 0$ such that

$$\|\mathcal{S}_\gamma(c_\mu, \mu_1, \mu_2) - \mathcal{S}_\gamma(c_\nu, \nu_1, \nu_2)\|_{L^2(X)} \leq C\|(c_\mu, \mu_1, \mu_2) - (c_\nu, \nu_1, \nu_2)\|_{\mathfrak{H}}^{\frac{1}{2}}$$

for all $(c_\mu, \mu_1, \mu_2), (c_\nu, \nu_1, \nu_2) \in \mathcal{C}_{\underline{c}}(X) \times \mathcal{M}^m_{\underline{\mu}}(X_1, X_2)$ with

$$\|(c_\mu, \mu_1, \mu_2)\|_{\mathfrak{H}}, \|(c_\nu, \nu_1, \nu_2)\|_{\mathfrak{H}} < \rho.$$

Proof. Given arbitrary points $(c_\mu, \mu_1, \mu_2), (c_\nu, \nu_1, \nu_2) \in \mathcal{C}_{\underline{c}}(X) \times \mathcal{M}^m_{\underline{\mu}}(X_1, X_2)$ with

$$\|(c_\mu, \mu_1, \mu_2) - (c_\nu, \nu_1, \nu_2)\|_{\mathfrak{H}} < \rho,$$

we set $\pi_\mu := \mathcal{S}_\gamma(c_\mu, \mu_1, \mu_2)$ and $\pi_\nu := \mathcal{S}_\gamma(c_\nu, \nu_1, \nu_2)$. By virtue of Theorem 3.9, there exist $\alpha_1^\mu, \alpha_1^\nu \in L^2(X_1)$ and $\alpha_2^\mu, \alpha_2^\nu \in L^2(X_2)$ such that $\pi_\mu = \frac{1}{\gamma}(\alpha_1^\mu \oplus \alpha_2^\mu - c_\mu)_+$ and $\pi_\nu = \frac{1}{\gamma}(\alpha_1^\nu \oplus \alpha_2^\nu - c_\nu)_+$ satisfy the equations (3.6b) and (3.6c), respectively. Hence,

$$\int_{X_2} \pi_\mu - \pi_\nu \, d\lambda_2 = \mu_1 - \nu_1 \quad \text{and} \quad \int_{X_1} \pi_\mu - \pi_\nu \, d\lambda_1 = \mu_2 - \nu_2. \tag{3.20}$$

Testing the first and second equation in (3.20) with $\alpha_1^\mu - \alpha_1^\nu$ and $\alpha_2^\mu - \alpha_2^\nu$, respectively, integrating and then adding the resulting equations, we arrive at

$$\begin{aligned}&\int_X (\pi_\mu - \pi_\nu)\big((\alpha_1^\mu - \alpha_1^\nu) \oplus (\alpha_2^\mu - \alpha_2^\nu)\big) \, d\lambda \\ &\quad = \int_{X_1} (\mu_1 - \nu_1)(\alpha_1^\mu - \alpha_1^\nu) \, d\lambda_1 + \int_{X_2} (\mu_2 - \nu_2)(\alpha_2^\mu - \alpha_2^\nu) \, d\lambda_2,\end{aligned}$$

which is equivalent to

$$\begin{aligned}&\int_X (\pi_\mu - \pi_\nu)\Big(\frac{1}{\gamma}(\alpha_1^\mu \oplus \alpha_2^\mu - c_\mu) - \frac{1}{\gamma}(\alpha_1^\nu \oplus \alpha_2^\nu - c_\nu)\Big) \, d\lambda \\ &\quad + \frac{1}{\gamma}\int_X (\pi_\mu - \pi_\nu)(c_\mu - c_\nu) \, d\lambda \\ &= \frac{1}{\gamma}\left(\int_{X_1} (\mu_1 - \nu_1)(\alpha_1^\mu - \alpha_1^\nu) \, d\lambda_1 + \int_{X_2} (\mu_2 - \nu_2)(\alpha_2^\mu - \alpha_2^\nu) \, d\lambda_2\right).\end{aligned}$$

Using (3.6a), the inequality $(a_+ - b_+)^2 \le (a_+ - b_+)(a - b)$ for all $a, b \in \mathbb{R}$ and the Cauchy-Schwarz inequality, this implies

$$\begin{aligned}&\|\pi_\mu - \pi_\nu\|^2_{L^2(X)} - \frac{1}{\gamma}\int_X |(\pi_\mu - \pi_\nu)(c_\mu - c_\nu)| \, d\lambda \\ &\le \frac{1}{\gamma}\big(\|\alpha_1^\mu - \alpha_1^\nu\|_{L^2(X_1)}\|\mu_1 - \nu_1\|_{L^2(X_1)} + \|\alpha_2^\mu - \alpha_2^\nu\|_{L^2(X_2)}\|\mu_2 - \nu_2\|_{L^2(X_2)}\big).\end{aligned} \tag{3.21}$$

By Young's inequality, see Lemma D.3,

$$\begin{aligned}\frac{1}{\gamma}\int_X |(\pi_\mu - \pi_\nu)(c_\mu - c_\nu)| \, d\lambda &\le \frac{1}{\gamma}\|\pi_\mu - \pi_\nu\|_{L^2(X)}\|c_\mu - c_\nu\|_{L^2(X)} \\ &\le \frac{1}{2}\|\pi_\mu - \pi_\nu\|^2_{L^2(X)} + \frac{2}{\gamma^2}\|c_\mu - c_\nu\|^2_{L^2(X)}\end{aligned}$$

Inserting this into (3.21), we arrive at

$$\begin{aligned}&\|\pi_\mu - \pi_\nu\|^2_{L^2(X)} \\ &\le \frac{4}{\gamma^2}\|c_\mu - c_\nu\|^2_{L^2(X)} + \frac{2}{\gamma}\sum_{i=1}^{2}\|\alpha_i^\mu - \alpha_i^\nu\|_{L^2(X_i)}\|\mu_i - \nu_i\|_{L^2(X_i)}\end{aligned}$$

$$\leq \max\Big\{\frac{4}{\gamma^2},\frac{2}{\gamma}\Big\}\Big(\|c_\mu - c_\nu\|^2_{L^2(X)} + \sum_{i=1}^{2}\|\alpha_i^\mu - \alpha_i^\nu\|_{L^2(X_i)}\|\mu_i - \nu_i\|_{L^2(X_i)}\Big)$$

or equivalently,

$$\|\pi_\mu - \pi_\nu\|_{L^2(X)} \leq \Big(\hat{C}\max\Big\{\frac{4\sqrt{3}}{\gamma^2},\frac{2\sqrt{3}}{\gamma}\Big\}\Big)^{\frac{1}{2}}\|(c_\mu,\mu_1,\mu_2) - (c_\nu,\nu_1,\nu_2)\|^{\frac{1}{2}}_{\mathfrak{H}}, \tag{3.22}$$

with

$$\hat{C} \coloneqq \max\Big\{\|c_\mu - c_\nu\|_{L^2(X)}, \|\alpha_1^\mu - \alpha_1^\nu\|_{L^2(X_1)}, \|\alpha_2^\mu - \alpha_2^\nu\|_{L^2(X_2)}\Big\} > 0.$$

By assumption,

$$\|c_\mu - c_\nu\|_{L^2(X)} \leq \|(c_\mu,\mu_1,\mu_2)\|_{\mathfrak{H}} + \|(c_\nu,\nu_1,\nu_2)\|_{\mathfrak{H}} < 2\rho.$$

Moreover, Lemma 3.24 provides a constant $C = C(\underline{c},\underline{\mu},m,\gamma) > 0$ such that

$$\begin{aligned}
\|\alpha_1^\mu\|_{L^2(X_i)}, \|\alpha_2^\mu\|_{L^2(X_i)} &\leq C\cdot\big(\|\mu_1\otimes\mu_2\|_{L^2(X)} + \|c_\mu\|_{L^2(X)} + 1\big)^6\\
&\leq C\cdot\big(\|\mu_1\|^2_{L^2(X_1)} + \|\mu_2\|^2_{L^2(X_2)} + \|c_\mu\|_{L^2(X)} + 1\big)^6\\
&\leq C\cdot\big(\rho\|\mu_1\|_{L^2(X_1)} + \rho\|\mu_2\|_{L^2(X_2)} + \|c_\mu\|_{L^2(X)} + 1\big)^6\\
&\leq C\cdot\big(\rho+1\big)^6,
\end{aligned}$$

Of course, we find the same bounds for α_1^ν and α_2^ν. Combining all of the above, shows that $\hat{C}$ can be estimated by

$$\hat{C} \leq C\cdot(\rho+1)^6,$$

where $C > 0$ is a constant solely depending on the radius ρ as well as the fixed parameters γ, $\underline{c}$, $\underline{\mu}$, and m. This together with (3.22) shows that the solution operator is (on bounded sets) Hölder continuous with exponent ½. □

We now have everything together to prove the existence of an optimal solution to the regularized bilevel Kantorovich problem. Therefore, we return to the setting of Subchapter 3.2 and recall the problem statement:

$$\begin{aligned}
\inf_{\pi,\mu_1,c}\quad & \mathcal{J}_\gamma(\pi,\mu_1,c) \coloneqq \mathcal{J}(\pi,\mu_1) + \tfrac{1}{\gamma}\|c - c_{\mathrm{d}}\|^p_{W^{1,p}(\Omega)}\\
\text{s.t.}\quad & c\in W^{1,p}(\Omega),\ \mu_1\in\mathfrak{P}(\Omega_1),\\
& \pi = (\mathcal{E}_\delta^*\circ\mathcal{S}_\gamma)\big(\mathcal{E}_\delta(c),\mathcal{T}_1^\delta(\mu_1),\mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big).
\end{aligned} \tag{$\mathrm{BK}_\gamma^\delta$}$$

Theorem 3.26. *Given the assumptions on the domains, the target marginal, the cost function, and the target functional from Subchapter 3.1 and Subchapter 3.2, for every $\gamma > 0$ and $\delta > 0$, there exists at least one optimal solution to ($\mathrm{BK}_\gamma^\delta$).*

Proof. Again, this proof is based on standard arguments: We show that the feasible set is non-empty and contains a sequence converging to the infimum of ($\mathrm{BK}_\gamma^\delta$) (i), verify the boundedness of that sequence (ii), argue that the limit point of a convergent subsequence is still contained in the feasible set (iii), and

apply the lower semi-continuity of the target functional, to show the optimality of the limit point (iv).

Ad (i): We abbreviate the feasible set of ($\mathrm{BK}_\gamma^\delta$) by

$$\mathcal{F} := \Big\{(\pi, \mu_1, c) \in \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1) \times W^{1,p}(\Omega) \colon \\ \mu_1 \in \mathfrak{P}(\Omega_1),\ \pi = \mathcal{E}_\delta^*(\tilde{\pi}),\ \tilde{\pi} = S_\gamma\big(\mathcal{E}_\delta(c), \mathcal{T}_1^\delta(\mu_1), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big)\Big\}.$$

Analogous to the proof of Theorem 3.5, we choose $\hat{\mu}_1 = \delta_{\hat{x}}$, the Dirac measure at some arbitrary point $\hat{x} \in \Omega_1$. By construction, $\hat{\mu}_1 \in \mathfrak{P}(\Omega_1)$. We choose $\hat{c} \equiv 0$ as a cost function on Ω. Trivially, its extension onto Ω^δ is given by $\mathcal{E}_\delta(\hat{c}) \equiv 0$. We know from Remark 3.14 that $\mathcal{T}_1^\delta(\hat{\mu}_1), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}}) \geq \delta \min\{|\Omega_1^\delta|, |\Omega_2^\delta|\} > 0$ and that $\int_{\Omega_1} \mathcal{T}_1^\delta(\hat{\mu}_1)\, d\lambda_1 = \int_{\Omega_2} \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\, d\lambda_2$. Lemma 3.8 therefore implies that $\tilde{\pi} = \mathcal{S}_\gamma\big(\mathcal{E}_\delta(\hat{c}), \mathcal{T}_1^\delta(\hat{\mu}_1), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big)$ exists and by setting $\hat{\pi} := \mathcal{E}_\delta^*(\pi)$, the triple $(\hat{\pi}, \hat{\mu}_1, \hat{c})$ is an element of $\mathcal{F}$. This shows that the feasible set is non-empty. Consequently, it must contain a minimizing sequence $(\pi_n, \mu_{1,n}, c_n)_{n\in\mathbb{N}}$ with

$$\lim_{n\to\infty} \mathcal{J}_\gamma(\pi_n, \mu_{1,n}, c_n) = \inf_{(\pi,\mu_1,c)\in\mathcal{F}} \mathcal{J}_\gamma(\pi, \mu_1, c) \in \mathbb{R} \cup \{-\infty\}. \tag{3.23}$$

Ad (ii): We now show that the minimizing sequence from step (i) is bounded. First, we notice that $\|\mu_{1,n}\|_{\mathfrak{M}(\Omega_1)} = 1$ for all $n \in \mathbb{N}$.

Also, for each and every $n \in \mathbb{N}$, we can find an optimal solution to the regularized Kantorovich problem, namely $\tilde{\pi}_n = \mathcal{S}_\gamma\big(\mathcal{E}_\delta(c_n), \mathcal{T}_1^\delta(\mu_{1,n}), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big)$, such that $\pi_n = \mathcal{E}_\delta^*(\tilde{\pi}_n)$. Thus,

$$\begin{aligned}
\|\pi_n\|_{\mathfrak{M}(\Omega)} = \mathcal{E}_\delta^*(\tilde{\pi}_n)(\Omega) &= \int_{\Omega_1}\int_{\Omega_2} \tilde{\pi}_n\, d\lambda_2\, d\lambda_1 \\
&\leq \int_{\Omega_1^\delta}\int_{\Omega_2^\delta} \tilde{\pi}_n\, d\lambda_2\, d\lambda_1 \\
&= \int_{\Omega_1^\delta} \varphi_1^\delta * \mu_{1,n} + \delta|\Omega_2^\delta|\, d\lambda_1 \\
&= \big\|\varphi_1^\delta\big\|_{L^1(B_1^\delta)} \|\mu_{1,n}\|_{\mathfrak{M}(\Omega_1)} + \delta|\Omega_1^\delta||\Omega_2^\delta| = 1 + \delta|\Omega_1^\delta||\Omega_2^\delta|,
\end{aligned}$$

where we used the feasibility and nonnegativity of $\tilde{\pi}_n$, Lemma A.4, and that φ_1^δ is a mollifier which supported on B_1^δ.

Owing to the $W^{1,p}$-penalty term in the target functional $\mathcal{J}_\gamma$ and the lower bound of $\mathcal{J}$, there exists some constant $C > 0$ such that $\|c_n\|_{W^{1,p}(\Omega)} \leq C$ for all $n \in \mathbb{N}$ (otherwise, $(\pi_n, \mu_{1,n}, c_n)_{n\in\mathbb{N}}$ cannot be a minimizing sequence).

Ad (iii): The boundedness of the sequence $(\pi_n, \mu_{1,n}, c_n)_{n\in\mathbb{N}}$ implies the existence of a subsequence $(\pi_{n_k}, \mu_{1,n_k}, c_{n_k})_{k\in\mathbb{N}}$ and a cluster point $(\bar{\pi}, \bar{\mu}_1, \bar{c}) \in \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1) \times W^{1,p}(\Omega)$ such that

$$(\pi_{n_k}, \mu_{1,n_k}) \xrightarrow[k\to\infty]{\mathfrak{M}(\Omega)\times\mathfrak{M}(\Omega_1)\searrow^*} (\bar{\pi}, \bar{\mu}_1) \quad \text{and} \quad c_{n_k} \xrightarrow[k\to\infty]{W^{1,p}(\Omega)\searrow} \bar{c}.$$

In particular, $c_{n_k} \to \bar{c}$ in $C(\Omega)$ as $k \to \infty$, where this convergence is understood for a selection of continuous representatives of the equivalence classes c_{n_k} and $\bar{c}$, see e.g. [1, Theorem 6.3 Part III].

By Lemma B.17, there exists, for any $B \in \mathfrak{B}(\Omega_1)$, a sequence of nonnegative functions $(v_m)_{m\in\mathbb{N}}$ with $\int_{\Omega_1} v_m \, d\bar{\mu}_1 \to \bar{\mu}_1(B)$ as $m \to \infty$. Owing to the nonnegativity of both v_m and μ_{1,n_k} for all m and all k, respectively, we find that

$$\bar{\mu}_1(B) = \lim_{m\to\infty} \int_{\Omega_1} v_m \, d\bar{\mu}_1 = \lim_{m\to\infty} \lim_{k\to\infty} \int_{\Omega_1} v_m \, d\mu_{1,n_k} \geq 0,$$

i.e., $\bar{\mu}_1$ is a nonnegative measure. Because of the weak* convergence $\mu_{1,n_k} \rightharpoonup^* \bar{\mu}_1$ it must hold that

$$1 = \|\mu_{1,n_k}\|_{\mathfrak{M}(\Omega_1)} = \langle \mu_{1,n_k}, \mathbb{1} \rangle_{C(\Omega_1)^*, C(\Omega_1)} \;\to\; \langle \bar{\mu}_1, \mathbb{1} \rangle_{C(\Omega_1)^*, C(\Omega_1)} = \|\bar{\mu}_1\|_{\mathfrak{M}(\Omega_1)}$$

as $k \to \infty$ so that $\|\bar{\mu}_1\|_{\mathfrak{M}(\Omega_1)} = 1$. This shows that $\mu_1 \in \mathfrak{P}(\Omega_1)$.

We define $\tilde{\pi} \coloneqq \mathcal{S}_\gamma\big(\mathcal{E}_\delta(\bar{c}), \mathcal{T}_1^\delta(\bar{\mu}_1), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big) \in L^2(\Omega^\delta)$, which exists by reason of the same arguments as in (i), and show in the following that $\bar{\pi} = \mathcal{E}_\delta^*(\tilde{\pi})$. Applying Lemma B.12, we find that

$$\begin{aligned}
&\big\|(\varphi_1^\delta * \mu_{1,n_k} + \delta|\Omega_2^\delta|) - (\varphi_1^\delta * \bar{\mu}_1 + \delta|\Omega_2^\delta|)\big\|^2_{L^2(\Omega_1^\delta)} \\
&\quad = \int_{\Omega_1^\delta} \left| \int_{\Omega_1} \varphi_1^\delta(y - x) \, d(\mu_{1,n_k} - \bar{\mu}_1)(y) \right|^2 dx \\
&\quad = \int_{\Omega_1^\delta} \left| \big\langle \mu_{1,n_k} - \bar{\mu}_1, \varphi_1^\delta \circ T_x \big\rangle_{C(\Omega_1)^*, C(\Omega_1)} \right|^2 dx,
\end{aligned}$$

where $T_x(y) \coloneqq y - x$. For any $x \in \Omega_1^\delta$, the composition $\varphi_1^\delta \circ T_x$ is a continuous function w.r.t. y and, owing to the weak* convergence $\mu_{1,n_k} \rightharpoonup^* \bar{\mu}_1$, the integrand (as a function of x) converges pointwisely to 0. Moreover, it is uniformly bounded by

$$\begin{aligned}
&\big|\big\langle \mu_{1,n_k} - \bar{\mu}_1, \varphi_1^\delta \circ T_x \big\rangle_{C(\Omega_1)^*, C(\Omega_1)}\big|^2 \\
&\quad \leq \|\mu_{1,n_k} - \bar{\mu}_1\|^2_{\mathfrak{M}(\Omega_1)} \big\|\varphi_1^\delta \circ T_x\big\|^2_{C(\Omega_1)} \leq 4\big\|\varphi_1^\delta\big\|^2_{C(\mathbb{R}^{d_1})} < \infty
\end{aligned}$$

for all $x \in \Omega_1^\delta$ and $k \in \mathbb{N}$. Therefore, Lebesgue's dominated convergence theorem implies the convergence of the convoluted marginals:

$$\mathcal{T}_1^\delta(\mu_{1,n_k}) = (\varphi_1^\delta * \mu_{1,n_k} + \delta|\Omega_2^\delta|) \xrightarrow[k\to\infty]{L^2(\Omega_1^\delta)} (\varphi_1^\delta * \bar{\mu}_1 + \delta|\Omega_2^\delta|) = \mathcal{T}_1^\delta(\bar{\mu}_1). \quad (3.24)$$

Also, if we set $m \coloneqq \delta|\Omega_1^\delta||\Omega_2^\delta|$, then $\big(\mathcal{T}_1^\delta(\bar{\mu}_1), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big) \in \mathcal{M}_\delta^m(\Omega_1^\delta, \Omega_2^\delta)$ and $\big(\mathcal{T}_1^\delta(\mu_{1,n_k}), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big) \in \mathcal{M}_\delta^m(\Omega_1^\delta, \Omega_2^\delta)$ for all $k \in \mathbb{N}$. Moreover, because $\mathcal{E}_\delta$ is the trivial extension to Ω^δ, we see that

$$c_{n_k} \xrightarrow[k\to\infty]{C(\Omega)} \bar{c} \implies \mathcal{E}_\delta(c_{n_k}) \xrightarrow[k\to\infty]{L^2(\Omega^\delta)} \mathcal{E}_\delta(\bar{c}) \quad (3.25)$$

and, in particular, $\mathcal{E}_\delta(c_{n_k}) \in \mathcal{C}_{\underline{\mathrm{C}}}(\Omega^\delta)$ for all $k \in \mathbb{N}$. Therefore, we are allowed to apply Proposition 3.25 which, in conjunction with (3.24) and (3.25), ensures that

$$\tilde{\pi}_{n_k} = \mathcal{S}_\gamma\big(\mathcal{E}_\delta(c_{n_k}), \mathcal{T}_1^\delta(\mu_{1,n_k}), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big) \;\to\; \mathcal{S}_\gamma\big(\mathcal{E}_\delta(\bar{c}), \mathcal{T}_1^\delta(\bar{\mu}_1), \mathcal{T}_2^\delta(\mu_2^{\mathrm{d}})\big) = \tilde{\pi}$$

in $L^2(\Omega^\delta)$ as $k \to \infty$. Because of $\pi_{n_k} = \mathcal{E}_\delta^*(\tilde{\pi}_{n_k})$, we find that

$$\begin{aligned}\langle \pi_{n_k}, \phi \rangle_{C(\Omega)^*, C(\Omega)} &= \int_\Omega \phi \, d\pi_{n_k} \\ &= \int_{\Omega^\delta} \mathcal{E}_\delta(\phi)\, \tilde{\pi}_{n_k} \, d\lambda \;\to \int_{\Omega^\delta} \mathcal{E}_\delta(\phi)\, \tilde{\pi} \, d\lambda \\ &= \int_\Omega \phi \, d\mathcal{E}_\delta^*(\tilde{\pi}) = \langle \mathcal{E}_\delta^*(\tilde{\pi}), \phi \rangle_{C(\Omega)^*, C(\Omega)}\end{aligned}$$

for all $\phi \in C(\Omega)$ as $k \to \infty$. The uniqueness of the weak* limit now implies that $\bar{\pi} = \mathcal{E}_\delta^*(\tilde{\pi})$, so that $(\bar{\pi}, \bar{\mu}_1, \bar{c})$ is indeed feasible for $(\mathrm{BK}_\gamma^\delta)$.

Ad (iv): By assumption, $\mathcal{J} \colon \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1) \to \mathbb{R}$ is weak* lower semicontinuous. The norm on $W^{1,p}(\Omega)$ is a convex and continuous functional and thus weakly lower semicontinuous. This is sufficient to conclude that

$$\begin{aligned}\mathcal{J}_\gamma(\bar{\pi}, \bar{\mu}_1, \bar{c}) &= \mathcal{J}(\bar{\pi}, \bar{\mu}_1) + \frac{1}{\gamma} \|\bar{c} - c_\mathrm{d}\|_{W^{1,p}(\Omega)}^p \\ &\le \liminf_{k\to\infty} \mathcal{J}(\pi_{n_k}, \mu_{1,n_k}) \;+\; \liminf_{k\to\infty} \frac{1}{\gamma} \|c_{n_k} - c_\mathrm{d}\|_{W^{1,p}(\Omega)} \\ &\le \liminf_{k\to\infty} \mathcal{J}_\gamma(\pi_{n_k}, \mu_{1,n_k}, c_{n_k}) \\ &= \lim_{k\to\infty} \mathcal{J}_\gamma(\pi_{n_k}, \mu_{1,n_k}, c_{n_k}) = \inf_{(\pi,\mu_1,c)\in\mathcal{F}} \mathcal{J}_\gamma(\pi, \mu_1, c),\end{aligned}$$

where, for the latter two equations, we used (3.23). This shows that $(\bar{\pi}, \bar{\mu}_1, \bar{c})$ is not only feasible but also optimal for $(\mathrm{BK}_\gamma^\delta)$, concluding the proof. □

Remark 3.27. The above proof, in particular step (iii), reveals that the convolution of a marginal with a fixed mollifier defines a compact operator from $\mathfrak{M}(\Omega_1)$ to $L^2(\Omega_1^\delta)$, or more general, a compact operator from $\mathfrak{M}(X)$ to $L^p(\mathbb{R}^d)$ for $X \subset \mathbb{R}^d$ compact and $p \in (1, \infty)$.

This is the crucial ingredient that allows us to ignore the missing weak* continuity of the solution operator of the regularized Kantorovich problem, see Example 3.18, and still obtain the admissibility of the cluster point $(\bar{\pi}, \bar{\mu}_1, \bar{c})$ for the regularized bilevel problem. ○

Now that we have found a positive answer to the well-posedness of the regularized bilevel Kantorovich problem, we will investigate in the next subchapter how we can approximate solutions of the non-regularized bilevel problem (BK) by solutions of the regularized bilevel problem $(\mathrm{BK}_\gamma^\delta)$.

3.4 Approximation of Solutions to the Bilevel Kantorovich Problem

In this subchapter we will show that, given suitably coupled vanishing sequences of regularization and smoothing parameters, we can find cluster points of the sequence of solutions to the corresponding regularized bilevel Kantorovich problems that are solutions to the non-regularized bilevel Kantorovich problem. In other words, we can use optimal solutions of $(\mathrm{BK}_\gamma^\delta)$ to approximate optimal solutions of (BK).

To this end, assume that we are given sequences of nonnegative regularization and smoothing parameters[3] $(\gamma_n)_{n\in\mathbb{N}} \subset \mathbb{R}_{>0}$ and $(\delta_n)_{n\in\mathbb{N}} \subset \mathbb{R}_{>0}$, respectively, that satisfy $\gamma_n, \delta_n \searrow 0$ as well as

$$0 < \delta_n \leq 1 \quad \text{for all } n \in \mathbb{N} \quad \text{and} \quad \frac{\gamma_n}{\delta_n^d} \to 0 \text{ as } n \to \infty. \tag{3.26}$$

For all $n \in \mathbb{N}$, Theorem 3.26 ensures the existence of a solution $(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n)$ to $(\mathrm{BK}_{\gamma_n}^{\delta_n})$. This defines a sequence $(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n)_{n\in\mathbb{N}}$ of regularized bilevel solutions, which will be the subject of our upcoming analysis.

Remark 3.28.

1. The above defined sequence of regularized bilevel solutions need not be unique as there may exist multiple solutions to each regularized bilevel problem.

2. To simplify the notation, from now on we will equip all entities and variables that depend on either γ_n or δ_n (or both) only with the identifier n. We write $\mathcal{S}_n$ instead of $\mathcal{S}_{\gamma_n}$, Ω_1^n instead of $\Omega_1^{\delta_n}$, (BK_n) instead of $(\mathrm{BK}_{\gamma_n}^{\delta_n})$, etc.

○

On the one hand, owing to the feasibility of $(\bar{\pi}_n, \bar{\mu}_{1,n})$ for (BK_n), we find that $\|\bar{\mu}_{1,n}\|_{\mathfrak{M}(\Omega_1)} = 1$ and

$$\begin{aligned} \|\bar{\pi}_n\|_{\mathfrak{M}(\Omega)} = \int_\Omega \tilde{\pi}_n \, d\lambda &\leq \int_{\Omega_1^n} \int_{\Omega_2^n} \tilde{\pi}_n \, d\lambda_2 \, d\lambda_1 = \int_{\Omega_1^n} \mathcal{T}_1^n(\mu_{1,n}) \, d\lambda_1 \\ &= \|\varphi_{1,n}\|_{L^1(B_1^n)} \|\bar{\mu}_{1,n}\|_{\mathfrak{M}(\Omega_1)} + \delta_n |\Omega_1^n| |\Omega_2^n| \\ &\leq 1 + |\Omega_1 + \overline{B(0;1)}| \, |\Omega_2 + \overline{B(0;1)}| < \infty \end{aligned} \tag{3.27}$$

for all $n \in \mathbb{N}$, where $\tilde{\pi}_n$ again denotes the nonnegative solution to the regularized Kantorovich problem (K_n) corresponding to $\bar{\pi}_n$ via $\bar{\pi}_n = \mathcal{E}_n^*(\tilde{\pi}_n)$. We thus can extract a subsequence (which we denote by the same symbol) and find a cluster point $(\bar{\pi}, \bar{\mu}_1) \in \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1)$ such that

$$(\bar{\pi}_n, \bar{\mu}_{1,n}) \xrightarrow[n\to\infty]{}{}^* (\bar{\pi}, \bar{\mu}_1) \quad \text{in } \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1).$$

On the other hand, for arbitrary but fixed $\mu_1 \in \mathfrak{P}(\Omega_1)$, we consider, for $n \in \mathbb{N}$, the regularized optimal transport plans

$$\pi_n = \big(\mathcal{E}_n \circ \mathcal{S}_n\big)\big(\mathcal{E}_n(c_\mathrm{d}), \mathcal{T}_1^n(\mu_1), \mathcal{T}_2^n(\mu_2^\mathrm{d})\big).$$

Then, the triple $(\pi_n, \mu_1, c_\mathrm{d})$ is feasible for (BK_n) and the sequence $(\pi_n, \mu_1)_{n\in\mathbb{N}}$ is bounded in $\mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1)$, see (3.27). Due to the optimality of $(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n)$ for (BK_n),

$$\begin{aligned} \mathcal{J}(\bar{\pi}_n, \bar{\mu}_{1,n}) + \frac{1}{\gamma_n} \|\bar{c}_n - c_\mathrm{d}\|_{W^{1,p}(\Omega)}^p &= \mathcal{J}_n(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n) \\ &\leq \mathcal{J}_n(\pi_n, \mu_1, c_\mathrm{d}) = \mathcal{J}(\pi_n, \mu_1) \end{aligned}$$

[3] For instance, one could choose $\gamma_n = n^{-2}$ and $\delta_n = n^{-1/d}$ for all $n \in \mathbb{N}$. These sequences satisfy all of the requirements.

and therefore after rearranging

$$\|\bar{c}_n - c_{\mathrm{d}}\|_{W^{1,p}(\Omega)} \leq \gamma_n^{\frac{1}{p}} \big(\mathcal{J}(\pi_n, \mu_1) - \mathcal{J}(\bar{\pi}_n, \bar{\mu}_{1,n})\big)^{\frac{1}{p}} \leq \gamma_n^{\frac{1}{p}} C$$

for all $n \in \mathbb{N}$ with some constant $C > 0$, because $\mathcal{J}$ is bounded on bounded sets, see (3.4). Consequently, because $\gamma_n \to 0$ as $n \to \infty$,

$$c_n \xrightarrow[n\to\infty]{} c_{\mathrm{d}} \quad \text{in } W^{1,p}(\Omega).$$

Having found the cluster point $(\bar{\pi}, \bar{\mu}_1, c_{\mathrm{d}})$ of the sequence of regularized solutions $(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n)_{n\in\mathbb{N}}$, we are going to show that $(\bar{\pi}, \bar{\mu}_1)$ is a solution the non-regularized bilevel Kantorovich problem (BK). As one would expect, we proceed in two steps:

1. Show that $(\bar{\pi}, \bar{\mu}_1)$ is feasible for (BK). In particular, this requires to show that $\bar{\pi}$ is not only feasible but also optimal for the non-regularized Kantorovich problem (K). This will be proven in Lemmas 3.30 – 3.33 and requires the technical Assumption 3.29.

2. Show that $(\bar{\pi}, \bar{\mu}_1)$ realizes the optimal value of (BK). This, however, requires the existence of a so-called recovery sequence, see Theorem 3.34.

We begin with the first point and the already mentioned technical assumption.

Assumption 3.29. We assume that there is some $\Delta > 0$ such that $\operatorname{supp}(\bar{\mu}_{1,n}) + B(0;\Delta) \subset \Omega_1$ for all $n \in \mathbb{N}$ and $\operatorname{supp}(\mu_2^{\mathrm{d}}) + B(0;\Delta) \subset \Omega_2$, i.e., the marginals that occur either as solutions of the regularized bilevel problems (BK_n) or as the fixed target marginal of (BK), are supported with a strictly positive distance from the boundary of their corresponding domains.

That the above assumption is not very restrictive is discussed in Subchapter 3.4.1 below. We need it straight away for the proof of the next lemma, which provides us with the feasibility of the cluster point $(\bar{\pi}, \bar{\mu}_1)$ for the non-regularized Kantorovich problem.

Lemma 3.30. *Let $(\pi_n, \mu_{1,n}, c_n)_{n\in\mathbb{N}} \subset \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1) \times W^{1,p}(\Omega)$ be a sequence of feasible points for the sequence of regularized bilevel problems $(\mathrm{BK}_n)_{n\in\mathbb{N}}$ that satisfies $\operatorname{supp}(\mu_{1,n}) + B(0;\Delta) \subset \Omega_1$ for all $n \in \mathbb{N}$. If $(\pi, \mu_1) \in \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1)$ is a weak* cluster point of the sequence $(\pi_n, \mu_{1,n})_{n\in\mathbb{N}}$, then π is a nonnegative coupling between μ_1 and μ_2^{d}, i.e., $\pi \geq 0$ and $P_{1\#}\pi = \mu_1$ as well as $P_{2\#}\pi = \mu_2^{\mathrm{d}}$.*

Proof. Recall that, for each $n \in \mathbb{N}$, there is an optimal solution to (K_n) with respect to $\mathcal{T}_1^n(\mu_{1,n})$, $\mathcal{T}_2^n(\mu_2^{\mathrm{d}})$, and $\mathcal{E}(c_n)$, namely $\tilde{\pi}$, such that $\pi_n = \mathcal{E}_n^*(\tilde{\pi}_n)$. The nonnegativity of π can be shown with the same arguments as the nonnegativity of $\bar{\mu}_1$ in step (iii) of the proof of Theorem 3.26. We therefore consider this proven and only verify the linear constraints of (K).

To this end, let $\phi_1 \in C(\Omega_1)$ be arbitrary and consider an extension of ϕ_1 to $\Omega_1 + \overline{B(0;1)}$, which we denote by $\mathcal{E}(\phi_1)$, that is continuous and is bounded by the same constant $C > 0$ as ϕ_1. This extension exists due to Tietze's extension theorem. Owing to (3.26),

$$\Omega_1^n = \Omega_1 + \overline{B(0;\delta_n)} \subset \Omega_1 + \overline{B(0;1)}$$

and therefore $\sup_{x\in\Omega_1^n}|\mathcal{E}(\phi_1)(x)|\leq C$ for all $n\in\mathbb{N}$. On the one hand, we find that

$$\begin{aligned}\int_{\Omega^n}\mathcal{E}(\phi_1)\tilde{\pi}_n\,d\lambda &= \int_{\Omega}\phi_1\tilde{\pi}_n\,d\lambda+\int_{\Omega^n\setminus\Omega}\mathcal{E}(\phi_1)\tilde{\pi}_n\,d\lambda\\ &= \langle\pi_n,\phi_1\circ P_1\rangle_{C(\Omega)^*,C(\Omega)}+r_n\end{aligned}$$

with $r_n\coloneqq\int_{\Omega^n\setminus\Omega}\mathcal{E}(\phi_1)\tilde{\pi}_n\,d\lambda$. Because of $(\Omega_1^n\times\Omega_2^n)\setminus(\Omega_1\times\Omega_2)=\big((\Omega_1^n\setminus\Omega_1)\times\Omega_2^n\big)\cup\big(\Omega_1^n\times(\Omega_2^n\setminus\Omega_2)\big)$[4] and the feasibility of $\tilde{\pi}_n$ for (K_n) w.r.t. $\mathcal{T}_1^n(\mu_{1,n})=\varphi_{1,n}*\mu_{1,n}+\delta_n|\Omega_2^n|$ and $\mathcal{T}_2^n(\mu_2^{\mathrm{d}})=\varphi_{2,n}*\mu_2^{\mathrm{d}}+\delta_n|\Omega_1^n|$,

$$\begin{aligned}|r_n| &\leq \int_{\Omega^n\setminus\Omega}|\mathcal{E}(\phi_1)|\tilde{\pi}_n\,d\lambda\\ &\leq \int_{(\Omega_1^n\setminus\Omega_1)\times\Omega_2^n}|\mathcal{E}(\phi_1)|\tilde{\pi}_n\,d\lambda+\int_{\Omega_1^n\times(\Omega_2^n\setminus\Omega_2)}|\mathcal{E}(\phi_1)|\tilde{\pi}_n\,d\lambda\\ &\leq C\bigg(\int_{(\Omega_1^n\setminus\Omega_1)}\int_{\Omega_2^n}\tilde{\pi}_n\,d\lambda_2\,d\lambda_1+\int_{(\Omega_2^n\setminus\Omega_2)}\int_{\Omega_1^n}\tilde{\pi}_n\,d\lambda_1\,d\lambda_2\bigg)\\ &= C\bigg(\int_{(\Omega_1^n\setminus\Omega_1)}(\varphi_{1,n}*\mu_{1,n}+\delta_n|\Omega_2^n|)\,d\lambda_1\\ &\qquad+\int_{(\Omega_2^n\setminus\Omega_2)}(\varphi_{2,n}*\mu_2^{\mathrm{d}}+\delta_n|\Omega_1^n|)\,d\lambda_2\bigg)\end{aligned}$$

for all $n\in\mathbb{N}$. Assumption 3.29 together with Theorem A.3, guarantees that $\operatorname{supp}(\varphi_{1,n}*\mu_{1,n})\subset\Omega_1$ and $\operatorname{supp}(\varphi_{2,n}*\mu_2^{\mathrm{d}})\subset\Omega_2$ for all n sufficiently large. Hence, $r_n\to 0$ as $n\to\infty$. Taking advantage of the weak* convergence $\pi_n\rightharpoonup^*\pi$ in $\mathfrak{M}(\Omega)$ and applying the transformation formula for push-forward measures, see e.g. [9, Theorem 3.6.1.], we obtain that

$$\int_{\Omega^n}\mathcal{E}(\phi_1)\tilde{\pi}_n\,d\lambda\;\to\;\langle\pi,\phi_1\circ P_1\rangle_{C(\Omega)^*,C(\Omega)}=\big\langle P_{1\#}\pi,\phi_1\big\rangle_{C(\Omega_1)^*,C(\Omega_1)}\tag{3.28}$$

as $n\to\infty$. On the other hand,

$$\begin{aligned}&\int_{\Omega^n}\mathcal{E}(\phi_1)\tilde{\pi}_n\,d\lambda\\ &\quad=\int_{\Omega_1^n}\mathcal{E}(\phi_1)(\varphi_{1,n}*\mu_{1,n})\,d\lambda_1+\delta_n|\Omega_2^n|\int_{\Omega_1^n}\mathcal{E}(\phi_1)\,d\lambda_1\to\langle\mu_1,\phi\rangle_{C(\Omega_1)^*,C(\Omega_1)},\end{aligned}\tag{3.29}$$

where we used Lemma A.5, the boundedness of $\mathcal{E}(\phi_1)$, and $\delta_n\to 0$ as $n\to\infty$. Comparing (3.28) with (3.29), we receive that

$$\big\langle P_{1\#}\pi,\phi_1\big\rangle_{C(\Omega_1)^*,C(\Omega_1)}=\langle\mu_1,\phi_1\rangle_{C(\Omega_1)^*,C(\Omega_1)}\quad\text{for all }\phi_1\in C(\Omega_1),$$

i.e., $P_{1\#}\pi=\mu_1$. An analogous argument for arbitrary $\phi_2\in C(\Omega_2)$ yields that $P_{2\#}\pi=\mu_2^{\mathrm{d}}$. □

[4] Caution: This decomposition is not disjoint!

We now come to an important approximation result which eventually guarantees that the weak* cluster point $\bar{\pi}$ is optimal for the Kantorovich problem (K), which in combination with lemma 3.30 corresponds to its feasibility for the bilevel problem (BK).

Its proof is based on the gluing lemma for measures and the equivalence of convergence in the Wasserstein 1-metric and weak* convergence of measures on compact sets.

Lemma 3.31. *Let $\pi \in \Pi(\mu_1, \mu_2)$ be a nonnegative coupling between the nonnegative marginals $\mu_1 \in \mathfrak{M}(\Omega_1)$ and $\mu_2 \in \mathfrak{M}(\Omega_2)$ and let $(\mu_{1,n})_{n\in\mathbb{N}} \subset \mathfrak{M}(\Omega_1)$ be a sequence of marginals such that $\mu_{1,n} \rightharpoonup^* \mu_1$ as $n \to \infty$. Then there exists a sequence of nonnegative couplings $(\pi_n)_{n\in\mathbb{N}} \subset \Pi(\mu_{1,n}, \mu_2)$ with $\pi_n \rightharpoonup^* \pi$.*

Proof. For each $n \in \mathbb{N}$, there exists an optimal transport plan $\theta_n \in \Pi(\mu_{1,n}, \mu_1)$ between $\mu_{1,n}$ and μ_1 with respect to the metric cost $\|x_1 - y_1\|$ on Ω_1. Following [76, Lemma 7.6], there exists a nonnegative measure $\sigma_n \in \mathfrak{M}(\Omega_1 \times \Omega_1 \times \Omega_2)$ such that $P_{12\#}\sigma_n = \theta_n$ and $P_{23\#}\sigma_n = \pi$.

In the above and for the rest of the proof the mapping $P_i \colon (x_1, x_2) \mapsto x_i$, $i = 1, 2$, refers to the projection of the tuple (x_1, x_2) to its i-th coordinate[5] and

$$P_{jk} \colon \Omega_1 \times \Omega_1 \times \Omega_2 \to \Omega_1 \times \Omega_l, \quad j, k = 1, 2, 3,\ j < k,\ l = k - 1,$$

refers to the projection onto the coordinates j and k.

Let us define

$$\pi_n \coloneqq P_{13\#}\sigma_n \in \mathfrak{M}(\Omega_1 \times \Omega_2).$$

By construction, for all $B_1 \in \mathfrak{B}(\Omega_1)$,

$$\begin{aligned}\big(P_{1\#}\pi_n\big)(B_1) &= \sigma_n\big(P_{13}^{-1}\big(P_1^{-1}(B_1)\big)\big)\\ &= \sigma_n(B_1 \times \Omega_1 \times \Omega_2)\\ &= \sigma_n\big(P_{12}^{-1}\big(P_1^{-1}(B_1)\big)\big) = \big(P_{1\#}\theta_n\big)(B_1) = \mu_{1,n}(B_1)\end{aligned}$$

and analogously, for all $B_2 \in \mathfrak{B}(\Omega_2)$,

$$\big(P_{2\#}\pi_n\big)(B_2) = \sigma_n(\Omega_1 \times \Omega_1 \times B_2) = \big(P_{2\#}\pi\big)(B_2) = \mu_2(B_2),$$

which yields that $\pi_n \in \Pi(\mu_{1,n}, \mu_2)$ as desired. The nonnegativity of π_n directly follows from the nonnegativity of σ_n.

The next argument, which we borrow from the proof of [10, Theorem 3.1], shows the weak* convergence of the sequence $(\pi_n)_{n\in\mathbb{N}}$ towards π. We consider the mapping

$$P_{1323} \colon \Omega_1 \times \Omega_1 \times \Omega_2 \to \Omega \times \Omega, \quad (x_1, y_1, x_2) \mapsto \big((x_1, x_2), (y_1, x_2)\big),$$

and define $\zeta \coloneqq P_{1323\#}\sigma_n$. We observe that $\zeta \in \mathfrak{M}(\Omega \times \Omega)$ and

$$\big(P_{1\#}\zeta\big)(B) = \zeta(B \times \Omega) = \sigma_n\big(P_{1323}^{-1}(B \times \Omega)\big) = \sigma_n\big(P_{13}^{-1}(B)\big) = \pi_n(B)$$

as well as

$$\big(P_{2\#}\zeta\big)(B) = \zeta(\Omega \times B) = \sigma_n\big(P_{1323}^{-1}(\Omega \times B)\big) = \sigma_n\big(P_{23}^{-1}(B)\big) = \pi(B)$$

[5] Here and in contrast to the projection map from Definition 3.1, the projection P_i will have the domains $\Omega_1 \times \Omega_1$, $\Omega_1 \times \Omega_2$, and $\Omega \times \Omega$. To ease notation, we denote it in all three cases by the same symbol.

for all $B \in \mathfrak{B}(\Omega)$ so that $\zeta \in \Pi(\pi_n, \pi)$. Again, the nonnegativity ζ directly follows from the nonnegativity of σ_n. We then estimate that

$$\begin{aligned}
0 \le W_1(\pi_n, \pi) &= \inf_{0 \le \theta \in \Pi(\pi_n, \pi)} \int_{\Omega \times \Omega} \|x - y\| \, d\theta(x, y) \\
&\le \int_{\Omega \times \Omega} \|x - y\| \, d\zeta(x, y) \\
&\le C \int_{\Omega \times \Omega} \|x_1 - y_1\| + \|x_2 - y_2\| \, d(P_{1323\#}\sigma_n)\big((x_1, x_2), (y_1, y_2)\big) \\
&= C \int_{\Omega_1 \times \Omega_1 \times \Omega_2} \|x_1 - y_1\| \, d\sigma_n(x_1, y_1, x_2) \\
&= C \int_{\Omega_1 \times \Omega_1} \|x_1 - y_1\| \, d(P_{12\#}\sigma_n)(x_1, y_1) \\
&= C \int_{\Omega_1 \times \Omega_1} \|x_1 - y_1\| \, d\theta_n = C \cdot W_1(\mu_{1,n}, \mu_1),
\end{aligned}$$

with some $C > 0$ that only depends on d_1 and d_2. Because of the weak* convergence $\mu_{1,n} \overset{*}{\rightharpoonup} \mu_1$, we find that

$$0 \xleftarrow[n\to\infty]{} W_1(\pi_n, \pi) \le C \cdot W_1(\mu_{1,n}, \mu_1) \xrightarrow[n\to\infty]{} 0$$

as $n \to \infty$ and therefore

$$\pi_n \xrightarrow[n\to\infty]{*} \pi \quad \text{in } \mathfrak{M}(\Omega),$$

see e.g. [75, Theorem 6.9]. □

A quick calculation shows, that a suitably chosen smoothing of transport plans preserves the linear constraints of (K_γ):

Lemma 3.32. *Let $\mu_1 \in \mathfrak{M}(\Omega_1)$, $\mu_2 \in \mathfrak{M}(\Omega_2)$, and $\pi \in \mathfrak{M}(\Omega)$ be such that $P_{i\#}\pi = \mu_i$ for $i = 1, 2$. If we define, for $\delta > 0$, the mollifier $\varphi^\delta := \varphi_1^\delta \otimes \varphi_2^\delta$, then*

$$\int_{\Omega_2^\delta} \varphi^\delta * \pi \, d\lambda_2 = \varphi_1^\delta * \mu_1 \quad \text{and} \quad \int_{\Omega_1^\delta} \varphi^\delta * \pi \, d\lambda_1 = \varphi_2^\delta * \mu_2.$$

Proof. We will only check the first equation. The second equation then follows analogously.

Let us begin by recalling that $\operatorname{supp}(\varphi_2^\delta) \subset B_2^\delta$, $\Omega_2^\delta = \Omega_2 + B_2^\delta$, and that $\int_{B_2^\delta} \varphi_2^\delta \, d\lambda_2 = 1$. Hence, the definition of the convolution of $\varphi_1^\delta \otimes \varphi_2^\delta$ with π together Fubini's theorem yields that

$$\begin{aligned}
\int_{\Omega_2^\delta} (\varphi^\delta * \pi)(x_1, x_2) \, dx_2 &= \int_{\Omega_2^\delta} \int_\Omega \varphi_1^\delta(x_1 - y_1) \varphi_2^\delta(x_2 - y_2) \, d\pi(y_1, y_2) \, dx_2 \\
&= \int_\Omega \varphi_1^\delta(x_1 - y_1) \int_{\Omega_2^\delta} \varphi_2^\delta(x_2 - y_2) \, dx_2 \, d\pi(y_1, y_2) \\
&= \int_\Omega \varphi_1^\delta(x_1 - y_1) \, d\pi(y_1, y_2) \int_{B_2^\delta} \varphi_2^\delta(x_2) \, dx_2 \\
&= \int_\Omega \varphi_1^\delta(x_1 - P_1(y_1, y_2)) \, d\pi(y_1, y_2)
\end{aligned}$$

$$= \int_{\Omega_1} \varphi_1^\delta(x_1 - y_1)\, d(P_{1\#}\pi)(y_1) = (\varphi_1^\delta * \mu_1)(x_1)$$

for all $x_1 \in \Omega_1^\delta$. □

We are now able to prove the feasibility of the cluster point $(\bar{\pi}, \bar{\mu}_1)$ of the sequence of regularized solutions for the non-regularized bilevel problem.

Lemma 3.33. *Let $(\pi_n, \mu_{1,n}, c_n)_{n\in\mathbb{N}} \subset \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1) \times W^{1,p}(\Omega)$ be a sequence of feasible points for the sequence of regularized bilevel problems $(\mathrm{BK}_n)_{n\in\mathbb{N}}$ which satisfies $\operatorname{supp}(\mu_{1,n}) + B(0;\Delta) \subset \Omega_1$ for all $n \in \mathbb{N}$. If $(\pi, \mu_1) \in \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1)$ is a weak* cluster point of the sequence $(\pi_n, \mu_{1,n})_{n\in\mathbb{N}}$ and $c_n \to c_\mathrm{d}$ in $W^{1,p}(\Omega)$ as $n \to \infty$, then (π, μ_1) is feasible for (BK), i.e., $\mu_1 \in \mathfrak{P}(\Omega_1)$ and π is optimal for (K) with respect to the marginals μ_1 and μ_2^d as well as the cost function c_d.*

Proof. The properties of μ_1 can be shown with the same arguments as in the proof of Theorem 3.5. Therefore, we consider this done.

As we have already seen in Lemma 3.30, π is feasible for (K) with respect to μ_1 and μ_2^d. Thus, it suffices to show its optimality w.r.t. c_d. Let us recall the target functionals of (K) and (K_n), namely

$$\mathcal{K}_c(\pi) = \langle \pi, c \rangle_{C(\Omega)^*, C(\Omega)} \quad \text{and} \quad \mathcal{K}_c^n(\pi) = (c, \pi)_{L^2(\Omega^n)} + \frac{\gamma_n}{2}\|\pi\|^2_{L^2(\Omega^n)},$$

respectively. We observe that

1. $\mathcal{K}_{c_n}(\pi_n) \to \mathcal{K}_{c_\mathrm{d}}(\pi)$, since $c_n \to c_\mathrm{d}$ in $C(\Omega)$ and $\pi_n \rightharpoonup^* \pi$ in $\mathfrak{M}(\Omega)$[6];
2. $\mathcal{K}_{c_n}(\pi_n) = (c_n, \tilde{\pi}_n)_{L^2(\Omega)}$ for all $n \in \mathbb{N}$, since π_n has the density $\tilde{\pi}_n$ on Ω.

Given μ_1, μ_2^d, and c_d, let π^* be an arbitrary optimal solution to (K), which exists because of Lemma 3.3. Owing to Lemma 3.31, there exists a sequence of nonnegative couplings $(\pi_n^*)_{n\in\mathbb{N}}$ with $\pi_n^* \in \Pi(\mu_{1,n}, \mu_2^\mathrm{d})$ for all $n \in \mathbb{N}$ that converges weakly* towards π^*. With the mollifier φ_n, which we introduced in Lemma 3.32, we define

$$\tilde{\pi}_n^* \coloneqq \varphi_n * \pi_n^* + \delta_n > 0$$

to receive that

$$\int_{\Omega_2^n} \tilde{\pi}_n^*\, d\lambda_2 = \varphi_{1,n} * \mu_{1,n} + \delta_n |\Omega_2^n| = \mathcal{T}_1^n(\mu_{1,n})$$

and

$$\int_{\Omega_1^n} \tilde{\pi}_n^*\, d\lambda_1 = \varphi_{2,n} * \mu_2^\mathrm{d} + \delta_n |\Omega_1^n| = \mathcal{T}_2^n(\mu_2^\mathrm{d}).$$

Hence, $\tilde{\pi}_n^*$ is feasible for the regularized Kantorovich problem (K_n) with respect to the marginals $\mathcal{T}_1^n(\mu_{1,n})$ and $\mathcal{T}_2^n(\mu_2^\mathrm{d})$.

Let us recall that $\mathcal{E}_n(c_n)$ is the trivial extension of c_n to Ω^n. Consequently,

$$\begin{aligned}\mathcal{K}_{c_n}(\pi_n) &= (c_n, \tilde{\pi}_n)_{L^2(\Omega)} \\ &\le (\mathcal{E}_n(c_n), \tilde{\pi}_n)_{L^2(\Omega^n)} + \frac{\gamma_n}{2}\|\tilde{\pi}_n\|^2_{L^2(\Omega^n)} = \mathcal{K}^n_{\mathcal{E}_n(c_n)}(\tilde{\pi}_n)\end{aligned}$$

[6] This comes straight from the properties of the dual product, see the proof of Lemma A.5.

for all $n \in \mathbb{N}$. Combining all of the above with the optimality of π^* and $\tilde{\pi}_n$ for (K) and (K_n), respectively, we receive that

$$\mathcal{K}_{c_\mathrm{d}}(\pi^*) \leq \mathcal{K}_{c_\mathrm{d}}(\pi) = \lim_{n\to\infty} \mathcal{K}_{c_n}(\pi_n) \leq \liminf_{n\to\infty} \mathcal{K}^n_{\mathcal{E}_n(c_n)}(\tilde{\pi}_n) \leq \liminf_{n\to\infty} \mathcal{K}^n_{\mathcal{E}_n(c_n)}(\tilde{\pi}^*_n). \tag{3.30}$$

It remains to show that

(i) $\lim_{n\to\infty} (\mathcal{E}_n(c_n), \tilde{\pi}^*_n)_{L^2(\Omega^n)} = \langle \pi^*, c_\mathrm{d} \rangle_{C(\Omega)^*, C(\Omega)} = \mathcal{K}_{c_\mathrm{d}}(\pi^*)$,

(ii) $\lim_{n\to\infty} \frac{\gamma_n}{2} \|\tilde{\pi}^*_n\|^2_{L^2(\Omega^n)} = 0$.

Inserting (i) and (ii) into (3.30) reveals that

$$\begin{aligned}\mathcal{K}_{c_\mathrm{d}}(\pi^*) \leq \mathcal{K}_{c_\mathrm{d}}(\pi) &\leq \liminf_{n\to\infty} \mathcal{K}^n_{\mathcal{E}_n(c_n)}(\tilde{\pi}^*_n) \\ &= \lim_{n\to\infty} \Big((\mathcal{E}_n(c_n), \tilde{\pi}^*_n)_{L^2(\Omega^n)} + \frac{\gamma_n}{2} \|\tilde{\pi}^*_n\|^2_{L^2(\Omega^n)} \Big) = \mathcal{K}_{c_\mathrm{d}}(\pi^*)\end{aligned}$$

and therefore $\mathcal{K}_{c_\mathrm{d}}(\pi) = \mathcal{K}_{c_\mathrm{d}}(\pi^*)$, which yields the proposed optimality of π for (K).

Ad (i): By assumption and Lemma C.1, for any $n \in \mathbb{N}$, the support of the coupling π^*_n (transporting $\mu_{1,n}$ onto μ^d_2) has a strictly positive distance to the boundary of Ω. Consequently, there exists a compact subset $K \subset \Omega$ with strictly positive distance to the boundary of Ω and $\operatorname{supp}(\pi^*_n) \subset K$. The definition of $\tilde{\pi}^*_n$ in conjunction with Fubini's theorem then yields that

$$\begin{aligned}(\mathcal{E}_n(c_n), \tilde{\pi}^*_n)_{L^2(\Omega^n)} &= \int_{\Omega^n} \mathcal{E}_n(c_n)(\varphi_n * \pi^*_n)\, d\lambda + \delta_n \int_{\Omega^n} \mathcal{E}_n(c_n)\, d\lambda \\ &= \int_\Omega \int_{\Omega^n} \mathcal{E}_n(c_n)(x)\varphi_n(x-y)\, dx\, d\pi^*_n(y) + r_n \\ &= \int_K \int_{\Omega \cap \overline{B(y;\delta_n)}} c_n(x)\varphi_n(x-y)\, dx\, d\pi^*_n(y) + r_n,\end{aligned} \tag{3.31}$$

with the remainder term

$$r_n := \delta_n \int_{\Omega^n} \mathcal{E}_n(c_n)\, d\lambda = \delta_n \int_\Omega c_n\, d\lambda \;\to\; 0 \cdot \int_\Omega c_\mathrm{d}\, d\lambda = 0, \tag{3.32}$$

owing to $\delta_n \searrow 0$ and the uniform convergence $c_n \to c_\mathrm{d}$ for $n \to \infty$. Choosing n large enough, we find that

$$\operatorname{supp}\big(\varphi_n(\,\cdot - y)\big) \subset \overline{B(y;\delta_n)} \subset \Omega \quad \text{for all } y \in K.$$

Using this and $\int_{\overline{B(y;\delta_n)}} \varphi_n(x-y)\, dx = 1$ yields that

$$\begin{aligned}&\sup_{y\in K} \left| \int_{\Omega \cap \overline{B(y;\delta_n)}} c_n(x)\varphi_n(x-y)\, dx - c_\mathrm{d}(y) \right| \\ &\quad\leq \sup_{y\in K} \int_{\overline{B(y;\delta_n)}} |c_n(x) - c_\mathrm{d}(y)| \varphi_n(x-y)\, dx \\ &\quad\leq \sup_{y\in K} \max_{x\in\overline{B(y;\delta_n)}} |c_n(x) - c_\mathrm{d}(y)|\end{aligned}$$

$$\leq \max_{x\in\Omega} |c_n(x) - c_{\mathrm{d}}(x)| + \sup_{y\in K} \max_{x\in\overline{B(y;\delta_n)}} |c_{\mathrm{d}}(x) - c_{\mathrm{d}}(y)|.$$

As $n \to \infty$, the first summand in the last line of above's estimate vanishes due to the uniform convergence $c_n \to c_{\mathrm{d}}$ in $C(\Omega)$. Likewise, the second summand vanishes, because according to the Heine-Cantor theorem the continuous function c_{d} is actually uniform continuous on the compact set Ω, i.e., for each $\varepsilon > 0$, there exists some $\rho > 0$ so that

$$|c_{\mathrm{d}}(x) - c_{\mathrm{d}}(y)| < \varepsilon \quad \text{as long as } \|x - y\| < \rho.$$

Consequently, for n large enough, $\|x - y\| < \rho$ for all $x \in \overline{B(y;\delta_n)}$ and all $y \in K$ and thus

$$\sup_{y\in K} \max_{x\in\overline{B(y;\delta_n)}} |c_{\mathrm{d}}(x) - c_{\mathrm{d}}(y)| < \varepsilon.$$

Altogether, this shows that

$$\int_{\Omega\cap\overline{B(y;\delta_n)}} c_n(x)\varphi_n(x - \cdot)\, dx \xrightarrow[n\to\infty]{C(K)} c_{\mathrm{d}},$$

which, in conjunction with the weak* convergence $\pi_n^* \rightharpoonup^* \pi^*$ in $\mathfrak{M}(K)$ and (3.31) as well as (3.32), implies that

$$(\mathcal{E}_n(c_n), \tilde{\pi}_n^*)_{L^2(\Omega^n)} \xrightarrow[n\to\infty]{} \langle \pi^*, c_{\mathrm{d}}\rangle_{C(K)^*,C(K)} = \langle \pi^*, c_{\mathrm{d}}\rangle_{C(\Omega)^*,C(\Omega)}$$

as claimed.

Ad (ii): We first note that $\operatorname{supp}(\varphi_n) \subset B_1^n \times B_2^n =: B_n$ and that

$$\Omega^n = (\Omega_1 + B_1^n) \times (\Omega_2 + B_2^n) = (\Omega_1 \times \Omega_2) + (B_1^n \times B_2^n) = \Omega + B_n,$$

where $B_i^n = \overline{B(0;\delta_n)} \subset \mathbb{R}^{d_i}$, for $i = 1, 2$. We then apply Theorem A.3 and Lemma A.4 to estimate

$$\begin{aligned}\frac{\gamma_n}{2}\|\tilde{\pi}_n^*\|_{L^2(\Omega^n)}^2 &\leq \gamma_n \|\varphi_n * \pi_n^*\|_{L^2(\Omega^n)}^2 + r_n \\ &\leq \gamma_n \|\varphi_n\|_{L^2(B_n)}^2 \|\pi_n^*\|_{\mathfrak{M}(\Omega)}^2 + r_n \\ &= \gamma_n \|\varphi_{1,n}\|_{L^2(B_1^n)}^2 \|\varphi_{2,n}\|_{L^2(B_2^n)}^2 + r_n.\end{aligned}$$

Obviously, $r_n := \gamma_n \delta_n^2 |\Omega^n| \to 0$ as $\gamma_n, \delta_n \searrow 0$. For $i = 1, 2$, we use Hölder's inequality to estimate that

$$\|\varphi_{i,n}\|_{L^2(B_i^n)}^2 = \int_{B_i^n} \big(\varphi_{i,n}\big)^2\, d\lambda_i \leq \|\varphi_{i,n}\|_{L^1(B_i^n)} \|\varphi_{i,n}\|_{L^\infty(B_i^n)} = \varphi_{i,n}(0) = \frac{C_i}{\delta_n^{d_i}},$$

where $C_i := k_i \exp(-1) > 0$, see Definition 3.12. Combining those estimates, we observe that

$$\frac{\gamma_n}{2}\|\tilde{\pi}_n^*\|_{L^2(\Omega^n)} \leq \gamma_n \frac{C_1 C_2}{\delta_n^d} + r_n \to 0 \quad \text{as} \quad \frac{\gamma_n}{\delta_n^d} \to 0,$$

see (3.26). This concludes the proof. □

Now, if we presuppose the existence of a so-called recovery sequence, we can show that the cluster point $(\bar{\pi}, \bar{\mu}_1)$ of the sequence of regularized solutions $(\bar{\pi}_n, \bar{\mu}_{1,n})_{n\in\mathbb{N}}$ is optimal for the non-regularized bilevel Kantorovich problem:

Theorem 3.34. *Let (π^*, μ_1^*) be a solution to the bilevel problem (BK) that is accompanied by a* recovery sequence*, i.e., a sequence $(\pi_n^*, \mu_{1,n}^*, c_n^*)_{n\in\mathbb{N}} \subset \mathfrak{P}(\Omega) \times \mathfrak{P}(\Omega_1) \times W^{1,p}(\Omega)$ such that*

1. *$(\pi_n^*, \mu_{1,n}^*, c_n^*)$ is feasible for (BK_n) for all $n \in \mathbb{N}$,*
2. *$\limsup_{n\to\infty} \mathcal{J}_n(\pi_n^*, \mu_{1,n}^*, c_n^*) \leq \mathcal{J}(\pi^*, \mu_1^*)$.*

Then the weak cluster point $(\bar{\pi}, \bar{\mu}_1)$ of the sequence of regularized solutions $(\bar{\pi}_n, \bar{\mu}_{1,n})_{n\in\mathbb{N}}$ is also a solution to (BK).*

Proof. Thanks to our preparatory work and the properties of the recovery sequence, the proof is fairly short. With a slight abuse of notation, we denote by $(\bar{\pi}_n, \bar{\mu}_{1,n})_{n\in\mathbb{N}}$ the subsequence that converges weakly* towards $(\bar{\pi}, \bar{\mu}_1)$. Because of the presupposed weak* lower semicontinuity of the target functional, we have

$$\begin{aligned}\mathcal{J}(\bar{\pi}, \bar{\mu}_1) &\leq \liminf_{n\to\infty} \mathcal{J}(\bar{\pi}_n, \bar{\mu}_{1,n}) \\ &\leq \liminf_{n\to\infty} \mathcal{J}(\bar{\pi}_n, \bar{\mu}_{1,n}) + \frac{1}{\gamma_n}\|\bar{c}_n - c_\mathrm{d}\|^p_{W^{1,p}(\Omega)} = \liminf_{n\to\infty} \mathcal{J}_n(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n).\end{aligned}$$

Because of the optimality of $(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n)$ for (BK_n), we observe that

$$\mathcal{J}(\bar{\pi}, \bar{\mu}_1) \leq \liminf_{n\to\infty} \mathcal{J}_n(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n) \leq \limsup_{n\to\infty} \mathcal{J}_n(\pi_n^*, \mu_{1,n}^*, c_n^*) \leq \mathcal{J}(\pi^*, \mu_1^*).$$

Thanks to Lemma 3.33, $(\bar{\pi}, \bar{\mu}_1)$ is feasible and because of (π^*, μ_1^*)'s optimality also optimal for (BK). □

Remark 3.35. In the context of the above result it is worth emphasizing that

1. the existence of a **single** recovery sequence implies the optimality of **every** cluster point (according to the arguments given at the beginning of this subchapter, several may exist) of the sequence of regularized solutions;
2. the notion of recovery sequences is closely related to the notion of Γ-convergence of functionals, since the former is essentially just a special case of the latter.

○

The proof of Theorem 3.34 shows that the cluster point of the sequence of regularized solutions must only provide feasibility for the non-regularized bilevel problem; its optimality is almost completely (aside from the lower semicontinuity of the objective function) ensured by the existence of the recovery sequence. This indicates that requiring an optimum that is accompanied by a recovery sequence is a strong assumption that may not be satisfied in general.

There are at least some cases where one can prove the existence of a recovery sequence and, consequently, the optimality of cluster points of the sequence of regularized solutions. A special case in which we can explicitly construct a recovery sequence even in the infinite-dimensional case will be covered in Subchapter 3.5.

First, however, we will convince ourselves in the following subchapter that Assumption 3.29 on the support of the marginals is not very restrictive.

3.4.1 Relaxation of the Assumption on the Support of the Marginals

The purpose of this subchapter is to ensure that the conditions of Assumption 3.29 can always be satisfied and therefore are not a real constraint.

To recall, we assumed at the beginning of Subchapter 3.1 that there exists some *distance parameter* $\Delta > 0$ such that the *extension domain*

$$\Omega^\Delta = \Omega_1^\Delta \times \Omega_2^\Delta, \quad \text{where } \Omega_i^\Delta = \Omega_i + \overline{B(0;\Delta)} \text{ for } i = 1, 2,$$

has a locally Lipschitz boundary. Moreover, the target functional $\mathcal{J}$ has an extension $\mathcal{J}^\Delta \colon \mathfrak{M}(\Omega^\Delta) \times \mathfrak{M}(\Omega_1^\Delta) \to \mathbb{R} \cup \{+\infty\}$ that is weak* lower semicontinuous, bounded ob bounded sets, and coincides with $\mathcal{J}$ for all measures π and μ_1 that are supported in Ω or Ω_1, respectively, see (3.5).

We define the *(trivial) extension* of μ_2^d to the domain Ω_2^Δ by

$$\mu_2^\Delta(B) \coloneqq \mu_2^\mathrm{d}(B \cap \Omega_2) \quad \text{for all } B \in \mathfrak{B}(\Omega_2^\Delta).$$

Moreover, let $[c^\Delta] \in W^{1,p}(\Omega^\Delta)$ be an extension of $[c_\mathrm{d}]$ to Ω^Δ and denote by c^Δ the continuous representative of $[c^\Delta]$. This extension and its continuous representative exist, because Ω and Ω^Δ have Lipschitz boundaries, see e.g. [1, Theorems 5.24 & 6.3, Part IV], respectively.

Having defined the above auxiliary data, we now consider the *regularized relaxing bilevel Kantorovich problem*

$$\begin{aligned} \inf_{\pi,\mu_1,c} \quad & \mathcal{J}_\gamma^\Delta(\pi,\mu_1,c) \coloneqq \mathcal{J}^\Delta(\pi,\mu_1) + \tfrac{1}{\gamma}\|c - c^\Delta\|_{W^{1,p}(\Omega^\Delta)}^p \\ \text{s.t.} \quad & c \in W^{1,p}(\Omega^\Delta),\ \mu_1 \in \mathfrak{P}(\Omega_1^\Delta),\ \operatorname{supp}(\mu_1) + B(0;\Delta) \subset \Omega_1^\Delta, \qquad (\mathrm{RBK}_\gamma^\delta) \\ & \pi = \big(\mathcal{E}_\delta^* \circ \mathcal{S}_\gamma\big)\big(\mathcal{E}_\delta(c), \mathcal{T}_1^\delta(\mu_1), \mathcal{T}_2^\delta(\mu_2^\Delta)\big). \end{aligned}$$

Remark 3.36.

1. Comparing ($\mathrm{RBK}_\gamma^\delta$) with ($\mathrm{BK}_\gamma^\delta$), we observe that ($\mathrm{RBK}_\gamma^\delta$) is defined on the larger domains Ω_1^Δ, Ω_2^Δ, and Ω^Δ who have all the same properties (i.e., non-emptiness, compactness, Lipschitz boundary of the Cartesian product) as their counterparts from Subchapter 3.1. We therefore can and will use the results of Subchapters 3.1 – 3.4 in the upcoming proofs.

2. Since Ω_2 is closed, for any $x \in \Omega_2^\Delta \setminus \Omega_2$ there exists an open (w.r.t. Ω_2^Δ) neighborhood N of x such that $N \cap \Omega_2 = \emptyset$. Consequently, $\mu_2^\Delta(N) = \mu_2^\mathrm{d}(\emptyset) = 0$ so that $x \notin \operatorname{supp}(\mu_2^\Delta)$. Thus, $\operatorname{supp}(\mu_2^\Delta) \subset \Omega_2$ and $\operatorname{supp}(\mu_2^\Delta) + B(0;\Delta) \subset \Omega_2^\Delta$. Analogously, if $\mu_1 \in \mathfrak{M}(\Omega_1^\Delta)$ is feasible for ($\mathrm{RBK}_\gamma^\delta$), then $\operatorname{supp}(\mu_1) + B(0;\Delta) \subset \Omega_1^\Delta$. Hence, an analogous assumption to Assumption 3.29 would be satisfied in the context of ($\mathrm{RBK}_\gamma^\delta$).

○

We are going to show that the relaxing regularized bilevel problem ($\mathrm{RBK}_\gamma^\delta$) admits a solution. First, however, we need the following lemma.

Lemma 3.37. *Consider a sequence of marginals $(\mu_{1,n})_{n\in\mathbb{N}} \subset \mathfrak{M}(\Omega_1^\Delta)$ satisfying both $\mu_{1,n} \rightharpoonup^* \mu_1 \in \mathfrak{M}(\Omega_1^\Delta)$ as $n \to \infty$ and $\operatorname{supp}(\mu_{1,n}) + B(0;\Delta) \subset \Omega_1^\Delta$ for all $n \in \mathbb{N}$. Then μ_1 is nonnegative and $\operatorname{supp}(\mu_1) + B(0;\Delta) \subset \Omega_1^\Delta$, i.e., the set*

$$\{\nu \in \mathfrak{M}(\Omega_1^\Delta) \colon \nu \geq 0,\ \operatorname{supp}(\nu) + B(0;\Delta) \subset \Omega_1^\Delta\}$$

is closed w.r.t. weak convergence.*

Proof. We first note that the nonnegativity of μ_1 follows from the exact same arguments as the nonnegativity of $\bar{\mu}_1$ in the proof of Theorem 3.26. Therefore, we consider this done.

Moreover, for arbitrary $n \in \mathbb{N}$,

$$\operatorname{supp}(\mu_{1,n}) \subset \Omega_1^\Delta \setminus \big(\partial\Omega_1^\Delta + B(0;\Delta)\big) =: M. \tag{3.33}$$

If the inclusion in (3.33) were not true, we could find $x \in \operatorname{supp}(\mu_{1,n}) \subset \Omega_1^\Delta$ as well as $a \in \partial\Omega_1^\Delta$ and $b \in B(0;\Delta)$ such that $x = a + b$. In particular, $b = x - a$. We would then define $\tilde{b} := -b + \alpha\vec{n}$ with $\alpha := {}^{(\Delta - \|b\|)}/_2 > 0$ and $\vec{n}$ being some normalized outward pointing vector at the point a, which exists because $\partial\Omega_1^\Delta$ locally takes the form of the graph of a Lipschitz continuous function. Obviously[7], $\|\tilde{b}\| \leq \|b\| + {}^{(\Delta - \|b\|)}/_2 < \Delta$ and $x + \tilde{b} = a + \alpha\vec{n} \notin \Omega_1^\Delta$. Clearly, the latter contradicts $\operatorname{supp}(\mu_{1,n}) + B(0;\Delta) \subset \Omega_1^\Delta$, so that (3.33) must be true.

We are now going to show that $\operatorname{supp}(\mu_1) \subset M$. We argue by contradiction and assume that there exists a point $x \in \operatorname{supp}(\mu_1)$ with $x \notin M$. Being the difference of a closed and an open set, M is closed. Hence, we can find a $\rho > 0$ such that $B(x;\rho) \cap M = \emptyset$. By definition of $\operatorname{supp}(\mu_1)$, we find that

$$\kappa := \mu_1\big(\overline{B(x;{}^\rho/_2)}\big) \geq \mu_1\big(B(x;{}^\rho/_2)\big) > 0.$$

Given $m \in \mathbb{N}$, Urysohn's lemma guarantees the existence of a continuous function $\phi_m \colon \Omega_1^\Delta \to [0,1]$ with

$$\phi_m \equiv \begin{cases} 1, & \text{on } \Omega_1^\Delta \cap \overline{B(x;{}^\rho/_2)}, \\ 0, & \text{on } \Omega_1^\Delta \setminus B(x;{}^\rho/_2 + {}^1/_m). \end{cases}$$

On the one hand, for all $m \in \mathbb{N}$, we apply the weak convergence $\mu_{1,n} \rightharpoonup^* \mu_1$ to find

$$\int_{\Omega_1^\Delta} \phi_m \, d\mu_{1,n} \xrightarrow[n\to\infty]{} \int_{\Omega_1^\Delta} \phi_m \, d\mu_1 \geq \int_{\overline{B(x;\rho/2)}} \phi_m \, d\mu_1 = \kappa$$

and therefore

$$\int_{\Omega_1^\Delta} \phi_m \, d\mu_{1,n} \geq \frac{\kappa}{2} \quad \text{for all } n \in \mathbb{N} \text{ sufficiently large.} \tag{3.34}$$

On the other hand, for all $n \in \mathbb{N}$,

$$\int_{\Omega_1^\Delta} \phi_m \, d\mu_{1,n} = 0 \quad \text{for all } m \in \mathbb{N} \text{ sufficiently large,} \tag{3.35}$$

because $B(x;{}^\rho/_2 + {}^1/_m) \cap M = \emptyset$, if $m > {}^2/_\rho$, and $\operatorname{supp}(\mu_{1,n}) \subset M$. Clearly, (3.34) and (3.35) contradict each other. Therefore, it must hold that $\operatorname{supp}(\mu_1) \subset M$ and

$$\operatorname{supp}(\mu_1) + B(0;\Delta) \subset M + B(0;\Delta) \subset \operatorname{int}\Omega_1^\Delta \subset \Omega_1^\Delta$$

as claimed. □

[7] Technically, $x + \tilde{b} \in \Omega_1^\Delta$ could be true, if Ω_1^Δ is non-convex and has a notch. However, in this case we can ensure the correctness of the argumentation by an additional scaling of the direction $\vec{n}$.

Theorem 3.38. *Considering the above assumptions on the domains, the target marginal, the cost function, and the target functional, for arbitrary $\gamma > 0$ and $\delta > 0$, there exists at least one optimal solution to* ($\mathrm{RBK}_\gamma^\delta$).

Proof. Analogous to the proof of Theorem 3.26, we choose a point $\hat{x} \in \Omega_1$ and set $\hat{\mu}_1 \coloneqq \delta_{\hat{x}} \in \mathfrak{P}(\Omega_1^\Delta)$ to be the Dirac measure of $\hat{x}$ in Ω_1^Δ. By construction,

$$\operatorname{supp}(\hat{\mu}_1) + B(0;\Delta) = \{\hat{x}\} + B(0;\Delta) \subset \Omega_1^\Delta,$$

i.e., feasibility of $\hat{\mu}_1$ for ($\mathrm{RBK}_\gamma^\delta$). We then again argue that the feasible set of ($\mathrm{RBK}_\gamma^\delta$) is non-empty and thus contains a minimizing sequence $(\pi_n, \mu_{1,n}, c_n)_{n\in\mathbb{N}}$. Regardless of the additional constraint on $\mu_{1,n}$, this sequence is still bounded and we can extract a subsequence (which we denote by the same symbol again) such that

$$(\pi_n, \mu_{1,n}) \rightharpoonup^* (\bar{\pi}, \bar{\mu}_1) \quad \text{in } \mathfrak{M}(\Omega^\Delta) \times \mathfrak{M}(\Omega_1^\Delta) \quad \text{and} \quad c_n \rightharpoonup \bar{c} \quad \text{in } W^{1,p}(\Omega^\Delta)$$

as $n \to \infty$. Lemma 3.37 then yields that $\bar{\mu}_1 \geq 0$ and $\operatorname{supp}(\bar{\mu}_1) + B(0;\Delta) \subset \Omega_1^\Delta$. With the usual arguments, we can then show that $\bar{\mu}_1 \in \mathfrak{P}(\Omega_1^\Delta)$ so that the cluster point $\bar{\mu}_1$ is feasible for ($\mathrm{RBK}_\gamma^\delta$).

The rest of the proof of Theorem 3.26 then carries, independently of the supports of $\bar{\mu}_1$ or $\mu_{1,n}$ for any $n \in \mathbb{N}$, over to the case of ($\mathrm{RBK}_\gamma^\delta$). □

In the following, we are going to see that we can use the solutions of the relaxing regularized bilevel Kantorovich problems ($\mathrm{RBK}_\gamma^\delta$) to approximate solutions to a bilevel problem that is very similar to (BK).

To this end, we once again choose sequences of regularization and smoothing parameters $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$, respectively, which satisfy the relation in (3.26). As was the case in Subchapter 3.4 we denote, for $n \in \mathbb{N}$, the (possibly nonunique) solution to ($\mathrm{RBK}_{\gamma_n}^{\delta_n}$) by $(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n)$. By repeating the same arguments, we can find a cluster point $(\bar{\pi}, \bar{\mu}_1) \in \mathfrak{M}(\Omega^\Delta) \times \mathfrak{M}(\Omega_1^\Delta)$ such that

$$(\bar{\pi}_n, \bar{\mu}_{1,n}) \rightharpoonup^* (\bar{\pi}, \bar{\mu}_1) \quad \text{in } \mathfrak{M}(\Omega^\Delta) \times \mathfrak{M}(\Omega_1^\Delta)$$

and, moreover, that

$$\bar{c}_n \to c^\Delta \quad \text{in } W^{1,p}(\Omega^\Delta).$$

Theorem 3.39. *The cluster point $(\bar{\pi}, \bar{\mu}_1)$ is feasible for the* relaxing bilevel Kantorovich problem

$$\begin{aligned} \inf_{\pi,\mu_1} \quad & \mathcal{J}^\Delta(\pi, \mu_1) \\ s.t. \quad & \mu_1 \in \mathfrak{P}(\Omega_1^\Delta), \ \operatorname{supp}(\mu_1) + B(0;\Delta) \subset \Omega_1^\Delta, \\ & \pi \in \arg\min\Big\{ \textstyle\int_{\Omega^\Delta} c^\Delta \, d\theta \colon \theta \in \Pi(\mu_1, \mu_2^\Delta), \ \theta \geq 0 \Big\}. \end{aligned} \tag{RBK}$$

Moreover, if there exists a solution (π^*, μ_1^*) to the relaxing bilevel problem (RBK) *which is accompanied by a recovery sequence, i.e., a sequence $(\pi_n^*, \mu_{1,n}^*, c_n^*)_{n\in\mathbb{N}} \subset \mathfrak{M}(\Omega^\Delta) \times \mathfrak{M}(\Omega_1^\Delta) \times W^{1,p}(\Omega^\Delta)$ such that*

1. *$(\pi_n^*, \mu_{1,n}^*, c_n^*)$ is feasible for ($\mathrm{RBK}_{\gamma_n}^{\delta_n}$) for all $n \in \mathbb{N}$,*
2. *$\limsup_{n\to\infty} \mathcal{J}_{\gamma_n}^\Delta(\pi_n^*, \mu_{1,n}^*, c_n^*) \leq \mathcal{J}^\Delta(\pi^*, \mu_1^*)$,*

then $(\bar{\pi}, \bar{\mu}_1)$ is optimal for (RBK).

Proof. As was the case in Subchapter 3.4, we denote all variables and entities only by the identifier n, if they depend on γ_n and/or δ_n, see Remark 3.28.

The constraints of (RBK$_n$) ensure that $\operatorname{supp}(\bar{\mu}_{1,n}) + B(0;\Delta) \subset \Omega_1^\Delta$ for all $n \in \mathbb{N}$, and, by the construction of μ_2^Δ at the beginning of this subchapter, $\operatorname{supp}(\mu_2^\Delta) + B(0;\Delta) \subset \Omega_2^\Delta$.

Because the sequence of relaxing regularized solutions $(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n)_{n\in\mathbb{N}}$ is not only feasible for the sequence of problems (RBK$_n$)$_{n\in\mathbb{N}}$ but also for the sequence of problems (BK$_n$)$_{n\in\mathbb{N}}$ w.r.t. the domains Ω_1^Δ, Ω_2^Δ, and Ω^Δ as well as the data c^Δ and μ_2^Δ, Lemma 3.33 yields that $(\bar{\pi}, \bar{\mu}_1)$ is feasible for (BK) w.r.t. the extended domains and data, i.e., $\bar{\mu}_1 \in \mathfrak{P}(\Omega_1^\Delta)$ and $\bar{\pi}$ is both feasible for (K) w.r.t. $\bar{\mu}_1$ and μ_2^Δ and optimal w.r.t. c^Δ. Moreover, Lemma 3.37 implies that $\operatorname{supp}(\bar{\mu}_1) + B(0;\Delta) \subset \Omega_1^\Delta$ so that $(\bar{\pi}, \bar{\mu}_1)$ is a feasible point for (RBK).

The proof of Theorem 3.34 directly translates into the setting of (RBK): because of the weak* lower semicontinuity of $\mathcal{J}^\Delta$, we find that

$$\begin{aligned}\mathcal{J}^\Delta(\bar{\pi}, \bar{\mu}_1) &\le \liminf_{n\to\infty} \mathcal{J}^\Delta(\bar{\pi}_n, \bar{\mu}_{1,n}) + \frac{1}{\gamma_n}\|\bar{c}_n - c^\Delta\|^p_{W^{1,p}(\Omega^\Delta)} \\ &= \liminf_{n\to\infty} \mathcal{J}_n^\Delta(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n).\end{aligned}$$

Because of the optimality of $(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n)$ for (RBK$_n$) and the properties of the recovery sequence,

$$\begin{aligned}\mathcal{J}^\Delta(\bar{\pi}, \bar{\mu}_1) &\le \liminf_{n\to\infty} \mathcal{J}_n^\Delta(\bar{\pi}_n, \bar{\mu}_{1,n}, \bar{c}_n) \\ &\le \limsup_{n\to\infty} \mathcal{J}_n^\Delta(\pi_n^*, \mu_{1,n}^*, c_n^*) \le \mathcal{J}^\Delta(\pi^*, \mu_1^*),\end{aligned}$$

i.e., optimality of $(\bar{\pi}, \bar{\mu}_1)$. □

It remains to argue that the problems (BK) and (RBK) are not only similar but also, in some sense, equivalent.

Lemma 3.40. *The bilevel Kantorovich problem* (BK) *is equivalent to the relaxing bilevel Kantorovich problem* (RBK) *in the sense that*

- *if π and μ_1 solve* (BK)*, then their trivial extensions π^Δ and μ_1^Δ solve* (RBK) *and*
- *if $\tilde{\pi}$ and $\tilde{\mu}_1$ solve* (RBK)*, then their restrictions $\tilde{\pi}|_{\mathfrak{M}(\Omega)}$ and $\tilde{\mu}_1|_{\mathfrak{M}(\Omega_1)}$ solve* (BK)*.*

Proof. We first convince ourselves that there is a one-to-one correspondence between the feasible sets of (BK) and (RBK). To this end, consider some arbitrary $\mu_1 \in \mathfrak{M}(\Omega_1)$ and its trivial extension which is defined by $\mu_1^\Delta(B_1) \coloneqq \mu_1(B_1 \cap \Omega_1)$ for all $B_1 \in \mathfrak{B}(\Omega_1^\Delta)$.

On the one hand, given some coupling $\theta \in \Pi(\mu_1, \mu_2^{\mathrm{d}})$ with $\theta \ge 0$, we consider its trivial extension $\theta^\Delta(B) \coloneqq \theta(B \cap \Omega)$ for all $B \in \mathfrak{B}(\Omega^\Delta)$. By construction, $\theta^\Delta \ge 0$ and

$$\begin{aligned}\theta^\Delta(B_1 \times \Omega_2^\Delta) &= \theta\big((B_1 \times \Omega_2^\Delta) \cap (\Omega_1 \times \Omega_2)\big) \\ &= \theta\big((B_1 \cap \Omega_1) \times (\Omega_2^\Delta \cap \Omega_2)\big) = \mu_1(B_1 \cap \Omega_1) = \mu_1^\Delta(B_1)\end{aligned}$$

for all $B_1 \in \mathfrak{B}(\Omega_1^\Delta)$. Analogously, we find that $\theta^\Delta(\Omega_1^\Delta \times B_2) = \mu_2^\Delta(B_2)$ for all $B_2 \in \mathfrak{B}(\Omega_2^\Delta)$, so that $\theta^\Delta \in \Pi(\mu_1^\Delta, \mu_2^\Delta)$.

On the other hand, given some coupling $\tilde{\theta} \in \Pi(\mu_1^\Delta, \mu_2^\Delta)$ with $\tilde{\theta} \geq 0$, we consider its restriction $\tilde{\theta}|_{\mathfrak{M}(\Omega)} \colon \mathfrak{B}(\Omega) \ni B \mapsto \tilde{\theta}(B) \in \mathbb{R}_+$. By construction, for $i = 1, 2$, $\operatorname{supp}(\mu_i^\Delta) \subset \Omega_i$, see the arguments in the second part of Remark 3.36. Lemma C.1 then ensures that

$$\operatorname{supp}(\tilde{\theta}) \subset \operatorname{supp}(\mu_1^\Delta) \times \operatorname{supp}(\mu_2^\Delta) \subset \Omega_1 \times \Omega_2 = \Omega.$$

Because, for arbitrary $B_1 \in \mathfrak{B}(\Omega_1)$,

$$\big(B_1 \times (\Omega_2^\Delta \setminus \Omega_2)\big) \cap \operatorname{supp}(\tilde{\theta}) = \emptyset,$$

Lemma B.18 yields that $\tilde{\theta}\big(B_1 \times (\Omega_2^\Delta \setminus \Omega_2)\big) = 0$ and we find that

$$\tilde{\theta}|_{\mathfrak{M}(\Omega)}(B_1 \times \Omega_2) = \tilde{\theta}(B_1 \times \Omega_2) = \tilde{\theta}(B_1 \times \Omega_2^\Delta) = \mu_1^\Delta(B_1) = \mu_1(B_1).$$

Analogously, $\tilde{\theta}|_{\mathfrak{M}(\Omega)}(\Omega_1 \times B_2) = \mu_2^{\mathrm{d}}(B_2)$ for all $B_2 \in \mathfrak{B}(\Omega_2)$, so that $\tilde{\theta}|_{\mathfrak{M}(\Omega)} \in \Pi(\mu_1, \mu_2^{\mathrm{d}})$. This shows that there is a one-to-one correspondence between the sets $\{\theta \in \Pi(\mu_1, \mu_2^{\mathrm{d}}) \colon \theta \geq 0\}$ and $\{\tilde{\theta} \in \Pi(\mu_1^\Delta, \mu_2^\Delta) \colon \tilde{\theta} \geq 0\}$ via the trivial extension to Ω^Δ and its inverse, the restriction to Ω.

Moreover, given $\theta \in \Pi(\mu_1, \mu_2^{\mathrm{d}})$ and $\theta^\Delta \in \Pi(\mu_1^\Delta, \mu_2^\Delta)$, we observe that

$$\int_{\Omega^\Delta} c^\Delta \, d\theta^\Delta = \int_{\Omega^\Delta \setminus \Omega} c^\Delta \, d\theta^\Delta + \int_\Omega c^\Delta \, d\theta^\Delta = \int_\Omega c_{\mathrm{d}} \, d\theta,$$

because $\operatorname{supp}(\theta^\Delta) \subset \Omega$ and c^Δ was defined to coincide with c_{d} on Ω, see its definition at the beginning of this subchapter. This shows that if (π, μ_1) is feasible for (BK), then $(\pi^\Delta, \mu_1^\Delta)$ is feasible for (RBK). Repeating the same argument for $(\tilde{\pi}, \tilde{\mu}_1)$ and its restriction $(\tilde{\pi}|_{\mathfrak{M}(\Omega)}, \tilde{\mu}_1|_{\mathfrak{M}(\Omega_1)})$ then proves the claimed one-to-one correspondence from the beginning of this proof.

Realizing that the values of the upper-level target functionals $\mathcal{J}$ and $\mathcal{J}^\Delta$ coincide for each pair (π, μ_1) and $(\pi^\Delta, \mu_1^\Delta)$ as well as $(\tilde{\pi}, \tilde{\mu}_1)$ and $(\tilde{\pi}|_{\mathfrak{M}(\Omega)}, \tilde{\mu}_1|_{\mathfrak{M}(\Omega_1)})$ then establishes the assertion of the lemma. □

To summarize, we have shown that Assumption 3.29 from Subchapter 3.4 is by no means a limitation for the analysis performed there: we can always resort to a similar problem on a larger domain that guarantees said assumption and perform the approximation for that problem. In the end, we obtain a solution to a problem that is actually equivalent to the non-regularized bilevel problem and we even know how to convert the solutions into each other.

In the following subchapter, we will finally present a setting for which we can guarantee the existence of a (trivial) recovery sequence and therefore the approximation property we have seen in Theorem 3.34.

3.5 Existence of a Recovery Sequence for the Bilevel Kantorovich Problem

As announced at the end of the previous subchapter, we now present a setting in which we can guarantee the existence of an optimal solution for (BK) that is accompanied by a recovery sequence, as has been assumed in Theorem 3.34.

We again consider the setting of Subchapter 3.4, i.e., we assume that we are given suitably coupled vanishing sequences of regularization parameters and

smoothing parameters, $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$, respectively, as well as a sequence of regularized solutions $(\pi_n, \mu_{1,n}, c_n)_{n\in\mathbb{N}}$ to the sequence of problems $(\mathrm{BK}^{\delta_n}_{\gamma_n})_{n\in\mathbb{N}}$ which has the cluster point $(\bar{\pi}, \bar{\mu}_1, c_\mathrm{d})$ w.r.t. the right topologies.

Corollary 3.41. *Assume that $\Omega_1 = \Omega_2$ and that μ_2^d is absolutely continuous w.r.t. λ_2. Moreover, assume that (BK)'s cost function c_d takes the form $c_\mathrm{d}(x_1, x_2) = \|x_1 - x_2\|^\rho$, for some $\rho > 1$, and that the objective $\mathcal{J}\colon \mathfrak{M}(\Omega) \times \mathfrak{M}(\Omega_1) \to \mathbb{R} \cup \{+\infty\}$ is, in addition to its other properties, weak* continuous w.r.t. π.*

Then, the cluster point $(\bar{\pi}, \bar{\mu}_1)$ is an optimum of the non-regularized bilevel Kantorovich problem (BK).

Remark 3.42. There are a few points in the formulation of Corollary 3.41 that need some additional notes.

- The assumption on the domains automatically implies that $d_1 = d_2$, $\lambda_1 = \lambda_2$, as well as $\mathfrak{B}(\Omega_1) = \mathfrak{B}(\Omega_2)$.
- A measure μ is said to be *absolutely continuous* w.r.t. some other measure ν if it holds that $\mu(B) = 0$ for every measurable set B with $\nu(B) = 0$.
- The cost function c_d is strictly convex, symmetric, and superlinear in the sense of [76, p. 90].
- The weak* continuity of $\mathcal{J}$ w.r.t. π still allows the target functional to be only (weak*) semicontinuous w.r.t. to its second variable.

○

Proof of Corollary 3.41. According to Theorem 3.5, there exists a solution of (BK) which we denote by (π^*, μ_1^*). Following the argumentation in Subchapter 3.4.1, we may assume that $\operatorname{supp}(\mu_1^*) + B(0;\Delta) \subset \Omega_1$. [76, Theorem 2.44] in combination with Lemma C.2 guarantees that π^* must be the unique optimal transport plan between μ_1^* and μ_2^d w.r.t. the cost c_d. We define

$$(\mu_{1,n}^*, c_n^*) \coloneqq (\mu_1^*, c_\mathrm{d}) \quad \text{and} \quad \pi_n^* \coloneqq \big(\mathcal{E}_n^* \circ \mathcal{S}_n\big)\big(\mathcal{E}_n(c_\mathrm{d}), \mathcal{T}_1^n(\mu_1^*), \mathcal{T}_2^n(\mu_2^\mathrm{d})\big)$$

for all $n \in \mathbb{N}$. By construction, the sequence $(\pi_n^*, \mu_{1,n}^*, c_n^*)_{n\in\mathbb{N}}$ is feasible for the sequence of problems $(\mathrm{BK}^{\delta_n}_{\gamma_n})_{n\in\mathbb{N}}$.

Applying Lemma 3.33 to the sequence $(\pi_n^*, \mu_{1,n}^*, c_n^*)_{n\in\mathbb{N}}$, extracting a subsequence (which we denote by the same symbol), and taking advantage of the uniqueness of π^*, we obtain that $\pi_n^* \rightharpoonup^* \pi^*$ as $n \to \infty$. The weak* continuity of $\mathcal{J}$ w.r.t. π is then sufficient to show that $(\pi_n^*, \mu_{1,n}^*, c_n^*)_{n\in\mathbb{N}}$ is an accompanying recovery sequence for the optimum (π^*, μ_1^*). Theorem 3.34 then establishes the statement of the corollary. □

Remark 3.43. Inspecting the presented proofs of the lemmas and theorems of the current chapter, we note that at no point there was a need to actually choose a cost function other than the fixed cost c_d. Even in the construction of the recovery sequence in the above proof, it is sufficient to fix the cost function to c_d.

However, this will change in Part II of this thesis. In Subchapter 4.5 of the finite-dimensional case, we will see an example of a comparable but slightly more

general setting than the one presented in Corollary 3.41, where we will indeed need the cost function to be an optimization variable to prove the existence of a recovery sequence.

It is therefore not unreasonable to assume that the ability to use the cost function as an optimization variable may prove to be handy in other infinite-dimensional examples, apart from the (trivial) setting of Corollary 3.41. ○

We conclude the present part of this thesis by presenting two examples of relevant applications, where the assumptions of Corollary 3.41 and the assumptions on the target functional from (3.3) – (3.5) are fulfilled. The first example is the infinite-dimensional analogon of the transportation identification problem (TIP) which we motivated in Chapter 1 and which we will consider as our numerical test case in Chapter 6.

Example 3.44. Consider the setting described in Corollary 3.41. Given $p > d_1 + d_2$, $p' = {}^{p}/_{p-1}$, as well as some non-empty open set $D \subset \Omega$ that has a locally Lipschitz boundary, an example for a weak* continuous (w.r.t. the first variable) target functional is given by the *tracking-type target functional*

$$\mathcal{J}(\pi, \mu_1) = \|\pi - \pi_\mathrm{d}\|_{W^{-1,p'}(D)} + \left\|\mu_1 - \mu_1^\mathrm{d}\right\|_{\mathfrak{M}(\Omega_1)},$$

where $\pi_\mathrm{d} \in \mathfrak{M}(D)$ and $\mu_1^\mathrm{d} \in \mathfrak{M}(\Omega_1)$ are given data.

That $\mathcal{J}$ indeed is weak* continuous, can be seen as follows. We know from [1, Theorem 6.3] that $W_0^{1,p}(D)$ embeds compactly in $C_0(D)$. Conversely, Schauder's theorem for adjoint operators, see e.g. [50, Theorem 8.2-5], yields that

$$\mathfrak{M}(D) \cong \left(C_0(D)\right)^* \overset{c}{\hookrightarrow} \left(W_0^{1,p}(D)\right)^* \cong W^{-1,p'}(D).$$

Therefore, if $\pi_n \rightharpoonup^* \pi$ in $\mathfrak{M}(D)$, then $\pi_{n_k} \to \tilde{\pi}$ in $W^{-1,p}(D)$ along some subsequence $(n_k)_{k\in\mathbb{N}}$. At the same time, $\pi_{n_k} \rightharpoonup^* \pi$ in $W^{-1,p}(D)$. Because the weak* limit is unique, $\tilde{\pi} = \pi$. If we apply this argument to an arbitrary subsequence of $(\pi_n)_{n\in\mathbb{N}}$, Lemma D.5 guarantees that $\pi_n \to \pi$ in $W^{-1,p^*}(D)$ and therefore $\|\pi_n - \pi_\mathrm{d}\|_{W^{-1,p'}(D)} \to \|\pi - \pi_\mathrm{d}\|_{W^{-1,p'}(D)}$ as $n \to \infty$.

To see that this already yields the weak* continuity of $\mathcal{J}$ w.r.t. π, we observe that restricting a sequence $(\pi_n)_{n\in\mathbb{N}} \subset \mathfrak{M}(\Omega)$ to $\mathfrak{M}(D)$ is a (weak*) continuous operation: if $\pi_n \rightharpoonup^* \pi$ in $\mathfrak{M}(\Omega)$, we find for arbitrary $\phi \in C_0(D)$ that

$$\begin{aligned}\langle \pi_n|_{\mathfrak{M}(D)}, \phi\rangle_{\mathfrak{M}(D),C_0(D)} &= \int_D \phi \, d\pi_n|_{\mathfrak{M}(D)} \\ &= \int_\Omega \mathcal{E}(\phi) \, d\pi_n \to \int_\Omega \mathcal{E}(\phi) \, d\pi = \langle \pi|_{\mathfrak{M}(D)}, \phi\rangle_{\mathfrak{M}(D),C_0(D)}\end{aligned}$$

as $n \to \infty$, where $\mathcal{E}(\phi)$ denotes the trivial (continuous) extension of ϕ to Ω. Note that this argument fails, if D is closed. In this case, $\phi \in C_0(D) = C(D)$ does in general not have a continuous extension to Ω that leaves the value of the integral unchanged when transitioning to the larger domain. The sequence defined by $\pi_n := \delta_{x_n}$, where the sequence of points $(x_n)_{n\in\mathbb{N}} \subset \Omega \setminus D$ converges to some $x \in D$, provides an easy counterexample.

$\mathcal{J}$'s boundedness on bounded sets and its weak* lower semicontinuity are evident from the properties of the norms on $\mathfrak{M}(\Omega_1)$ and $W^{-1,p'}(D)$, see Lemma D.4, and the compactness of the embedding of $\mathfrak{M}(D)$ into $W^{-1,p'}(D)$.

An extension of $\mathcal{J}$ to $\mathfrak{M}(\Omega^\Delta) \times \mathfrak{M}(\Omega_1^\Delta)$ in the sense of Subchapter 3.1 is given by

$$\mathcal{J}^\Delta(\pi, \mu_1) = \|\pi - \pi_\mathrm{d}\|_{W^{-1,p'}(D)} + \|\mu_1 - \mu_1^{\mathrm{d},\Delta}\|_{\mathfrak{M}(\Omega_1^\Delta)},$$

where $\mu_1^{\mathrm{d},\Delta} \in \mathfrak{M}(\Omega_1^\Delta)$ is the trivial extension of $\mu_1^\mathrm{d} \in \mathfrak{M}(\Omega_1)$ to Ω_1^Δ, see Subchapter 3.4.1. $\mathcal{J}^\Delta$ is bounded on bounded sets and weak* lower semicontinuous by the same reasons as $\mathcal{J}$. That $\mathcal{J}^\Delta$ indeed satisfies (3.5) can be seen by evaluating the (dual) norms on $W^{-1,p'}(D)$ and $\mathfrak{M}(\Omega_1^\Delta)$, see [44, Section 4.1] for a representation of the former.

Consequently, in this setting Corollary 3.41 guarantees that any cluster point of the sequence of solutions to the regularized bilevel problems ($\mathrm{BK}_\gamma^\delta$) is a solution to the non-regularized bilevel problem (BK). ◊

The second example represents a class of optimization problems for which the assumptions of Corollary 3.41 are already fulfilled by the very definition of the problem and which at the same time represents a relevant example of possible applications of (BK).

Example 3.45. Let $\Omega_* \subset \mathbb{R}^{d_*}$, for $d_* \in \mathbb{N}$, be a given compact domain such that both $\Omega := \Omega_* \times \Omega_*$ as well as, for some $\Delta > 0$, its extension $\Omega + \overline{B(0;\Delta)} \times \overline{B(0;\Delta)}$ have locally Lipschitz boundaries that are negligible w.r.t. to the Lebesgue measure on $\mathbb{R}^{d_*} \times \mathbb{R}^{d_*}$. Moreover, assume that we are given a compact linear operator G which maps $\mathfrak{M}(\Omega_*)$ to some Banach space Y and assume that G has an extension $G^\Delta \colon \mathfrak{M}(\Omega_*^\Delta) \to Y$, where $\Omega_*^\Delta := \Omega_* + \overline{B(0;\Delta)}$ denotes the extension of the domain Ω_*, that is a linear compact operator and satisfies $G^\Delta \mu = G\mu|_{\mathfrak{M}(\Omega_*)}$ for all $\mu \in \mathfrak{M}(\Omega_*^\Delta)$ with $\mathrm{supp}(\mu) \subset \Omega_*$. Let $\rho > 1$ as well as $\nu > 0$ be given parameters and consider the prior $\mu_\mathrm{d} \in \mathfrak{P}(\Omega_*)$, which is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$, as well as the target data $y_\mathrm{d} \in Y$. We then consider the *Wasserstein inverse problem*

$$\begin{aligned} \inf_{\mu} \quad & \tfrac{1}{2}\|G\mu - y_\mathrm{d}\|_Y^2 + \nu W_\rho(\mu, \mu_\mathrm{d})^\rho \\ \text{s.t.} \quad & \mu \in \mathfrak{P}(\Omega_*), \end{aligned} \tag{WIP}$$

where the distance between the variable μ and the data μ^d is measured by the Wasserstein ρ-distance.

To see that (WIP) exactly fits into the setting of (BK), we need to reformulate it a bit. For any $\mu \in \mathfrak{P}(\Omega_*)$, we find that

$$\int_{\Omega_*} \|x\|^\rho \, d\mu \leq C\|\mu\|_{\mathfrak{M}(\Omega_*)} = C,$$

where the constant $C > 0$ arises from the boundedness of the domain Ω_*. Therefore, $\mu \in \mathfrak{P}_\rho(\Omega_*)$, where the latter denotes the set of *probability measures with finite ρ-th moment*. Using the definition of the *Wasserstein ρ-distance*, i.e.,

$$W_\rho(\mu_1, \mu_2) := \min\left\{ \int_{\Omega_* \times \Omega_*} \|x - y\|^\rho_{\mathbb{R}^{d_*} \times \mathbb{R}^{d_*}} \, d\theta \colon \theta \in \Pi(\mu_1, \mu_2), \ \theta \geq 0 \right\}^{\frac{1}{\rho}}$$

for all $\mu_1, \mu_2 \in \mathfrak{P}_\rho(\Omega_*)$, see e.g. [68, Chapter 5], we see that (WIP) is equivalent

to

$$\begin{aligned} \inf_{\mu} \quad & \frac{1}{2}\|G\mu - y_d\|_Y^2 + \nu \int_\Omega \|x-y\|^\rho \, d\pi \\ \text{s.t.} \quad & \mu_1 \in \mathfrak{P}(\Omega_*), \\ & \pi \in \arg\min \left\{ \int_\Omega \|x-y\|^\rho \, d\theta \colon \theta \in \Pi(\mu, \mu_\mathrm{d}),\, \theta \geq 0 \right\}, \end{aligned}$$

which is a problem of the form (BK) with $\mu_1 = \mu$, $\mu_2^\mathrm{d} = \mu_\mathrm{d}$, the cost function $c_d(x-y) = \|x-y\|^\rho$, and the target functional

$$\mathcal{J}(\pi, \mu) = \nu \langle \pi, c_\mathrm{d} \rangle_{C(\Omega)^*, C(\Omega)} + \frac{1}{2}\|G\mu - y_\mathrm{d}\|_Y^2.$$

The continuity of both the compact operator G and the norm on Y ensure that $\mathcal{J}$ is weakly* continuous w.r.t. μ. $\mathcal{J}$ is weak* continuous w.r.t. π, because by definition the weak* convergence of the transport plans implies the convergence of the dual pairing. The boundedness of $\mathcal{J}$ on bounded sets is due to the boundedness of the dual pairing $\langle \pi, c_\mathrm{d} \rangle$ and the compactness of the operator G. Again, an (obvious) extension of $\mathcal{J}$ to $\mathfrak{M}(\Omega^\Delta) \times \mathfrak{M}(\Omega_1^\Delta)$ is given by

$$\mathcal{J}^\Delta(\pi, \mu) = \nu \langle \pi, c_\mathrm{d} \rangle_{C(\Omega^\Delta)^*, C(\Omega^\Delta)} + \frac{1}{2}\left\|G^\Delta \mu - y_\mathrm{d}\right\|_Y^2.$$

The boundedness of $\mathcal{J}^\Delta$ on bounded sets and its weak* lower semicontinuity follow from the same arguments as in the case of $\mathcal{J}$. That $\mathcal{J}^\Delta$ satisfies the property in (3.5) follows from the properties of both the dual pairing and the operator G^Δ.

Consequently, we can apply Corollary 3.41 to the above setting: if we consider the regularized bilevel Kantorovich problem ($\mathrm{BK}_\gamma^\delta$) with $\mathcal{J}$, c_d, and μ_2^d from above, we know that the sequence of regularized solutions has a (weak*) cluster point that is a solution to the Wasserstein inverse problem (WIP).

Natural choices for the operator G include, for instance:

– Convolutions with a fixed mollifier $\varphi \in C_c^\infty(\mathbb{R}^{d_*})$, i.e.,

$$G \colon \mathfrak{M}(\Omega_*) \to L^p(\mathbb{R}^{d_*}), \quad \mu \mapsto \varphi * \mu.$$

The linearity of the (signed) integration, see Lemma B.12, shows that it is a linear operator while we can prove its compactness analogously to the proof of Theorem 3.26. An extension of G that satisfies all of the presupposed properties is given by the (natural) extension $G^\Delta \colon \mathfrak{M}(\Omega_*^\Delta) \to L^p(\mathbb{R}^{d_*})$, $\mu \mapsto \varphi * \mu$, see Definition A.1.

– Solution operators of (elliptic) partial differential equations. With this choice, (WIP) becomes an optimal control problem, where the control μ "lives" on the measure space $\mathfrak{P}(\Omega_*)$. This particular case is, in more detail, discussed in [44, Section 4].

◊

In the next part of this thesis we will first derive the same results we found in the present part but in a finite-dimensional setting. Furthermore, we will deal with the differentiability of a regularized version of the dual of the regularized transport problem and derive an implicit programming approach which we will then implement and test numerically.

Part II

The Finite-Dimensional Case

Chapter 4

Bilevel Optimization of the Hitchcock Optimal Transport Problem

We now leave the infinite-dimensional setting and enter finite-dimensional terrain, which is of more application-oriented than theoretical interest. In particular, if we want to reproduce and solve the bilevel problem (BK) numerically on a computer, we inevitably end up in the finite-dimensional case after a discretization of the variables.

The present chapter is organized analogously to Chapter 3. This should simplify the comparison of the two scenarios and also show the (subtle) differences.

4.1 Problem Statement

Again, we start with a rigorous definition of the bilevel optimization problem we are interested in. First, we define the lower-level Hitchcock problem and then, based on this, formulate the bilevel optimization problem. We achieve the former by deriving the Hitchcock problem from the Kantorovich problem from Chapter 3.

To this end, we choose, for $n_1, n_2 \in \mathbb{N}$, the non-empty finite sets $\Omega_1 = \{1, \dots, n_1\}$ and $\Omega_2 = \{1, \dots, n_2\}$ and again abbreviate their Cartesian product by $\Omega \coloneqq \Omega_1 \times \Omega_2$. If we equip these sets with the discrete topology, then Ω_1 and Ω_2 are compact in $\mathbb{R}$, Ω is compact in $\mathbb{R}^2$, as well as the Borel sigma algebras $\mathfrak{B}(\Omega_1)$, $\mathfrak{B}(\Omega_2)$, and $\mathfrak{B}(\Omega)$ are just the power sets of Ω_1, Ω_2, as well as Ω, respectively.

According to Lemma B.15, we find that $\mathfrak{M}(\Omega_1) \cong \mathbb{R}^{n_1}$, $\mathfrak{M}(\Omega_2) \cong \mathbb{R}^{n_2}$, and $\mathfrak{M}(\Omega) \cong \mathbb{R}^{n_1 \times n_2}$. Moreover, because every subset of Ω (and as a consequence the preimage of every function $f \colon \Omega \to \mathbb{R}$) is open w.r.t. the discrete topology, we have that $C(\Omega) \cong \mathbb{R}^{n_1 \times n_2}$.

Because of the above choice of the *domains* Ω_1 and Ω_2 and the resulting isomorphisms between the measure spaces and the finite-dimensional vector spaces, we can always uniquely replace the marginals $\mu_1 \in \mathfrak{M}(\Omega_1)$ and $\mu_2 \in \mathfrak{M}(\Omega_2)$ by vectors $\mu_1 \in \mathbb{R}^{n_1}$ and $\mu_2 \in \mathbb{R}^{n_2}$, respectively, which we denote identically for

simplicity. This has several consequences:

- Since, by Lemma B.15, $\|\mu_i\|_{\mathfrak{M}(\Omega_i)} = \|\mu_i\|_1$ for $i = 1, 2$, the marginals μ_1 and μ_2 are compatible if and only if the vectors μ_1 and μ_2 are elementwise nonnegative and

$$\|\mu_1\|_1 = \mathbb{1}^\top \mu_1 = \mathbb{1}^\top \mu_2 = \|\mu_2\|_1,$$

  where $\mathbb{1}$ denotes the vector in $\mathbb{R}^{n_1}$ or $\mathbb{R}^{n_2}$ (we use the same symbol for both) whose components are all equal to 1. In the following, we will use the term *marginals* indistinguishable for both the elements of the measure spaces as well as their vector representations.

- The set of transport plans (couplings) between μ_1 and μ_2 takes the form

$$\Pi(\mu_1, \mu_2) = \{\pi \in \mathbb{R}^{n_1 \times n_2} \colon \pi\mathbb{1} = \mu_1,\ \pi^\top \mathbb{1} = \mu_2\},$$

  where the equality of the sets on the left-hand side and the right-hand side of the equation is understood w.r.t. the isometric isomorphism between $\mathfrak{M}(\Omega)$ and $\mathbb{R}^{n_1 \times n_2}$. Again, we will use the term *transport plan* (*coupling*) for both the measure on Ω as well as its matrix representation.

- Consider some arbitrary cost function $c \colon \Omega \to \mathbb{R}$. Because we equipped Ω with the discrete topology, the function c is at the same time continuous, $\mathfrak{B}(\Omega)$-$\mathfrak{B}(\mathbb{R})$-measurable, and a simple function. Hence, we can express the total cost of transportation realized by some transport plan $\pi \in \mathfrak{M}(\Omega)$ by

$$\int_\Omega c \, d\pi = \sum_{(i_1, i_2) \in \Omega} c_{i_1, i_2} \pi_{i_1, i_2} = (c, \pi)_F,$$

  where $c \in \mathbb{R}^{n_1 \times n_2}$ and $\pi \in \mathbb{R}^{n_1 \times n_2}$ denote the matrix representations of the cost function c, which we will call *cost matrix*, and the transport plan π, respectively.

Consequently, in the above setting, given the marginals μ_1 and μ_2 as well as the cost matrix c the Kantorovich problem (K) is equivalent to the problem

$$\begin{aligned} \inf_{\pi} \quad & (\pi, c)_F \\ \text{s.t.} \quad & \pi \in \mathbb{R}^{n_1 \times n_2},\ \pi\mathbb{1} = \mu_1,\ \pi^\top \mathbb{1} = \mu_2,\ \pi \geq 0, \end{aligned} \tag{H}$$

which is also known as the *Hitchcock problem (of optimal transportation)*, see e.g. [47] or [37], and plays a significant role in many fields including economics, logistics, integrated circuit design, and image processing, see the references in the introduction of this thesis.

Remark 4.1. In the situation of (H), we note the following:

- Contrary to the widely spread convention of denoting real matrices with capital letters, we stick to the notation of Part I and denote the cost matrices and (discrete) transport plans by the lowercase letters c and π, respectively.

– As is the case with any other norm defined on an finite-dimensional vector space, the Frobenius norm is equivalent to the 1-norm. Because of the subadditivity of the square root,

$$\|A\|_F = \sqrt{\sum_{(i_1,i_2)\in\Omega} A_{i_1,i_2}^2} \leq \sum_{(i_1,i_2)\in\Omega} \sqrt{A_{i_1,i_2}^2} = \sum_{(i_1,i_2)\in\Omega} |A_{i_1,i_2}| = \|A\|_1$$

and, because of the Cauchy-Schwarz inequality,

$$\begin{aligned}\|A\|_1 &= \big((\operatorname{sgn} A_{i_1,i_2})_{(i_1,i_2)\in\Omega}, A\big)_F \\ &\leq \big\|(\operatorname{sgn} A_{i_1,i_2})_{(i_1,i_2)\in\Omega}\big\|_F \|A\|_F = \sqrt{n_1 n_2}\|A\|_F\end{aligned}$$

for all $A = (A_{i_1,i_2})_{(i_1,i_2)\in\Omega} \in \mathbb{R}^{n_1\times n_2}$.

○

We could content ourselves at this point with citing the existence result of the Kantorovich problem (K) from Chapter 3 to show the existence of solutions to the Hitchcock problem (H). However, this is not necessary since the Hitchcock problem (H) is an optimization problem with a compact feasible set and a continuous objective function and therefore takes its minimum and maximum. This immediately gives us the following lemma:

Lemma 4.2. *For every choice of compatible marginals μ_1 and μ_2 and every cost matrix c, the Hitchcock problem (H) possesses at least one optimal solution.*

Similar to the infinite-dimensional case of Chapter 3, we will now formulate the bilevel problem that will be of interest for the rest of this chapter.

To this end and for the rest of this chapter, we fix a target marginal $\mu_2^{\mathrm{d}} \in \mathbb{R}^{n_2}$ with $\mu_2^{\mathrm{d}} \geq 0$ and $\mathbb{1}^\top \mu_2^{\mathrm{d}} = 1$ as well as some matrix $c_{\mathrm{d}} \in \mathbb{R}^{n_1\times n_2}$, the cost matrix of the Kantorovich problem. Given this data, the *bilevel Hitchcock (optimal transport) problem* reads

$$\begin{aligned}\inf_{\pi,\mu_1}\quad & \mathcal{J}(\pi,\mu_1) \\ \text{s.t.}\quad & \mu_1 \in \mathbb{R}^{n_1}, \quad \mu_1 \geq 0, \quad \mathbb{1}^\top\mu_1 = 1, \\ & \pi \in \arg\min\big\{(\theta, c_{\mathrm{d}})_F \colon \theta \in \mathbb{R}^{n_1\times n_2},\ \theta \geq 0,\ \theta\mathbb{1} = \mu_1,\ \theta^\top\mathbb{1} = \mu_2^{\mathrm{d}}\big\}\end{aligned} \tag{BH}$$

where $\mathcal{J}\colon \mathbb{R}^{n_1\times n_2}\times\mathbb{R}^{n_1} \to \mathbb{R}$ is a lower semi-continuous target function which is *bounded on bounded sets*, i.e., for all $M > 0$ it holds that

$$\sup_{\|(\pi,\mu_1)\|<M} \mathcal{J}(\pi,\mu_1) < \infty.$$

Remark 4.3. We could state the bilevel Hitchcock problem (BH) without explicitly specifying the constraints for μ_1, i.e., $\mu_1 \geq 0$ and $\mathbb{1}^\top\mu_1 = 1$, as these are implicitly implied by the constraints of (H). However, we prefer to specify them anyway to rule out the possibility that the feasible set of (H) becomes empty. ○

Before we proceed to prove the existence of solutions to (BH), we first prove the following lemma, which can be seen as the finite-dimensional counterpart of Lemma 3.31 and will be useful throughout the entire subchapter.

Lemma 4.4. *Consider the marginals $\mu \in \mathbb{R}^{n_1}$ and $\nu \in \mathbb{R}^{n_2}$ with $\mu, \nu \geq 0$ and $\mathbb{1}^\top \mu = \mathbb{1}^\top \nu = 1$ as well as the coupling $\pi \in \Pi(\mu, \nu) \subset \mathbb{R}^{n_1 \times n_2}$. If $(\mu_k)_{k\in\mathbb{N}}$ is a given sequence of nonnegative marginals with $\mathbb{1}^\top \mu_k = 1$ for all $k \in \mathbb{N}$ and $\mu_k \to \mu$ for $k \to \infty$, then there exists a sequence of couplings $(\pi_k)_{k\in\mathbb{N}}$ with $\pi_k \in \Pi(\mu_k, \nu)$ for all $k \in \mathbb{N}$ and $\pi_k \to \pi$ for $k \to \infty$.*

Proof. Although at this point we could simply refer to the proof of Lemma 3.31, here we give a slightly simpler proof.

For every $k \in \mathbb{N}$, there exists a nonnegative optimal transport plan $\theta_k \in \Pi(\mu_k, \mu) \subset \mathbb{R}^{n_1 \times n_1}$ between μ_k and μ with respect to the cost function

$$c(i, j) \coloneqq |i - j| \quad \text{for all } i, j \in \{1, \dots, n_1\}.$$

Let us, for $k \in \mathbb{N}$, define the coupling

$$\pi_k^{i_1,i_2} \coloneqq \sum_{l=1}^{n_1} \frac{\theta_k^{i_1,l} \pi^{l,i_2}}{\mu^l} \quad \text{for all } (i_1, i_2) \in \Omega.$$

Its nonnegativity comes straight from the nonnegativity of π, θ, and μ. Also,

$$(\pi_k \mathbb{1})^{i_1} = \sum_{i_2=1}^{n_2} \pi_k^{i_1,i_2} = \sum_{l=1}^{n_1} \frac{\theta_k^{i_1,l}}{\mu^l} \sum_{i_2=1}^{n_2} \pi^{l,i_2} = \sum_{l=1}^{n_1} \theta_{i_1,l}^{k} = \mu_k^{i_1}$$

and

$$(\pi_k^\top \mathbb{1})^{i_2} = \sum_{i_1=1}^{n_1} \pi_k^{i_1,i_2} = \sum_{l=1}^{n_1} \frac{\pi^{l,i_2}}{\mu^l} \sum_{i_1=1}^{n_1} \theta_k^{i_1,l} = \sum_{l=1}^{n_1} \pi^{l,i_2} = \nu^{i_2}$$

for all $(i_1, i_2) \in \Omega$, i.e., π_k indeed is a coupling between μ_k and ν.

One easily verifies that

$$\bar{\theta} \coloneqq \operatorname{diag}(\mu) \in \Pi(\mu, \mu) \quad \text{with} \quad \sum_{i,j=1}^{n_1} c(i, j)\, \bar{\theta} = 0$$

is the unique optimal transport plan between μ and itself. Because of $\|\theta_k\|_1 = 1$ for all $k \in \mathbb{N}$, there exists a convergent subsequence (which we denote by the same symbol) such that $\theta_k \to \tilde{\theta} \in \mathbb{R}^{n_1 \times n_1}$ as $k \to \infty$. The stability theorem from [68, Theorem 1.50] then ensures that $\tilde{\theta} = \bar{\theta}$ so that the whole sequence converges to $\bar{\theta}$, see Lemma D.5.

Hence, it follows from the definition of π_k and the convergence $\theta_k \to \bar{\theta}$ that

$$\lim_{k\to\infty} \pi_k^{i_1,i_2} = \sum_{l=1}^{n_1} \frac{\lim_{k\to\infty} \theta_k^{i_1,l} \pi^{l,i_2}}{\mu^l} = \sum_{l=1}^{n_1} \frac{\bar{\theta}^{i_1,l} \pi^{l,i_2}}{\mu^l} = \frac{\bar{\theta}^{i_1,i_1} \pi^{i_1,i_2}}{\mu^{i_1}} = \pi^{i_1,i_2},$$

for all $(i_1, i_2) \in \Omega$, as claimed. □

The following proof is quite short and manages to use, besides above's lemma, only standard arguments.

Theorem 4.5. *With the target marginal and the cost function given as above, the bilevel Hitchcock problem (BH) possesses at least one optimal solution.*

Proof. As usual, we denote the feasible set of (BH) by $\mathcal{F}$. The marginal $\hat{\mu}_1 \coloneqq (1, 0, \dots, 0)^\top \in \mathbb{R}^{n_1}$ is feasible for (BH). The feasible set of (H) w.r.t. $\hat{\mu}_1$ and μ_2^{d} consists solely of the matrix $\hat{\pi}$ whose first row equals μ_2^{d} and who is zero otherwise. Thereby, $\hat{\pi}$ is the unique solution to (H) and hence $(\hat{\pi}, \hat{\mu}_1) \in \mathcal{F} \neq \emptyset$.

We are going to show that $\mathcal{F}$ is compact. Its boundedness comes straight from the linear constraints of (H). To show that it is closed, consider a sequence $(\pi_k, \mu_{1,k}) \subset \mathcal{F}$ with $(\pi_k, \mu_{1,k}) \to (\pi, \mu_1)$ as $k \to \infty$. Passing to the limit in the linear constraints on $\mu_{1,k}$ immediately yields that $\mu_1 \geq 0$ and $\mathbb{1}^\top \mu_1 = 1$. We similarly obtain that $\pi \geq 0$, $\pi \mathbb{1} = \mu_1$, and $\pi^\top \mathbb{1} = \mu_2^{\mathrm{d}}$, i.e., feasibility of π for (H) w.r.t. to the marginals μ_1 and μ_2^{d}. To show its optimality w.r.t. c_{d}, let π^* be an arbitrary solution to (H) w.r.t. μ_1, μ_2^{d}, and c_{d}, which exists due to Lemma 4.2. Then, by Lemma 4.4, there exists a sequence of couplings $(\pi_k^*)_{k \in \mathbb{N}}$ with $\pi_k \in \Pi(\mu_{1,k}, \mu_2^{\mathrm{d}})$ and $\pi_k^* \to \pi^*$. The continuity of (H)'s target function then yields that

$$(\pi, c_{\mathrm{d}})_F = \lim_{k \to \infty} (\pi_k, c_{\mathrm{d}})_F \leq \lim_{k \to \infty} (\pi_k^*, c_{\mathrm{d}})_F = (\pi^*, c_{\mathrm{d}})_F,$$

i.e., the optimality of π for (H) w.r.t. μ_1, μ_2^{d}, and c_{d}. Consequently, $(\pi, \mu_1) \in \mathcal{F}$, which proves the closedness and in turn the compactness of $\mathcal{F}$.

The statement of the theorem then follows because (BH)'s target function $\mathcal{J}$ is lower semicontinuous. □

If we compare (BH) with (BK), we are facing the exact same difficulties:

- (H) is a linear program (short: LP) and can therefore have more than one optimal solution for a given pair of marginals μ_1 and μ_2; this prevents us from using the implicit programming approach which we will introduce in Chapter 5.
- To compute the solution to the subproblems, one in general needs to use an appropriate LP solver like, for example, the simplex method; the solution is directly determinable only in a handful of special cases (e.g. one marginal is a unit vector).
- Because of the curse of dimensionality, the number of unknowns for π is roughly quadratic to the number of unknowns of μ_1 and μ_2.

4.2 Quadratic Regularization of the Hitchcock Problem

To overcome the mentioned difficulties, we sneak a peek at (K_γ) and copy the idea of adding a quadratic regularization term to (H)'s target function. However, unlike in the infinite-dimensional case in Subchapter 3.2, there no need to improve the regularity of the marginals and transport plans in the finite-dimensional setting.

This approach then results in the *(quadratically) regularized Hitchcock problem*:

$$\begin{aligned} \inf_{\pi} \quad & (\pi, c)_F + \tfrac{\gamma}{2} \|\pi\|_F^2 \\ \text{s.t.} \quad & \pi \in \mathbb{R}^{n_1 \times n_2},\ \pi \mathbb{1} = \mu_1,\ \pi^\top \mathbb{1} = \mu_2,\ \pi \geq 0. \end{aligned} \tag{H_γ}$$

In the above, μ_1 and μ_2 are arbitrary but compatible marginals, c is an arbitrary cost matrix, and $\gamma > 0$ is an arbitrary *regularization parameter*.

The goal of this subchapter is to derive results for the regularized Hitchcock problem (H_γ) similar to those we quoted in Subchapter 3.2. Even though we have seen that, by the choice of domains Ω_1 and Ω_2, the Hitchcock problem is just a special case of the Kantorovich problem (this is also true for the corresponding bilevel problems), we do not simply quote the results but prove the desired properties directly, because

1. the corresponding proofs in the finite-dimensional case are typically shorter and less involved compared to the infinite-dimensional case;

2. the proof of the dual representation of the optimal transport plan in Theorem 3.9 heavily relies on the properties of the integral w.r.t. the Lebesgue measure; this becomes evident when comparing its results with Theorem 4.9 below: while the former requires the marginals to have a strictly positive lower bound for the dual representation to hold, the latter can be stated without such an assumption.

Again, (H_γ)'s feasible set is compact and its target function is continuous (and strictly convex). We therefore immediately have the following lemma:

Lemma 4.6. *For each $\gamma > 0$ and for every choice of compatible marginals μ_1 and μ_2 and every cost matrix c, the quadratically regularized Hitchcock problem (H_γ) possesses a unique solution.*

We now aim to obtain a dual representation of the solution to (H_γ), which will later be an essential ingredient for the construction of recovery sequences. To achieve this, however, we must first set up (H_γ)'s optimality system.

Lemma 4.7. *$\pi \in \mathbb{R}^{n_1 \times n_2}$ is the unique solution to (H_γ) if and only if there exist* dual variables *$\alpha_1 \in \mathbb{R}^{n_1}$ and $\alpha_2 \in \mathbb{R}^{n_2}$ such that*

$$\pi \mathbb{1} = \mu_1, \ \pi^\top \mathbb{1} = \mu_2, \ \pi \geq 0, \tag{4.1a}$$

$$c + \gamma\pi - \alpha_1 \oplus \alpha_2 \geq 0, \tag{4.1b}$$

$$(c + \gamma\pi - \alpha_1 \oplus \alpha_2, \pi)_F = 0. \tag{4.1c}$$

Remark 4.8. The operator $\oplus\colon \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_1 \times n}$ from Lemma 4.7 which is defined by

$$u \oplus v \coloneqq (u_{i_1} + v_{i_2})_{(i_1,i_2)\in\Omega} = \begin{pmatrix} u_1 + v_1 & \dots & u_1 + v_{n_2} \\ \vdots & \ddots & \vdots \\ u_{n_1} + v_1 & \dots & u_m + v_{n_2} \end{pmatrix}$$

for all $(u, v) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, refers to the *outer sum* of the vectors u and v. A straightforward calculation shows that its adjoint operator $\oplus^*\colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ is given by $\oplus^*(\theta) \coloneqq (\theta\mathbb{1}, \theta^\top \mathbb{1})$: for all $(u, v) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and $\theta \in \mathbb{R}^{n_1 \times n_2}$ it holds that

$$\begin{aligned} \big((u, v), \oplus^*(\theta)\big)_{\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}} &= \big(\oplus(u, v), \theta\big)_F \\ &= \sum_{(i_1,i_2)\in\Omega} (u_{i_1} + v_{i_2})\theta_{i_1,i_2} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i_1 \in \Omega_1} \Big(u_{i_1} \sum_{i_2 \in \Omega_2} \theta_{i_1,i_2} \Big) + \sum_{i_2 \in \Omega_2} \Big(v_{i_2} \sum_{i_1 \in \Omega_1} \theta_{i_1,i_2} \Big) \\
&= (u, \theta \mathbb{1})_{\mathbb{R}^{n_1}} + \big(v, \theta^\top \mathbb{1}\big)_{\mathbb{R}^{n_2}} \\
&= \big((u,v), (\theta \mathbb{1}, \theta^\top \mathbb{1})\big)_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}}
\end{aligned}
$$

and therefore the claimed representation of $\oplus^*$.

Moreover, the *nonnegative part* and the *nonpositive part* of a matrix θ, which we will need in Theorem 4.9, are defined by

$$
\theta_+ \coloneqq \big(\max\{0, \theta_{i_1,i_2}\}\big)_{(i_1,i_2)\in\Omega} = \begin{pmatrix} \max\{0,\theta_{1,1}\} & \dots & \max\{0,\theta_{1,n_2}\} \\ \vdots & \ddots & \vdots \\ \max\{0,\theta_{n_1,1}\} & \dots & \max\{0,\theta_{n_1,n_2}\} \end{pmatrix}
$$

and

$$
\theta_- \coloneqq \big(-\min\{0, \theta_{i_1,i_2}\}\big)_{(i_1,i_2)\in\Omega} = \begin{pmatrix} -\min\{0,\theta_{1,1}\} & \dots & -\min\{0,\theta_{1,n_2}\} \\ \vdots & \ddots & \vdots \\ -\min\{0,\theta_{n_1,1}\} & \dots & -\min\{0,\theta_{n_1,n_2}\} \end{pmatrix},
$$

respectively. Note that $\theta = \theta_+ - \theta_-$ and $|\theta| = \theta_+ + \theta_-$ for all $\theta \in \mathbb{R}^{n_1 \times n_2}$. ∘

Proof of Lemma 4.7. By reshaping the matrices $\pi \in \mathbb{R}^{n_1 \times n_2}$ and $c \in \mathbb{R}^{n_1 \times n_2}$ into vectors $\vec{\pi} \in \mathbb{R}^{n_1 n_2}$ and $\vec{c} \in \mathbb{R}^{n_1 n_2}$ (stacking π's and c's columns on top of each other in order), respectively, (H_γ) is equivalent to the problem

$$
\begin{aligned}
\inf_{\vec{\pi}} \quad & \vec{\pi}^\top \vec{c} + \tfrac{\gamma}{2} \|\vec{\pi}\|^2_{\mathbb{R}^{n_1 n_2}} \\
\text{s.t.} \quad & \vec{\pi} \in \mathbb{R}^{n_1 n_2}, \ \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \vec{\pi} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \vec{\pi} \geq 0,
\end{aligned}
\tag{$\mathrm{H}_\gamma^{\mathrm{vec}}$}
$$

where

$$
A_1 \coloneqq \begin{pmatrix} 1 & \dots & 0 & 1 & \dots & 0 & & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 & \dots & 1 & & 0 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{n_1 \times n_1 n_2}
$$

and

$$
A_2 \coloneqq \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 & & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & & \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & & 1 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{n_2 \times n_1 n_2}.
$$

The vectorized problem ($\mathrm{H}_\gamma^{\mathrm{vec}}$) is a convex optimization problem in $\mathbb{R}^{n_1 n_2}$. Therefore, according to [53, pp. 382 – 384], $\vec{\pi}$ solves ($\mathrm{H}_\gamma^{\mathrm{vec}}$) if and only if there exist $\lambda \in \mathbb{R}^{n_1 n_2}$ and $\nu \in \mathbb{R}^{n_1 + n_2}$ such that

$$
\begin{aligned}
\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \vec{\pi} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \vec{\pi} \geq 0, \\
\vec{c} + \gamma \vec{\pi} - \lambda + \begin{pmatrix} A_1^\top & A_2^\top \end{pmatrix} \nu = 0,
\end{aligned}
$$

$$\lambda \geq 0,\ \lambda^\top \vec{\pi} = 0,$$

which is, after redefining $\nu \coloneqq (-\alpha_1, -\alpha_2)^\top$ with $\alpha_1 \in \mathbb{R}^{n_1}$ and $\alpha_2 \in \mathbb{R}^{n_2}$, equivalent to

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \vec{\pi} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\ \vec{\pi} \geq 0,$$
$$\vec{c} + \gamma\vec{\pi} - \begin{pmatrix} A_1^\top & A_2^\top \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \geq 0,$$
$$\left(\vec{c} + \gamma\vec{\pi} - \begin{pmatrix} A_1^\top & A_2^\top \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \right)^\top \vec{\pi} = 0.$$

If we convert the above system back to matrix notation, we obtain the system from (4.1). □

Using the optimality system from (4.1), we can now derive a similar dual representation of the solution of (H_γ) to the one that we have already seen for the regularized Kantorovich problem (K_γ) in the infinite-dimensional case.

Theorem 4.9. *π solves (H_γ) w.r.t. μ_1, μ_2, and c if and only if there exist dual variables $\alpha_1 \in \mathbb{R}^{n_1}$ and $\alpha_2 \in \mathbb{R}^{n_2}$ such that*

$$\pi = \frac{1}{\gamma}(\alpha_1 \oplus \alpha_2 - c)_+, \tag{4.2a}$$
$$\sum_{i_2=1}^{n_2} \pi_{i_1,i_2} = \mu_1^{i_1} \quad \text{for all } i_1 = 1, \dots, n_1, \tag{4.2b}$$
$$\sum_{i_1=1}^{n_1} \pi_{i_1,i_2} = \mu_2^{i_2} \quad \text{for all } i_2 = 1, \dots, n_2. \tag{4.2c}$$

Proof. If π solves (H_γ), then by Lemma 4.7 there exist $\alpha_1 \in \mathbb{R}^{n_1}$ and $\alpha_2 \in \mathbb{R}^{n_2}$ such that the system in (4.1) is satisfied. The complementary slackness condition in (4.1c) together with the nonnegativity of π and $c + \gamma\pi - \alpha_1 \oplus \alpha_2$ from (4.1a) and (4.1b), respectively, implies the following: On the one hand, if $\pi_{i_1,i_2} > 0$, then

$$c_{i_1,i_2} + \gamma\pi_{i_1,i_2} - (\alpha_1 \oplus \alpha_2)_{i_1,i_2} = 0 \quad \iff \quad \pi_{i_1,i_2} = \frac{1}{\gamma}\big((\alpha_1 \oplus \alpha_2)_{i_1,i_2} - c_{i_1,i_2}\big)_+.$$

On the other hand, if $\pi_{i_1,i_2} = 0$, then

$$c_{i_1,i_2} - (\alpha_1 \oplus \alpha_2)_{i_1,i_2} \geq 0 \quad \iff \quad (\alpha_1 \oplus \alpha_2)_{i_1,i_2} - c_{i_1,i_2} \leq 0$$

and thus $\gamma\pi_{i_1,i_2} = 0 = \big((\alpha_1 \oplus \alpha_2)_{i_1,i_2} - c_{i_1,i_2}\big)_+$. Because $i_1 \in \Omega_1$ and $i_2 \in \Omega_2$ were arbitrary, all of the above implies (4.2a). (4.2b) and (4.2b) are just the rephrased feasibility constraints from (4.1a).

Now, assume that $\pi \in \mathbb{R}^{n_1 \times n_2}$, $\alpha_1 \in \mathbb{R}^{n_1}$, and $\alpha_2 \in \mathbb{R}^{n_2}$ satisfy the system in (4.2). Because of (4.2b), (4.2c), and the $(\cdot)_+$-operator, π is feasible for (H_γ). We abbreviate (H_γ)'s target function by $f(\pi) \coloneqq (\pi, c)_F + \frac{\gamma}{2}\|\pi\|_F^2$ and consider an arbitrary feasible point $\tilde{\pi}$. Because $f\colon \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$ is convex and differentiable with derivative $f'(\pi) = c + \gamma\pi$, we find that

$$f(\tilde{\pi}) - f(\pi) \geq f'(\pi; \tilde{\pi} - \pi) = (c + \gamma\pi, \tilde{\pi} - \pi)_F, \tag{4.3}$$

see e.g. [58, Theorem 3.8.1]. Taking advantage of $\tilde{\pi}$'s feasibility and $(\alpha_1 \oplus \alpha_2 - c)_- \geq 0$, we estimate that

$$\begin{aligned}(c + \gamma\pi, \tilde{\pi})_F &\geq \big(c + (\alpha_1 \oplus \alpha_2 - c)_+, \tilde{\pi}\big)_F - \big((\alpha_1 \oplus \alpha_2 - c)_-, \tilde{\pi}\big)_F \\ &= (c + \alpha_1 \oplus \alpha_2 - c, \tilde{\pi})_F \\ &= (\alpha_1, \tilde{\pi}\mathbb{1})_{\mathbb{R}^{n_1}} + (\alpha_2, \mathbb{1}^\top \tilde{\pi})_{\mathbb{R}^{n_2}} = (\alpha_1, \mu_1)_{\mathbb{R}^{n_1}} + (\alpha_2, \mu_2)_{\mathbb{R}^{n_2}},\end{aligned} \tag{4.4}$$

Also, for all $A \in \mathbb{R}^{n_1 \times n_2}$ we find that $(A_+, A_+)_F = (A_+, A)_F$. Hence,

$$\begin{aligned}(c + \gamma\pi, \pi)_F &= \big(c + (\alpha_1 \oplus \alpha_2 - c)_+, \pi\big)_F \\ &= (c + \alpha_1 \oplus \alpha_2 - c, \pi)_F \\ &= (\alpha_1, \pi\mathbb{1})_{\mathbb{R}^{n_1}} + (\alpha_2, \mathbb{1}^\top \pi)_{\mathbb{R}^{n_2}} = (\alpha_1, \mu_1)_{\mathbb{R}^{n_1}} + (\alpha_2, \mu_2)_{\mathbb{R}^{n_2}}.\end{aligned} \tag{4.5}$$

Plugging (4.4) and (4.5) into (4.3) then yields that $f(\tilde{\pi}) \geq f(\pi)$ for all feasible $\tilde{\pi}$, i.e., optimality of π for (H_γ). □

We can also characterize, similarly to the authors of [52], the dual problem to (H_γ):

Lemma 4.10. *The* (Lagrangian) dual problem *to* (H_γ) *is equivalent to*

$$\begin{aligned}\sup_{\alpha_1, \alpha_2} \quad & (\alpha_1, \mu_1)_{\mathbb{R}^{n_1}} + (\alpha_2, \mu_2)_{\mathbb{R}^{n_2}} - \tfrac{1}{2\gamma}\|(\alpha_1 \oplus \alpha_2 - c)_+\|_F^2, \\ \textit{s.t.} \quad & \alpha_1 \in \mathbb{R}^{n_1},\ \alpha_2 \in \mathbb{R}^{n_2}.\end{aligned} \tag{HD_γ}$$

For each $\gamma > 0$, (HD_γ) *admits an optimal solution. If (α_1, α_2) is a solution to* (HD_γ)*, then $(\alpha_1 + a, \alpha_2 - a)$, for arbitrary $a \in \mathbb{R}$, is also a solution with the same optimal value. Moreover, there is no duality gap, i.e.,* $\inf(\mathrm{H}_\gamma) = \sup(\mathrm{HD}_\gamma)$.

Proof. The Lagrangian function $\mathcal{L}\colon \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ corresponding to (H_γ) is given by

$$\begin{aligned}\mathcal{L}(\pi, \lambda, \rho_1, \rho_2) \coloneqq (\pi, c)_F + \frac{\gamma}{2}\|\pi\|_F^2 \\ - (\lambda, \pi)_F + (\rho_1, \pi\mathbb{1} - \mu_1)_{\mathbb{R}^{n_1}} + (\rho_2, \pi^\top \mathbb{1} - \mu_2)_{\mathbb{R}^{n_2}},\end{aligned}$$

see e.g. [53, p. 383]. The *Lagrangian dual problem* is then defined to be

$$\sup_{\substack{0 \leq \lambda \in \mathbb{R}^{n_1 \times n_2}, \\ \rho_1 \in \mathbb{R}^{n_1},\, \rho_2 \in \mathbb{R}^{n_2}}} \quad \inf_{\pi \in \mathbb{R}^{n_1 \times n_2}} \quad \mathcal{L}(\pi, \lambda, \rho_1, \rho_2).$$

Setting $(\alpha_1, \alpha_2) \coloneqq (-\rho_1, -\rho_2)$, this can be equivalently written as

$$\sup_{\substack{\alpha_1 \in \mathbb{R}^{n_1}, \\ \alpha_2 \in \mathbb{R}^{n_2}}} \Big((\alpha_1, \mu_1)_{\mathbb{R}^{n_1}} + (\alpha_2, \mu_2)_{\mathbb{R}^{n_2}} + \sup_{\lambda \geq 0} \inf_{\pi} (c + \gamma/2\pi - \lambda - \alpha_1 \oplus \alpha_2, \pi)_F\Big).$$

The inner unconstrained minimization problem is quadratic w.r.t. π and therefore solved by $\pi^* = \frac{1}{\gamma}(\lambda + \alpha_1 \oplus \alpha_2 - c)$. Consequently, the Lagrangian dual problem is equivalent to

$$\sup_{\substack{\alpha_1 \in \mathbb{R}^{n_1}, \\ \alpha_2 \in \mathbb{R}^{n_2}}} \Big((\alpha_1, \mu_1)_{\mathbb{R}^{n_1}} + (\alpha_2, \mu_2)_{\mathbb{R}^{n_2}} - \inf_{\lambda \geq 0} \frac{1}{2\gamma}\|\lambda + \alpha_1 \oplus \alpha_2 - c\|_F^2\Big).$$

Again, the inner problem is solved by $\lambda^* = (\alpha_1 \oplus \alpha_2 - c)_- \geq 0$ and takes the optimal value $\frac{1}{2\gamma}\|(\alpha_1 \oplus \alpha_2 - c)_+\|_F^2$, which shows the equivalence of the Lagrangian dual problem to (HD_γ).

The remaining statements in the formulation of the lemma are either trivial or well-known results of finite-dimensional optimization, see e.g. [53, Strong Duality Theorem, p. 393]. □

The property that the regularized Hitchcock problem (H_γ) admits a unique solution, see Lemma 4.6, implicitly defines the *solution operator*

$$\mathcal{S}_\gamma \colon \mathbb{R}^{n_1 \times n_2} \times \mathcal{M}_0 \to \mathbb{R}^{n_1 \times n_2}, \quad (c, \mu_1, \mu_2) \mapsto \pi, \tag{4.6}$$

where π is unique the solution to (H_γ) w.r.t. the marginals μ_1 and μ_2 as well as the cost matrix c and where $\mathcal{M}_0$ is the *set of compatible marginals*, i.e.,

$$\mathcal{M}_0 \coloneqq \{(\mu_1, \mu_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \colon \mu_1, \mu_2 \geq 0,\ \mathbb{1}^\top \mu_1 = \mathbb{1}^\top \mu_2\}.$$

Similarly to the infinite-dimensional case, we can now replace the lower-level Hitchcock problem in (BH) by the solution operator of its regularized counterpart to receive, for a given regularization parameter $\gamma > 0$, the *regularized (& penalized) bilevel Hitchcock problem*:

$$\begin{aligned} \inf_{\pi, \mu_1, c} \quad & \mathcal{J}_\gamma(\pi, \mu_1, c) \coloneqq \mathcal{J}(\pi, \mu_1) + \tfrac{1}{2\gamma}\|c - c_\mathrm{d}\|_F^2 \\ \text{s.t.} \quad & c \in \mathbb{R}^{n_1 \times n_2},\ \mu_1 \in \mathbb{R}^{n_1},\ \mu_1 \geq 0,\ \mathbb{1}^\top \mu_1 = 1, \\ & \pi = \mathcal{S}_\gamma(c, \mu_1, \mu_2^\mathrm{d}). \end{aligned} \tag{BH_γ}$$

In the above, μ_2^d and c_d are given as in (BH). We will show in the remainder of this chapter that

- this problem admits a solution for any given regularization parameter $\gamma > 0$, see Subchapter 4.3.
- solutions to (BH_γ) can be used to approximate certain solutions of the non-regularized bilevel problem (BH), given that there exists a solution of the non-regularized bilevel problem that is accompanied by a recovery sequence, see Subchapter 4.4.
- there exist at least some cases in which we can construct a recovery sequence directly, see Subchapter 4.5.

4.3 Existence of Solutions to the Regularized Bilevel Hitchcock Problem

In contrast to the infinite-dimensional case, the existence proof of the regularized bilevel problem is much simpler and almost identical to the non-regularized case.

Theorem 4.11. *With the target marginal and the cost function given as in* (BH)*, for arbitrary $\gamma > 0$, there exists at least one optimal solution to* (BH_γ).

Proof. Let us again denote the feasible set of (BH_γ) by $\mathcal{F}$. It is straightforward to see that $\mathcal{F}$ is non-empty: choose $\hat{\mu}_1 = (1, 0, \dots, 0)^\top \in \mathbb{R}^{n_1}$ and $\hat{c} \in \mathbb{R}^{n_1 \times n_2}$ arbitrarily, then Lemma 4.6 yields the existence of a unique $\hat{\pi} = \mathcal{S}_\gamma(\hat{c}, \hat{\mu}_1, \mu_2^{\mathrm{d}})$ and hence $(\hat{\pi}, \hat{\mu}_1, \hat{c}) \in \mathcal{F}$.

With the same arguments as in the proof of Theorem 4.5, we can then show that $\mathcal{F}$ is compact. The assertion of the theorem then follows directly from $\mathcal{J}_\gamma$'s lower semicontinuity.

□

Remark 4.12. Comparing the proofs of Theorem 3.26 and Theorem 4.11, we note that the latter is entirely based on standard arguments and also does not require any preliminaries, as was the case in Subchapter 3.3.

This is essentially because we have shown in the above proof (more precisely, in the proof of Theorem 4.5) that the solution operator of the regularized Hitchcock problem is continuous w.r.t. the first marginal and the cost function.

In contrast, the solution operator of the regularized Kantorovich problem is not continuous with respect to the proper topology, see Example 3.18. We therefore had to rely on the compactness of the smoothing operators $\mathcal{T}_i \colon \mathfrak{M}(\Omega_i) \to L^2(\Omega_i^\delta)$, $i = 1, 2$, and had to prove that the solution operator $\mathcal{S}_\gamma \colon \mathcal{C}_{\underline{\mathrm{c}}}(X) \times \mathcal{M}^m_{\underline{\mu}}(X_1, X_2) \to L^2(X)$ is Hölder continuous, which led to the lengthy preliminary work. ○

4.4 Approximation of Solutions to the Bilevel Hitchcock Problem

This subchapter is devoted to showing that, under certain conditions, we can find sequences of solutions to the regularized bilevel problems (BH_γ) that converge against solutions to the non-regularized bilevel problem (BH).

To that end, we consider an arbitrary vanishing sequence of regularization parameters $(\gamma_k)_{k\in\mathbb{N}} \searrow 0$. Thanks to Theorem 4.11, we can find a sequence of regularized solutions $(\bar{\pi}_k, \bar{\mu}_{1,k}, \bar{c}_k)_{k\in\mathbb{N}}$ to the sequence of problems $(\mathrm{BH}_{\gamma_k})_{k\in\mathbb{N}}$. Note that this sequence need not be unique as the solution to the regularized bilevel problems may not be unique. Nevertheless, each such sequence has at least one cluster point:

For all $k \in \mathbb{N}$, the linear constraints of (BH_{γ_k}) yield that

$$\|\bar{\mu}_{1,k}\|_{\mathbb{R}^{n_1}} \leq \|\bar{\mu}_{1,k}\|_1 = 1 \quad \text{and} \quad \|\bar{\pi}_k\|_F \leq \|\bar{\pi}_k\|_1 = \mathbb{1}^\top \bar{\mu}_{1,k} = 1,$$

i.e., the boundedness of the sequence $(\bar{\pi}_k, \bar{\mu}_{1,k})_{k\in\mathbb{N}}$. Consequently, $(\bar{\pi}_k, \bar{\mu}_{1,k}) \to (\bar{\pi}, \bar{\mu}_1) \in \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1}$ after possibly extracting a subsequence.

Furthermore, given an arbitrary $\tilde{\mu}_1 \in \mathbb{R}^{n_1}$ with $\tilde{\mu}_1 \geq 0$ and $\mathbb{1}^\top \tilde{\mu}_1 = 1$, we consider the sequence of optimal transport plans defined via $\tilde{\pi}_k \coloneqq \mathcal{S}_{\gamma_k}(c_{\mathrm{d}}, \tilde{\mu}_1, \mu_2^{\mathrm{d}})$ for $k \in \mathbb{N}$. By construction, the triple $(\tilde{\pi}_k, \tilde{\mu}_1, c_{\mathrm{d}})$ is feasible for (BH_{γ_k}) for all $k \in \mathbb{N}$ and, owing to $(\bar{\pi}_k, \bar{\mu}_{1,k}, \bar{c}_k)$'s optimality,

$$\mathcal{J}(\bar{\pi}_k, \bar{\mu}_{1,k}) + \frac{1}{\gamma_k}\|\bar{c}_k - c_{\mathrm{d}}\|_F^2 = \mathcal{J}_{\gamma_k}(\bar{\pi}_k, \bar{\mu}_{1,k}, \bar{c}_k) \leq \mathcal{J}_{\gamma_k}(\tilde{\pi}_k, \tilde{\mu}_1, c_{\mathrm{d}}) = \mathcal{J}(\tilde{\pi}_k, \tilde{\mu}_1).$$

Thus and because $\mathcal{J}$ is bounded on both of the bounded sets $\{(\tilde{\pi}_k, \tilde{\mu}_1)\}_{k\in\mathbb{N}}$ and $\{(\bar{\pi}_k, \bar{\mu}_{1,k})\}_{k\in\mathbb{N}}$,

$$\|\bar{c}_k - c_{\mathrm{d}}\|_F \leq \gamma_k^{\frac{1}{2}} \big(\mathcal{J}(\tilde{\pi}_k, \tilde{\mu}_1) - \mathcal{J}(\bar{\pi}_k, \bar{\mu}_{1,k})\big)^{\frac{1}{2}} \leq \gamma_k^{\frac{1}{2}} C,$$

for some constant $C > 0$. This shows that $\bar{c}_k \to c_\mathrm{d}$ as $k \to \infty$. To summarize, we have found a cluster point $(\bar{\pi}, \bar{\mu}_1, c_\mathrm{d}) \in \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_1 \times n_2}$ so that

$$(\bar{\pi}_k, \bar{\mu}_{1,k}, \bar{c}_k) \xrightarrow[k\to\infty]{} (\bar{\pi}, \bar{\mu}_1, c_\mathrm{d}),$$

after possibly extracting a subsequence that we denote by the same symbol.

Compared to the infinite-dimensional setting from Chapter 3, it takes much less effort in the finite-dimensional setting to show the feasibility of the cluster point $(\bar{\pi}, \bar{\mu}_1)$ for (BH). The proof follows the same reasoning as the proof of Lemma 3.33, but is at the same time much less technical since we do not have to deal with smoothed marginals and their properties. Besides, the main difficulty has already been provided by the proof of Lemma 4.4.

Lemma 4.13. *The cluster point $(\bar{\pi}, \bar{\mu}_1)$ of the sequence of regularized solutions $(\bar{\pi}_k, \bar{\mu}_{1,k})_{k\in\mathbb{N}}$ is feasible for the non-regularized bilevel problem (BH), i.e., $\bar{\mu}_1 \geq 0$ and $\mathbb{1}^\top \bar{\mu}_1 = 1$ and $\bar{\pi}$ is an optimal coupling between $\bar{\mu}_1$ and μ_2^d w.r.t. c_d.*

Proof. As before, the feasibility of $\bar{\mu}_1$ for (BH) is clear due to the linearity of the corresponding constraints. That $\bar{\pi}$ is feasible for (H) w.r.t. to the marginals $\bar{\mu}_1$ and μ_2^d follows directly from passing to the limit in the linear equations

$$\bar{\pi}_k \mathbb{1} = \bar{\mu}_{1,k} \quad \text{and} \quad \bar{\pi}_k^\top \mathbb{1} = \mu_2^\mathrm{d} \quad \text{for all } k \in \mathbb{N}.$$

Let π^* once again be an optimal solution to (H) w.r.t. $\bar{\mu}_1$, μ_2^d, and c_d. By Lemma 4.4, there exists a sequence $(\pi_k^*)_{k\in\mathbb{N}}$ where, for all $k \in \mathbb{N}$, π_k^* is a nonnegative coupling between $\bar{\mu}_{1,k}$ and μ_2^d as well as $\pi_k^* \to \pi^*$ as $k \to \infty$. In particular, π_k^* is feasible for (H_{γ_k}) w.r.t. the right marginals. The optimality of $\bar{\pi}_k$ for (H_γ) w.r.t. $\bar{\mu}_{1,k}$, μ_2^d, and $\bar{c}_k$ then implies that

$$\begin{aligned}(c_\mathrm{d}, \pi^*)_F \leq (c_\mathrm{d}, \bar{\pi})_F &= \lim_{k\to\infty} \Big((\bar{c}_k, \bar{\pi}_k)_F + \frac{\gamma_k}{2} \|\bar{\pi}_k\|_F^2 \Big) \\ &\leq \lim_{k\to\infty} \Big((\bar{c}_k, \pi_k^*)_F + \frac{\gamma_k}{2} \|\pi_k^*\|_F^2 \Big) = (c_\mathrm{d}, \pi^*)_F,\end{aligned}$$

i.e., $(c_\mathrm{d}, \pi^*)_F = (c_\mathrm{d}, \bar{\pi})_F$. Consequently, $\bar{\pi}$ is optimal for (H) w.r.t. $\bar{\mu}_1$, μ_2^d, and c_d and the cluster point $(\bar{\pi}, \bar{\mu}_1)$ therefore feasible for (BH). □

If we presuppose the existence of a recovery sequence, we can show that the cluster point $(\bar{\pi}, \bar{\mu}_1)$ of the sequence of regularized solutions $(\bar{\pi}_k, \bar{\mu}_{1,k})_{k\in\mathbb{N}}$ is optimal for the non-regularized bilevel problem.

Theorem 4.14. *Let (π^*, μ_1^*) be a solution to the bilevel problem (BH) that is accompanied by a* recovery sequence*, i.e., a sequence $(\pi_k^*, \mu_{1,k}^*, c_k^*)_{k\in\mathbb{N}} \subset \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_1 \times n_2}$ such that*

1. *$(\pi_k^*, \mu_{1,k}^*, c_k^*)$ is feasible for (BH_{γ_k}) for all $k \in \mathbb{N}$,*
2. *$\limsup_{k\to\infty} \mathcal{J}_{\gamma_k}(\pi_k^*, \mu_{1,k}^*, c_k^*) \leq \mathcal{J}(\pi^*, \mu_1^*)$.*

Then, the cluster point $(\bar{\pi}, \bar{\mu}_1)$ of the sequence of regularized solutions $(\bar{\pi}_k, \bar{\mu}_{1,k})_{k\in\mathbb{N}}$ is also a solution to (BH).

Proof. This proof is just a brazen copy of the proof of Theorem 3.34 in the infinite-dimensional case:

With a slight abuse of notation, we denote by $(\bar{\pi}_k, \bar{\mu}_{1,k})_{k\in\mathbb{N}}$ the subsequence that converges towards $(\bar{\pi}, \bar{\mu}_1)$. Because of the presupposed lower semicontinuity of the target function,

$$\begin{aligned}\mathcal{J}(\bar{\pi}, \bar{\mu}_1) &\leq \liminf_{k\to\infty} \mathcal{J}(\bar{\pi}_k, \bar{\mu}_{1,k}) \\ &\leq \liminf_{k\to\infty} \mathcal{J}(\bar{\pi}_k, \bar{\mu}_{1,k}) + \frac{1}{\gamma_k}\|\bar{c}_k - c_\mathrm{d}\|_F^2 = \liminf_{k\to\infty} \mathcal{J}_{\gamma_k}(\bar{\pi}_k, \bar{\mu}_{1,k}, \bar{c}_k).\end{aligned}$$

Because of the optimality of $(\bar{\pi}_k, \bar{\mu}_{1,k}, \bar{c}_k)$ for (BH_{γ_k}),

$$\mathcal{J}(\bar{\pi}, \bar{\mu}_1) \leq \liminf_{k\to\infty} \mathcal{J}_{\gamma_k}(\bar{\pi}_k, \bar{\mu}_{1,k}, \bar{c}_k) \leq \limsup_{k\to\infty} \mathcal{J}_{\gamma_k}(\pi_k^*, \mu_{1,k}^*, c_k^*) \leq \mathcal{J}(\pi^*, \mu_1^*).$$

Thanks to Lemma 4.13, $(\bar{\pi}, \bar{\mu}_1)$ is feasible for (BH) and, because of (π^*, μ_1^*)'s optimality, also optimal. □

Similar to the infinite-dimensional case, we want to emphasize that

1. the existence of a recovery sequence implies the optimality of **every** cluster point of the sequence $(\bar{\pi}_k, \bar{\mu}_{1,k}, \bar{c}_k)_{k\in\mathbb{N}}$ (since there may be more than one);
2. the assumption of finding an optimum that is accompanied by a recovery sequence is relatively strong and may be not satisfied in the general case. Nevertheless, we are able to explicitly construct a recovery sequence in a setting that is slightly more general than what was described in Corollary 3.41. This will be the topic of the following subchapter.

4.5 Existence of a Recovery Sequence for the Bilevel Hitchcock Problem

As already indicated in the previous subchapter, the purpose of the present subchapter is to explicitly construct a recovery sequence in the following setting:

Assumption 4.15. For the entire subchapter, we assume that $n_2 \geq n_1 \geq 2$ and that (BH)'s cost matrix $c_\mathrm{d} \in \mathbb{R}^{n_1\times n_2}$ takes the form $(c_\mathrm{d})_{i_1,i_2} = |i_1 - i_2|^\rho$ with $\rho \geq 1$ for all $(i_1, i_2) \in \Omega = \{1, \dots, n_1\} \times \{1, \dots, n_2\}$. Moreover, we assume that there exists a solution to (BH), namely (π^*, μ_1^*), in a way that there exists a monotone *assignment function* $j^*\colon \Omega_1 \to \Omega_2$ with $j^*(1) = 1$ and

$$\pi^*_{i_1,i_2} \begin{cases} \geq 0, & \text{if } i_2 = j^*(i_1), \\ = 0, & \text{if } i_2 \neq j^*(i_1), \end{cases} \tag{4.7}$$

for all $(i_1, i_2) \in \Omega$.

Remark 4.16. 1. If $n_1 = 1$ or $n_2 = 1$, then the Hitchcock problem (H) would possess only trivial solutions. Consequently, the regularization approach would be pointless and the bilevel problem (BH) would also be of no interest. Further, the assumption $n_2 \geq n_1$ serves to avoid additional case distinctions in the subsequent analysis. One can always consider the reverse case by exploiting the symmetry of the optimal transport problem, see Lemma C.2.

2. In particular the case $\rho = 1$ often leads to nonunique solutions of the Hitchcock problem[1], making the arguments from the proof of Corollary 3.41, which relies heavily on the uniqueness of the optimal transport plan, invalid.

3. Illustratively, the condition in (4.7) states that π^* shall be a sparse matrix whose nonzero entries are subject to a monotonic ordering. This is related to Brenier's theorem in the infinite-dimensional case, see e.g. [76, Theorem 2.12 (ii)], which states that (under certain conditions on the data) there exists an optimal transport plan which is supported on the graph of a monotone function.

4. The assumption that $j^*(1) = 1$ is only for convenience. The arguments in this subchapter apply even without this normalization, but additional case distinctions would then have to be made, which would make the following (already nontrivial) calculations even more opaque.

○

In contrast to the infinite-dimensional case, we will now exploit the fact that the cost variable is an optimization variable of (BH_γ). The main advantage is that this allows us to construct a recovery sequence where the transportation plans and first marginals are constant. We do so by simply hiding the γ-dependent parts in the cost variable, as the proof of the following lemma shows.

Lemma 4.17. *If there exists a vector $b \in \mathbb{R}^{n_2}$ satisfying the system*

$$b_{i_2} - b_{j^*(i_1)} \leq (c_\mathrm{d})_{i_1,i_2} - (c_\mathrm{d})_{i_1,j^*(i_1)} \tag{4.8}$$

for all $i_1 \in \Omega_1$ and $i_2 \in \Omega_2 \setminus \{j^*(i_1)\}$, then there exists a recovery sequence $(\pi_k^*, \mu_{1,k}^*, c_k^*)_{k\in\mathbb{N}}$ accompanying the solution (π^*, μ_1^*) in the sense of Theorem 4.14.

Proof. Given the vanishing sequence of regularization parameters $(\gamma_k)_{k\in\mathbb{N}}$ from Subchapter 4.4 as well as c_d and π^* from Assumption 4.15, we define $c_k^* \coloneqq c_\mathrm{d} - \gamma_k \pi^* \in \mathbb{R}^{n_1 \times n_2}$. Moreover, we set $\alpha_2 \coloneqq b$ and define $\alpha_1 \in \mathbb{R}^{n_1}$ by

$$\alpha_1^{i_1} \coloneqq -\alpha_2^{j^*(i_1)} + (c_\mathrm{d})_{i_1,j^*(i_1)} \quad \text{for all } i_1 \in \Omega_1.$$

By construction, α_1 and α_2 satisfy

$$\begin{aligned} \alpha_1^{i_1} + \alpha_2^{j^*(i_1)} &= (c_\mathrm{d})_{i_1,j^*(i_1)} = (c_k^*)_{i_1,j^*(i_1)} + \gamma_k \pi^*_{i_1,j^*(i_1)} \\ \alpha_1^{i_1} + \alpha_2^{i_2} &= b_{i_2} - b_{j^*(i_1)} + (c_\mathrm{d})_{i_1,j^*(i_1)} \leq (c_k^*)_{i_1,i_2} \end{aligned}$$

for all $i_1 \in \Omega_1$ and $i_2 \in \Omega_2 \setminus \{j^*(i_1)\}$, where we used (4.8) for the inequalities. We then define

$$\pi_k^* \coloneqq \frac{1}{\gamma_k}(\alpha_1 \oplus \alpha_2 - c_k^*)_+.$$

[1] Imagine having a bookshelf with four compartments, where the first three compartments are each occupied by a book, and wanting to free the first compartment. If we measure the effort in metric costs, i.e., we are solely interested in the total distance covered when rearranging the books, then it makes no difference whether we move each book one position to the right or whether we place the first book directly in the last shelf and leave the others untouched.

By construction, $\pi_k^* = \pi^*$ for all $k \in \mathbb{N}$. In light of Theorem 4.9, π_k^* is the unique optimal solution of (H_{γ_k}) w.r.t. the marginals μ_1^* and μ_2^{d} and the cost c_k^*. Thus, $(\pi_k^*, \mu_{1,k}^*, c_k^*)$ with $\mu_{1,k}^* = \mu_1^*$ is feasible for (BH_{γ_k}) for all $k \in \mathbb{N}$.

Because the sequence $(\pi_k^*, \mu_{1,k}^*)$ is constant and because of $c_k^* \to c_{\mathrm{d}}$, it holds that

$$\limsup_{k\to\infty} \mathcal{J}_{\gamma_k}(\pi_k^*, \mu_{1,k}^*, c_k^*) = \lim_{k\to\infty} \mathcal{J}(\pi^*, \mu_1^*) + \frac{1}{2\gamma_k}\|c_k^* - c_{\mathrm{d}}\|_F^2 = \mathcal{J}(\pi^*, \mu_1^*).$$

Hence, $(\pi_k^*, \mu_{1,k}^*, c_k^*)$ is a recovery sequence in the sense of Theorem 4.14. □

We now aim to reformulate the system from (4.8) to make it more handable. To this end, we need the following definitions.

Definition 4.18. For $m, n \in \mathbb{N}$ with $n \geq m \geq 2$, consider a monotone assignment function $j\colon \{1, \dots, m\} \to \{1, \dots, n\}$ with $j(1) = 1$ and some cost matrix $\bar{c} \in \mathbb{R}^{m\times n}$. We define the corresponding *reduced system matrix* by

$$A \coloneqq \begin{pmatrix} E_{j(1)} \\ \vdots \\ E_{j(m)} \end{pmatrix} \in \mathbb{R}^{(m(n-1))\times n},$$

where, for $l \in \{1, \dots, n\}$,

$$E_l \coloneqq \begin{pmatrix} e_1 \ \dots \ e_{l-1} & \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix} & e_l \ \dots \ e_{n-1} \end{pmatrix} \in \mathbb{R}^{(n-1)\times n}.$$

In the above, $e_1, \dots, e_{n-1}$ denote the unit vectors of $\mathbb{R}^{n-1}$. Moreover, we define the *reduced cost vector* corresponding to j and $\bar{c}$ by

$$c \coloneqq \begin{pmatrix} \left(\bar{c}_{1,l_1^1} - \bar{c}_{1,j(1)}, \ \dots, \ \bar{c}_{1,l_1^{n-1}} - \bar{c}_{1,j(1)}\right)^\top \\ \vdots \\ \left(\bar{c}_{m,l_m^1} - \bar{c}_{m,j(m)}, \ \dots, \ \bar{c}_{m,l_m^{n-1}} - \bar{c}_{m,j(m)}\right)^\top \end{pmatrix} \in \mathbb{R}^{m(n-1)},$$

where for $l_{i_1}^1, \dots, l_{i_1}^{n-1} \in \{1, \dots, n\} \setminus \{j(i_1)\}$, $i_1 \in \{1, \dots, m\}$, it holds that $l_{i_1}^1 < \dots < l_{i_1}^{n-1}$.

Remark 4.19. By construction, the reduced cost vector c associated with some assignment function j and cost matrix $\bar{c}$ satisfies

$$c_{(i_1-1)(n-1)+i_2} = \begin{cases} \bar{c}_{i_1,i_2} - \bar{c}_{i,j(i_1)}, & \text{if } i_2 < j(i_1), \\ \bar{c}_{i_1,i_2+1} - \bar{c}_{i_1,j(i_1)}, & \text{if } i_2 \geq j(i_1), \end{cases} \tag{4.9}$$

for all $i_1 \in \{1, \dots, m\}$ and $i_2 \in \{1, \dots, n-1\}$. ○

Let A^* and c^* be the reduced system matrix and the reduced cost vector corresponding to j^* and c_{d}, respectively. We then find that the system from (4.8) is equivalent to $A^* b \leq c^*$. Our strategy for proving the existence of a vector $b^* \in \mathbb{R}^{n_2}$ that satisfies $A^* b^* \leq c^*$ and therefore (4.8) includes the following induction argument:

We start from the (simplest possible) assignment function j_0 (with A_0 and c_0 being the associated reduced system matrix and reduced cost function, respectively), show that the system $A_0 b \leq c_0$ has a solution b_0, see Example 4.22, and then prove that small (but significant) changes in the assignment function do not affect the solvability of the corresponding linear inequality system, see Lemma 4.23 and Lemma 4.24. This is then sufficient to show that, after multiple applications of the aforementioned lemmas, the system $A^* b \leq c^*$ also admits a solution b^*, see Theorem 4.25.

For the outlined chain of arguments, the following version of Farka's lemma will be useful, as it provides the lever with which we can prove the existence of solutions for a system $Ab \leq c$.

Lemma 4.20 ([53, p. 34])**.** *For $m, n \in \mathbb{N}$, some matrix $A \in \mathbb{R}^{m \times n}$, as well as a vector $c \in \mathbb{R}^m$, there exists a solution $b \in \mathbb{R}^n$ to the linear inequality system $Ab \leq c$ if and only if for all $d \geq 0$ with $A^\top d = 0$ it holds that $d^\top c \geq 0$.*

Shortly, we will give an example to illustrate the inductive argument that we mentioned above. First, however, we will prove the following lemma, which will be used several times in the remainder of this subchapter.

Lemma 4.21. *For $m, n \in \mathbb{N}$ and $\rho \geq 1$, consider the cost matrix defined by $\bar{c}_{i_1,i_2} \coloneqq |i_1 - i_2|^\rho$ for all $i_1 \in \{1, \dots, m\}$ and $i_2 \in \{1, \dots, n\}$. Further, fix some $N \in \{1, \dots, n\}$. Then it holds that*

$$\bar{c}_{1,i_2} - \bar{c}_{1,N} \leq \bar{c}_{2,i_2} - \bar{c}_{2,N} \leq \dots \leq \bar{c}_{m,i_2} - \bar{c}_{m,N} \tag{4.10}$$

for all $i_2 \in \{1, \dots, N-1\}$ and

$$\bar{c}_{1,i_2} - \bar{c}_{1,N} \geq \bar{c}_{2,i_2} - \bar{c}_{2,N} \geq \dots \geq \bar{c}_{m,i_2} - \bar{c}_{m,N} \tag{4.11}$$

for all $i_2 \in \{N+1, \dots, n\}$.

Proof. Let $\rho > 1$ and abbreviate $f(x) \coloneqq |x - i_2|^\rho - |x - N|^\rho$. Then, f is differentiable with $f'(x) = \rho\big(\operatorname{sgn}(x - i_2)|x - i_2|^{\rho-1} - \operatorname{sgn}(x - N)|x - N|^{\rho-1}\big)$ and, by the mean value theorem, $f(x) - f(a) = f'(\xi)(x - a)$ for all $x, a \in \mathbb{R}$ and some $\xi \in (a, x)$.

Now, let $i_2 \in \{1, \dots, N-1\}$ and $i_1 \in \{1, \dots, m-1\}$ be arbitrary. Then there exists some $\xi \in (i_1, i_1 + 1)$ such that

$$\begin{aligned} &\bar{c}_{i_1+1,i_2} - \bar{c}_{i_1+1,N} - \big(\bar{c}_{i_1,i_2} - \bar{c}_{i_1,N}\big) \\ &\quad = f(i_1 + 1) - f(i_1) \\ &\quad = f'(\xi) = \rho\big(\operatorname{sgn}(\xi - i_2)|\xi - i_2|^{\rho-1} - \operatorname{sgn}(\xi - N)|\xi - N|^{\rho-1}\big). \end{aligned}$$

If $i_1 \in \{1, \dots, i_2 - 1\}$, then $\xi < i_1 + 1 \leq i_2 < N$ and consequently

$$\bar{c}_{i_1+1,i_2} - \bar{c}_{i_1+1,N} - \big(\bar{c}_{i_1,i_2} - \bar{c}_{i_1,N}\big) = \rho\big((N - \xi)^{\rho-1} - (i_2 - \xi)^{\rho-1}\big) > 0,$$

where the estimate holds because the mapping $x \mapsto x^{\rho-1}$ is increasing for non-negative values of x and $0 < i_2 - \xi < N - \xi$. If $i_2 \in \{i_2, \dots, N-1\}$, then $i_2 \leq i_1 < \xi$ as well as $\xi < i_1 + 1 \leq N$ and consequently

$$\bar{c}_{i_1+1,i_2} - \bar{c}_{i_1+1,N} - \big(\bar{c}_{i_1,i_2} - \bar{c}_{i_1,N}\big) = \rho\big(|\xi - i_2|^{\rho-1} + |\xi - N|^{\rho-1}\big) > 0.$$

If $i_1 \in \{N, \dots, m-1\}$, then $\xi > i_1 \geq N > i_2$ and consequently

$$\bar{c}_{i_1+1,i_2} - \bar{c}_{i_1+1,N} - \big(\bar{c}_{i_1,i_2} - \bar{c}_{i_1,N}\big) = \rho\big((\xi - i_2)^{\rho-1} - (\xi - N)^{\rho-1}\big) > 0,$$

again because of monotonicity and $\xi - i_2 > \xi - N > 0$. This proves (4.10). The proof of (4.11) is analogous.

The case of $\rho = 1$ can be proven with the same distinction of cases w.r.t. i_1, but without using the mean value theorem. □

Example 4.22. For $m, n \in \mathbb{N}$ with $n \geq m \geq 2$ and $\rho \geq 1$, we consider the cost matrix given by $\bar{c}_{i_1,i_2} = |i_1 - i_2|^\rho$, for all $i_1 \in \{1, \dots, m\}$ and $i_2 \in \{1, \dots, n\}$ as well as the monotone assignment function $j_0(1) = \cdots = j_0(m) = 1$, the latter of which corresponds to a matrix of the form

$$\pi_0 = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ p_m & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{m \times n},$$

with $p_1, \dots, p_m \geq 0$. The definition of the reduced system matrix A_0 from Definition 4.18 implies that $d_0 \in \mathbb{R}^{m(n-1)}$ solves $A_0^\top d_0 = 0$ if and only if

$$\begin{pmatrix} -1 & \dots & -1 & -1 & \dots & -1 & & -1 & \dots & -1 \\ 1 & \dots & 0 & 1 & \dots & 0 & & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 & \dots & 1 & & 0 & \dots & 1 \end{pmatrix} d_0 = 0$$

$$\iff \begin{pmatrix} 1 & \dots & 0 & 1 & \dots & 0 & & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & 0 & \dots & 1 & & 0 & \dots & 1 \end{pmatrix} d_0 = 0.$$

Consequently, $A_0^\top d_0 = 0$ if and only if

$$\sum_{i_1=1}^{m} d_0^{(i_1-1)(n-1)+i_2} = 0 \quad \text{for all } i_2 \in \{1, \dots, n-1\}.$$

Now, if $d_0 \geq 0$, then the above implies that $d_0 = 0$ and thus $d_0^\top c_0 = 0$, with c_0 being the reduced cost vector corresponding to j_0 and $\bar{c}$. Lemma 4.20 therefore ensures that there exists a solution $b_0 \in \mathbb{R}^n$ to the system $A_0 b \leq c_0$.

We now want to prove the same property for the system $A_1 b \leq c_1$, where A_1 and c_1 are associated to the assignment function j_1 (and to the cost matrix $\bar{c}$) with

$$j_1(1) = \cdots = j_1(m-1) = 1 \quad \text{and} \quad j_1(m) = 2,$$

i.e., j_1 corresponds to a matrix $\pi_1 \in \mathbb{R}^{m \times n}$ which looks very similar to π_0:

$$\pi_1 = \begin{pmatrix} p_1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{m-1} & 0 & 0 & \dots & 0 \\ 0 & p_m & 0 & \dots & 0 \end{pmatrix}.$$

Obviously, π_1 results from π_0 by shifting the nonzero element in π_0's last row one column to the right.

We observe that $d_1 \in \mathbb{R}^{m(n-1)}$ solves $A_1^\top d_1 = 0$ if and only if

$$\begin{pmatrix} -1 & \dots & \dots & -1 & & -1 & \dots & \dots & -1 & 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & & 1 & 0 & \dots & 0 & -1 & \dots & \dots & -1 \\ 0 & 1 & \dots & 0 & \dots & 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \end{pmatrix} d_1 = 0$$

$$\iff \begin{pmatrix} 1 & 0 & \dots & 0 & & 1 & 0 & \dots & 0 & -1 & \dots & \dots & -1 \\ 0 & 1 & \dots & 0 & & 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \end{pmatrix} d_1 = 0.$$

Consequently, $A_1^\top d_1 = 0$ if and only if

$$d_1^1 = \sum_{i_1=2}^{m-1} -d_1^{(i_1-1)(n-1)+1} + \sum_{i_2=1}^{n-1} d_1^{(m-1)(n-1)+i_2}$$

and

$$\sum_{i_1=1}^{m} d_1^{(i_1-1)(n-1)+i_2} = 0 \quad \text{for all } i_2 \in \{2, \dots, n-1\}.$$

Now, if $d_1 \geq 0$, then the above implies that

$$d_1^{(i_1-1)(n-1)+i_2} = 0 \quad \text{for all } i_1 \in \{1, \dots, m\} \text{ and } i_2 \in \{2, \dots, n-1\}. \tag{4.12}$$

Thus,

$$d_1^1 = \sum_{i_1=2}^{m-1} -d_1^{(i_1-1)(n-1)+1} + d_1^{(m-1)(n-1)+1}. \tag{4.13}$$

Using (4.12), (4.13), and $c_1^1 = \bar{c}_{1,2} - \bar{c}_{1,1} = 1 > 0$, we are able to estimate the scalar product of d_1 and c_1 by

$$\begin{aligned} d_1^\top c_1 &= d_1^1 c_1^1 + \sum_{i_1=2}^{m-1} d_1^{(i_1-1)(n-1)+1} c_1^{(i_1-1)(n-1)+1} \\ &\quad + d_1^{(m-1)(n-1)+1} c_1^{(m-1)(n-1)+1} \\ &\geq \min_{i_1 \in \{2,\dots,m-1\}} c_1^{(i_1-1)(n-1)+1} \cdot \sum_{i_1=2}^{m-1} d_1^{(i_1-1)(n-1)+1} \\ &\quad + d_1^{(m-1)(n-1)+1} c_1^{(m-1)(n-1)+1} \\ &\geq \Big(\min_{i_1 \in \{2,\dots,m-1\}} c_1^{(i_1-1)(n-1)+1} + c_1^{(m-1)(n-1)+1} \Big) \\ &\quad \cdot \sum_{i_1=2}^{m-1} d_1^{(i_1-1)(n-1)+1}. \end{aligned}$$

We use Remark 4.19 and Lemma 4.21 (with $N = 1$ and $i_2 = 2$) to see that

$$c_1^{(i_1-1)(n-1)+1} + c_1^{(m-1)(n-1)+1} = \bar{c}_{i_1,2} - \bar{c}_{i_1,1} - \left(\bar{c}_{m,2} - \bar{c}_{m,1}\right) \geq 0$$

for all $i_1 \in \{2, \dots, m-1\}$. Consequently, $d_1^\top c_1 \geq 0$. Again, Lemma 4.20 ensures the existence of some vector $b_1 \in \mathbb{R}^n$ with $A_1 b_1 \leq c_1$. ◊

The above example demonstrates two properties that are important for the analysis of this subchapter:

1. the solvability of the system $Ab \leq c$ does not depend on the actual value of the nonzero entries of the matrix π, but only on their positions (encoded by j) within the matrix;
2. if we are given a monotone assignment function j whose corresponding linear system $Ab \leq c$ admits a solution, we can increase $j(m)$ and the resulting system will still admit a solution. This observation is made rigorous in the following lemma.

Lemma 4.23. *Let $m, n \in \mathbb{N}$ with $n \geq m \geq 2$ be given and consider the cost matrix $\bar{c}$ from Example 4.22. For $p \in \{0, 1\}$, consider the monotone assignment functions $j_p \colon \{1, \dots, m\} \to \{1, \dots, n\}$ and denote their associated reduced system matrices and reduced cost vectors by A_p and c_p, respectively.*

Assume that $j_0(1) = 1$ as well as $N \coloneqq j_0(m) < n$ and, moreover, that $j_1|_{\{1,\dots,m-1\}} \equiv j_0|_{\{1,\dots,m-1\}}$ as well as $j_1(m) = j_0(m) + 1$.

Then, if the system $A_0 b \leq c_0$ has a solution, so does the system $A_1 b \leq c_1$.

Proof. We have already examined the case $N = 1$ in the previous example. Therefore, we assume that $N \geq 2$. Given $d_1 \in \mathbb{R}^{m(n-1)}$, an arbitrary non-negative solution of $A_1^\top d = 0$, we then define the vector $d_0 \in \mathbb{R}^{m(n-1)}$ via

$$d_0^{(i_1-1)(n-1)+i_2} \coloneqq \begin{cases} d_1^{(i_1-1)(n-1)+N-1} + d_1^{(i_1-1)(n-1)+N}, & \text{if } i_1 \notin j_0^{-1}(N),\ i_2 = N-1, \\ 0, & \text{if } i_2 = N, \\ d_1^{(i_1-1)(n-1)+i_2}, & \text{else,} \end{cases} \tag{4.14}$$

for all $i_1 \in \{1, \dots, m\}$ and $i_2 \in \{1, \dots, n-1\}$. We will show in the following that $A_0^\top d_0 = 0$.

By construction, $d_0 \geq 0$ and

$$\sum_{i_2=1}^{n-1} d_0^{(i_1-1)(n-1)+i_2} = \sum_{i_2=1}^{n-1} d_1^{(i_1-1)(n-1)+i_2} \quad \text{for all } i_1 \notin j_0^{-1}(N). \tag{4.15}$$

The structure of the reduced system matrices A_p, $p \in \{0, 1\}$, yields that

$$\begin{aligned} \left(A_p^\top d_p\right)_l = &\sum_{i_1 \colon j_p(i_1) < l} d_p^{(i_1-1)(n-1)+l-1} \\ &+ \sum_{i_1 \colon j_p(i_1) = l} \sum_{i_2=1}^{n-1} -d_p^{(i_1-1)(n-1)+i_2} \\ &+ \sum_{i_1 \colon j_p(i_1) > l} d_p^{(i_1-1)(n-1)+l} \end{aligned} \tag{4.16}$$

for all $l \in \{1,\dots,n\}$. Using the definition of j_1, we observe that

$$\begin{aligned}\{i_1 \colon j_0(i_1) < l\} &= \{i_1 \colon j_1(i_1) < l\},\\ \{i_1 \colon j_0(i_1) = l\} &= \{i_1 \colon j_1(i_1) = l\}, \qquad \text{for all } l \in \{1,\dots,N-1\}.\\ \{i_1 \colon j_0(i_1) > l\} &= \{i_1 \colon j_1(i_1) > l\},\end{aligned} \tag{4.17}$$

If $l \in \{1,\dots,N-2\}$, then $j_0^{-1}(l) \cap j_0^{-1}(N) = \emptyset$. Hence, we use (4.14) – (4.17) to find that

$$\begin{aligned}(A_0^\top d_0)_l = &\sum_{i_1 \colon j_1(i_1) < l} d_1^{(i_1-1)(n-1)+l-1}\\ &+ \sum_{i_1 \colon j_1(i_1) = l} \sum_{i_2=1}^{n-1} -d_1^{(i_1-1)(n-1)+i_2}\\ &+ \sum_{i_1 \colon j_1(i_1) > l} d_1^{(i_1-1)(n-1)+l} = (A_1^\top d_1)_l = 0,\end{aligned}$$

where the last equality follows from the assumption on d_1. Moreover, we find that $\big(\{i_1 \colon j_0(i_1) < N-1\} \cup \{i_1 \colon j_0(i_1) = N-1\}\big) \cap j_0^{-1}(N) = \emptyset$ and $\{i_1 \colon j_0(i_1) > N-1\} = \{i_1 \colon j_0(i_1) \geq N\} = j_0^{-1}(N)$, because j_0 is monotone and $N = j_0(m)$. Similarly to before, we use (4.16) and (4.17) for $l = N-1$ together with the definition of d_0 to obtain that

$$\begin{aligned}(A_0^\top d_0)_{N-1} = &\sum_{i_1 \colon j_1(i_1) < N-1} d_1^{(i_1-1)(n-1)+N-2}\\ &+ \sum_{i_1 \colon j_1(i_1) = N-1} \sum_{i_2=1}^{n-1} -d_1^{(i_1-1)(n-1)+i_2}\\ &+ \sum_{i_1 \colon j_1(i_1) > N-1} d_1^{(i_1-1)(n-1)+N-1} = (A_1^\top d_1)_{N-1} = 0.\end{aligned}$$

By the properties of j_0, it holds that $\{1,\dots,m\} = \{i_1 \colon j_0(i_1) < N\} \cup j_0^{-1}(N)$ and therefore $\{i_1 \colon j_0(i_1) > N\} = \emptyset$. Thus, for $l = N$, we calculate that

$$\begin{aligned}&(A_0^\top d_0)_N\\ &= \sum_{i_1 \notin j_0^{-1}(N)} \big(d_1^{(i_1-1)(n-1)+N-1} + d_1^{(i_1-1)(n-1)+N}\big)\\ &\quad + \sum_{i_1 \in j_0^{-1}(N)} \sum_{i_2=1}^{n-1} -d_1^{(i_1-1)(n-1)+i_2}\\ &= d_1^{N-1} + d_1^N + \sum_{i_1 \in \{2,\dots,m\} \setminus \{j_0^{-1}(N)\}} \big(d_1^{(i_1-1)(n-1)+N-1} + d_1^{(i_1-1)(n-1)+N}\big)\\ &\quad + \sum_{i_1 \in j_0^{-1}(N)} d_1^{(i_1-1)(n-1)+N} + \sum_{i_1 \in j_0^{-1}(N)} -d_1^{(i_1-1)(n-1)+N}\\ &\quad + \sum_{i_1 \in j_0^{-1}(N)} \sum_{i_2 \in \{1,\dots,n-1\} \setminus \{N\}} -d_1^{(i_1-1)(n-1)+i_2}\end{aligned}$$

$$= d_1^{N-1} + d_1^N + \sum_{i_1 \in \{2,\dots,m\} \setminus \{j_0^{-1}(N)\}} d_1^{(i_1-1)(n-1)+N-1} + u + v,$$

where

$$\begin{aligned} u &:= \sum_{i_1 \in \{2,\dots,m\} \setminus \{j_0^{-1}(N)\}} d_1^{(i_1-1)(n-1)+N} + \sum_{i_1 \in j_0^{-1}(N)} d_1^{(i_1-1)(n-1)+N} \\ &= \sum_{i_1=2}^{m} d_1^{(i_1-1)(n-1)+N} + d_1^{(m-1)(n-1)+N} \end{aligned}$$

and

$$\begin{aligned} v &:= \sum_{i_1 \in j_0^{-1}(N)} -d_1^{(i_1-1)(n-1)+N} + \sum_{i_1 \in j_0^{-1}(N)} \sum_{i_2 \in \{1,\dots,n-1\} \setminus \{N\}} -d_1^{(i_1-1)(n-1)+i_2} \\ &= \sum_{i_1 \in j_0^{-1}(N) \setminus \{m\}} \sum_{i_2=1}^{n-1} -d_1^{(i_1-1)(n-1)+i_2} + \sum_{i_2=1}^{n-1} -d_1^{(m-1)(n-1)+i_2}. \end{aligned}$$

Further, we take a close look at the system $A_1^\top d_1 = 0$ to find that

$$\begin{aligned} &\sum_{i_1 \in \{2,\dots,m\} \setminus \{j_0^{-1}(N)\}} d_1^{(i_1-1)(n-1)+N-1} + u + v \\ &\quad = -\Bigg(\sum_{i_1 \in \{2,\dots,m\} \setminus \{j_0^{-1}(N)\}} -d_1^{(i_1-1)(n-1)+N-1} \\ &\qquad - d_1^{(m-1)(n-1)+N} + \sum_{i_1 \in j_0^{-1}(N) \setminus \{m\}} \sum_{i_2=1}^{n-1} d_1^{(i_1-1)(n-1)+i_2} \Bigg) \\ &\qquad - \Bigg(\sum_{i_1=2}^{m} -d_1^{(i_1-1)(n-1)+N} + \sum_{i_2=1}^{n-1} d_1^{(m-1)(n-1)+i_2} \Bigg) \\ &\quad = -d_1^{N-1} - d_1^N. \end{aligned}$$

Consequently, $(A_0^\top d_0)_N = 0$. Moreover, because of $\{i_1 \colon j_0(i_1) < N+1\} = \{1, \dots, m\}$, we immediately receive that (choosing $l = N+1$ in (4.16))

$$(A_0^\top d_0)_{N+1} = \sum_{i_1=1}^{m} d_0^{(i_1-1)(n-1)+N} = 0$$

and, for $l \in \{N+2, \dots, n\}$,

$$(A_0^\top d_0)_l = \sum_{i_1=1}^{m} d_0^{(i_1-1)(n-1)+l-1} = \sum_{i_1=1}^{m} d_1^{(i_1-1)(n-1)+l-1} = 0,$$

which again can be justified by a close look at the system $A_1^\top d_1 = 0$. To summarize all of the above, we have shown that $A_0^\top d_0 = 0$.

Let us now assume that the system $A_0 b \leq c_0$ has a solution. Then, by Lemma 4.20, $d_0^\top c_0 \geq 0$. A comparison of c_0 with c_1, see their representations

from (4.9), yields that

$$c_1^{(i_1-1)(n-1)+i_2} = c_0^{(i_1-1)(n-1)+i_2} + \begin{cases} 0, & \text{if } i_1 \in \{1,\dots,m-1\}, \\ 2\left(\bar{c}_{m,N} - \bar{c}_{m,N+1}\right), & \text{if } i_1 = m,\ i_2 = N, \\ \bar{c}_{m,N} - \bar{c}_{m,N+1}, & \text{else.} \end{cases} \tag{4.18}$$

For all $i_1 \in \{1,\dots,m-1\}$ and $i_2 = N$, this yields that

$$c_1^{(i_1-1)(n-1)+N} = \bar{c}_{i_1,N+1} - \bar{c}_{i_1,j_0(i_1)},$$

whereas

$$c_1^{(m-1)(n-1)+N} = \bar{c}_{m,N} - \bar{c}_{m,N+1}.$$

Moreover, for $i_1 \notin j_0^{-1}(N)$ and $i_2 = N-1$,

$$c_1^{(i_1-1)(n-1)+N-1} = \bar{c}_{i_1,N} - \bar{c}_{i_1,j_0(i_1)}.$$

This, together with the definition of d_0, see (4.14), and (4.18) leads to

$$\begin{aligned} & d_1^\top c_1 - d_0^\top c_0 \\ &= \sum_{i_1 \notin j_0^{-1}(N)} d_1^{(i_1-1)(n-1)+N}\left(c_1^{(i_1-1)(n-1)+N} - c_1^{(i_1-1)(n-1)+N-1}\right) \\ &\quad + \sum_{i_1 \in j_0^{-1}(N)\setminus\{m\}} d_1^{(i_1-1)(n-1)+N} c_1^{(i_1-1)(n-1)+N} \\ &\quad + \sum_{i_2 \in \{1,\dots,n-1\}\setminus\{N\}} d_1^{(m-1)(n-1)+i_2}\left(c_1^{(m-1)(n-1)+i_2} - c_0^{(m-1)(n-1)+i_2}\right) \\ &\quad + d_1^{(m-1)(n-1)+N} c_1^{(m-1)(n-1)+N} \\ &= d_1^N\left(\bar{c}_{1,N+1} - \bar{c}_{1,N}\right) \\ &\quad + \sum_{i_1 \in \{2,\dots,m-1\}} d_1^{(i_1-1)(n-1)+N}\left(\bar{c}_{i_1,N+1} - \bar{c}_{i_1,N}\right) \\ &\quad + \sum_{i_2=1}^{n-1} d_1^{(m-1)(n-1)+i_2}\left(\bar{c}_{m,N} - \bar{c}_{m,N+1}\right). \end{aligned} \tag{4.19}$$

The equation $\left(A_1^\top d_1\right)_{N+1} = 0$ reveals that

$$d_1^N = \sum_{i_1=2}^{m-1} -d_1^{(i_1-1)(n-1)+N} + \sum_{i_2=1}^{n-1} d_1^{(m-1)(n-1)+i_2} \geq 0. \tag{4.20}$$

Adding $d_1^N\left(\bar{c}_{m,N} - \bar{c}_{m,N+1}\right) - d_1^N\left(\bar{c}_{m,N} - \bar{c}_{m,N+1}\right)$ to the right-hand side of the equation in (4.19) and using the relation in (4.20), one finds that

$$\begin{aligned} d_1^\top c_1 - d_0^\top c_0 = {} & d_1^N\left(\bar{c}_{1,N+1} - \bar{c}_{1,N} + \bar{c}_{m,N} - \bar{c}_{m,N+1}\right) \\ & + \sum_{i_1=2}^{m-1} d_1^{(i_1-1)(n-1)+N}\left(\bar{c}_{i_1,N+1} - \bar{c}_{i_1,N} + \bar{c}_{m,N} - \bar{c}_{m,N+1}\right) \end{aligned}$$

$$\geq \sum_{i_1=2}^{m-1} d_1^{(i_1-1)(n-1)+N}\big(\bar{c}_{i_1,N+1} - \bar{c}_{i_1,N} - (\bar{c}_{m,N+1} - \bar{c}_{m,N})\big) \geq 0,$$

where the last estimate stems from the nonnegativity of d_1 and an application of Lemma 4.21 (for $i_2 = N+1$). Consequently, $d_1^\top c_1 \geq d_0^\top c_0 \geq 0$ which, owing to Lemma 4.20, completes the proof. □

The just proven lemma states that, for a given matrix, we can always "advance" the nonzero entry of its last row by one column without sacrificing the solvability of the associated linear inequality system. We will see in the next lemma that we can, in the same sense, "move up" nonzero entries of the rows above.

Lemma 4.24. *Let $m, n \in \mathbb{N}$ with $n \geq m \geq 3$ be given and consider the cost matrix $\bar{c}$ from Example 4.22. For $p \in \{0,1\}$, consider the monotone assignment functions $j_p\colon \{1,\dots,m\} \to \{1,\dots,n\}$ and denote their associated reduced system matrices and reduced cost vectors by A_p and c_p, respectively.*

Assume that $j_0(1) = 1$ as well as $N \coloneqq j_0(m) \geq 2$. Abbreviate $I \coloneqq \max\{i_1\colon i_1 \notin j_0^{-1}(N)\}$ and assume that $I > 1$ as well as $j_0(I) = N-1$ and, moreover, that $j_1|_{\{1,\dots,m\}\setminus\{I\}} \equiv j_0|_{\{1,\dots,m\}\setminus\{I\}}$ and $j_1(I) = N = j_0(I)+1$.

Then, if the system $A_0 b \leq c_0$ has a solution, so does the system $A_1 b \leq c_1$.

Proof. Assume that the system $A_0 b \leq c_0$,

$$b_{i_2} - b_{j_0(i_1)} \leq \bar{c}_{i_1,i_2} - \bar{c}_{i_1,j_0(i_1)}, \quad i_1 \in \{1,\dots,m\},\ i_2 \in \{1,\dots,n\}\setminus\{j_0(i_1)\},$$

has a solution. Then the subsystem

$$b_{i_2} - b_{j_0(i_1)} \leq \bar{c}_{i_1,i_2} - \bar{c}_{i_1,j_0(i_1)}, \quad i_1 \in \{1,\dots,I\},\ i_2 \in \{1,\dots,n\}\setminus\{j_0(i_1)\},$$

has the same solution. Applying Lemma 4.23 to the restriction $j_0|_{\{1,\dots,I\}}$, we obtain that the system

$$b_{i_2} - b_{j_1(i_1)} \leq \bar{c}_{i_1,i_2} - \bar{c}_{i_1,j_1(i_1)} \quad i_1 \in \{1,\dots,I\},\ i_2 \in \{1,\dots,n\}\setminus\{j_1(i_1)\}, \tag{4.21}$$

with $j_1|_{\{1,\dots,I-1\}} \equiv j_0|_{\{1,\dots,I-1\}}$ and $j_1(I) = j_0(I)+1 = N$, admits at least one solution $b' \in \mathbb{R}^n$. We then define the vector $b_1 \in \mathbb{R}^n$ by

$$b_1^{i_2} \coloneqq \begin{cases} b'_{i_2}, & \text{if } i_2 \leq N, \\ b'_{i_2} - \big(\bar{c}_{I,i_2} - \bar{c}_{I,N} - (\bar{c}_{m,i_2} - \bar{c}_{m,N})\big), & \text{if } i_2 > N, \end{cases} \tag{4.22}$$

for all $i_2 \in \{1,\dots,n\}$.

On the one hand, let $i_1 \in \{1,\dots,I-1\}$ and $i_2 \in \{1,\dots,n\}\setminus\{j_1(i_1)\}$ be arbitrary. By construction of j_1, we find that $j_1(i_1) < N$. If $i_2 \leq N$, because of (4.21) and (4.22),

$$b_1^{i_2} - b_1^{j_1(i_1)} = b'_{i_2} - b'_{j_1(i_1)} \leq \bar{c}_{i_1,i_2} - \bar{c}_{i_1,j_1(i_1)}.$$

If $i_2 > N$, we additionally apply (4.11) to receive

$$\begin{aligned} b_1^{i_2} - b_1^{j_1(i_1)} &= b'_{i_2} - \big(\bar{c}_{I,i_2} - \bar{c}_{I,N} - (\bar{c}_{m,i_2} - \bar{c}_{m,N})\big) - b'_{j_1(i_1)} \\ &\leq b'_{i_2} - b'_{j_1(i_1)} \leq \bar{c}_{i_1,i_2} - \bar{c}_{i_1,j_1(i_1)}. \end{aligned}$$

On the other hand, let $i_1 \in \{I, \dots, m\}$ be arbitrary. Then, $j_1(i_1) = j_1(I) = N$. For $i_2 \in \{1, \dots, N-1\}$,

$$\begin{aligned} b_1^{i_2} - b_1^{j_1(i_1)} = b'_{i_2} - b'_{j_1(I)} &\leq \bar{c}_{I,i_2} - \bar{c}_{I,N} \\ &\leq \bar{c}_{i_1,i_2} - \bar{c}_{i_1,N} = \bar{c}_{i_1,i_2} - \bar{c}_{i_1,j_1(i_1)}, \end{aligned}$$

where the second estimate holds because of (4.10). If $i_2 > N$, we use (4.11) to find that

$$\begin{aligned} b_1^{i_2} - b_1^{j_1(i_1)} &= b'_{i_2} - b'_{j_1(I)} - \big(\bar{c}_{I,i_2} - \bar{c}_{I,N} - (\bar{c}_{m,i_2} - \bar{c}_{m,N})\big) \\ &\leq \bar{c}_{I,i_2} - \bar{c}_{I,N} - \big(\bar{c}_{I,i_2} - \bar{c}_{I,N} - (\bar{c}_{m,i_2} - \bar{c}_{m,N})\big) \\ &= \bar{c}_{m,i_2} - \bar{c}_{m,N} \leq \bar{c}_{i_1,i_2} - \bar{c}_{i_1,N} = \bar{c}_{i_1,i_2} - \bar{c}_{i_1,j_1(i_1)}. \end{aligned}$$

Thus, we have shown that for all $i_1 \in \{1, \dots, m\}$ and all $i_2 \in \{1, \dots, n\} \backslash \{j_1(i_1)\}$,

$$b_1^{i_2} - b_1^{j_1(i_1)} \leq \bar{c}_{i_1,i_2} - \bar{c}_{i_1,j_1(i_1)},$$

or equivalently, $A_1 b_1 \leq c_1$ as claimed. □

The following result elaborates on the inductive argument we mentioned earlier and which allows us to prove the existence of solutions to systems $Ab \leq c$ that correspond to (almost) arbitrary monotone assignment functions j.

Theorem 4.25. *Let $m, n \in \mathbb{N}$ with $n \geq m \geq 2$ be given and consider the cost matrix $\bar{c}$ from Example 4.22. Consider the monotone assignment function $j\colon \{1, \dots, m\} \to \{1, \dots, n\}$ with $j(1) = 1$ and denote its associated reduced system matrix and reduced cost vector by A and c, respectively.*

Then, there exists a solution $b \in \mathbb{R}^n$ to the system $Ab \leq c$.

Proof. In Example 4.22, we have shown that the system $A_0 b \leq c_0$ belonging to the monotone assignment function $j_0\colon \{1, \dots, m\} \to \{1, \dots, n\}$ defined by $j_0(1) = \dots = j_0(m) = 1$ admits a solution.

If $m = 2$, applying Lemma 4.23 a total of $j(2) - 1$ times, starting from the system $A_0 b \leq c_0$, yields the claim.

If $m = 3$, we alternately apply Lemma 4.23 and Lemma 4.24 a total of $j(3) - 1$ and $j(2) - 1$ times, respectively, starting with the former at the system $A_0 b \leq c_0$. This yields the claim.

The procedure for the case $m = 3$ describes a method by which we can prove the assertion for all other cases where $m > 3$: beginning with the system $A_0 b \leq c_0$, we apply Lemma 4.23 and then, if necessary, Lemma 4.24 up to $m-2$ times to the resulting system. We repeat this process a total of $j(m) - 2$ times to prove the claim. □

Applying Theorem 4.25 to the solution from Assumption 4.15 finally yields the desired approximation result.

Corollary 4.26. *Consider the solution (π^*, μ_1^*) from Assumption 4.15 and the vanishing sequence of regularization parameters $(\gamma_k)_{k \in \mathbb{N}} \searrow 0$ and the sequence of solutions $(\bar{\pi}_k, \bar{\mu}_1^k, \bar{c}_k)_{k \in \mathbb{N}}$ to the regularized bilevel problems from Subchapter 4.4. Then, every cluster point $(\bar{\pi}, \bar{\mu}_1)$ of that sequence is a solution to (BH).*

Proof. The existence of a cluster point $(\bar{\pi}, \bar{\mu}_1)$ was discussed at the beginning of Subchapter 4.4. By Theorem 4.25, there exists a solution $b^* \in \mathbb{R}^{n_2}$ to the system $A^* b \leq c^*$, where A^* and c^* denote the reduced system matrix and the reduced cost vector associated with the monotone assignment function j^*, respectively. By definition of A^* and c^*, this shows that b^* satisfies the system from (4.8), which in turn yields the existence of a recovery sequence $(\pi_k^*, \mu_{1,k}^*, c_k^*)$ accompanying the solution (π^*, μ_1^*), see Lemma 4.17. The claim of the corollary then follows from Theorem 4.14. □

We conclude this chapter with a comparison of the approximation results of the infinite-dimensional case of Part I and the finite-dimensional case of Part II.

- In both cases, if we presuppose the existence of a solution of the non-regularized bilevel problem that is accompanied by a recovery sequence, we can show that solutions of the regularized bilevel problems converge (w.r.t. the proper topology) towards solutions of the non-regularized bilevel problems. In other words, we can approximate certain solutions of the non-regularized problems arbitrarily accurately. However, since we do not need to smooth the marginals in the finite-dimensional case of Part II, a single regularization parameter is sufficient in this case.
- If we compare the scenarios for which we can prove the existence of a recovery sequence, we find that the scenario described in Corollary 3.41 is slightly less general than the one from Assumption 4.15:
  - In Corollary 3.41, we assume that the domains of the marginals coincide, while in Assumption 4.15 both of the domains can be sets with an arbitrary (finite) number of points.
  - The parameter ρ defining the cost function in Corollary 3.41 is restricted to $\rho > 1$, while in Assumption 4.15 we also include the case $\rho = 1$.
  - In contrast to the infinite-dimensional case, there are no additional assumptions on the objective function in the finite-dimensional case, other than its lower semi-continuity and boundedness on bounded sets, which is required anyway for the approximation results of both cases.
  - The proof of the existence of a recovery sequence in Corollary 3.41 relies heavily on the uniqueness of the optimal transport plan, whereas the recovery sequence in Assumption 4.15 can be constructed without this property. This, however, goes along with the fact that the latter proof is considerably more complex than in the infinite-dimensional case.

Chapter 5

Towards Implicit Programming

Implicit programming (IP) is an approach to the numerical treatment of optimization problems whose constraints include some sort of defining relation between some input variable (often called "control") and some output variable (often called "state"), such as solution operators of variational inequalities, complementarity systems, or, like in the case of the bilevel Hitchcock problem, (non-)linear optimization problems. Evolving from the implicit programming problem, an optimization problem whose constraints are implicitly defined by the problem's solution itself, see e.g. [36] or [35], the IP approach is frequently applied in the context of (but not limited to) mathematical programming with equilibrium constraints (MPECs), see e.g. [54], [60], [46], or [59].

The IP approach typically consists of replacing the control-to-state relation from the problems constraints by the corresponding solution mapping (often called "control-to-state mapping"), which then can be plugged directly into the upper-level target functional, leading to a (possibly unconstrained) optimization problem with fewer optimization variables. Given sufficient smoothness of the target functional, one can then employ (typically nonsmooth) optimization methods to solve the original optimization problem. This, however, generally requires that

1. the solution operator is single-valued (i.e., in particular not set-valued), since the higher-order objective functional typically operates only on single elements and not on sets;

2. the solution mapping itself must satisfy some notion of differentiability.

We already tried to indicate in Subchapter 4.1 that both of the above requirements are not satisfied in the context of the bilevel Hitchcock problem: In general, the solution of the Hitchcock problem is not unique. Therefore, we can neither find a single-valued solution operator nor differentiate it in a sense that would be useful for our purposes.

How we can still put the IP approach into practice in the context of the bilevel Hitchcock problem will be the topic of the present chapter.

5.1 Regularization of the Dual Problem of the Regularized Hitchcock Problem

We have mentioned at several points in this thesis that we would like to use the implicit programming approach to solve the bilevel Hitchcock problem (BH). As noted in the introduction of this chapter, the Hitchcock problem itself is not necessarily uniquely solvable, ruling out the existence of a single-valued solution operator.

For this and other reasons, we have introduced the quadratic regularization of the Hitchcock problem in Chapter 4, which guarantees the uniqueness of the optimal solution and therefore implicitly defines a solution operator. However, as can be seen in Theorem 4.9, the dual variables α_1 and α_2 corresponding to a solution π of (H_γ) are not unique for two reasons: First, because of the outer sum of the dual variables, namely $\alpha_1 \oplus \alpha_2$, one can constantly shift the dual variables in the opposite direction, i.e., consider $(\alpha_1 + a, \alpha_2 - a)$ for $a \in \mathbb{R}$, and this will not affect the system in (4.2), and second, the $(\,\cdot\,)_+$-operator provides the potential for arbitrary deviations of the expression $\alpha_1 \oplus \alpha_2 - c$ where it is negative.

We therefore consider an additional regularization of the dual problem (HD_γ) corresponding to the solution of (H_γ). Given $\gamma, \varepsilon > 0$, some cost matrix $c \in \mathbb{R}^{n_1 \times n_2}$, as well as the (arbitrary) marginals μ_1 and μ_2, we consider the *regularized dual problem* of the regularized Hitchcock problem:

$$\begin{aligned} \sup_{\alpha_1, \alpha_2} \quad & (\alpha_1, \mu_1)_{\mathbb{R}^{n_1}} + (\alpha_2, \mu_2)_{\mathbb{R}^{n_2}} \\ & -\tfrac{1}{2\gamma}\|(\alpha_1 \oplus \alpha_2 - c)_+\|_F^2 - \tfrac{\varepsilon}{2}\big(\|\alpha_1\|_{\mathbb{R}^{n_1}}^2 + \|\alpha_2\|_{\mathbb{R}^{n_2}}^2\big) \\ \text{s.t.} \quad & \alpha_1 \in \mathbb{R}^{n_1},\ \alpha_2 \in \mathbb{R}^{n_2}. \end{aligned} \tag{$\mathrm{HD}_\gamma^\varepsilon$}$$

Without having to do extensive preliminary work, we immediately arrive at the following result.

Lemma 5.1. *For any $\gamma, \varepsilon > 0$, the regularized dual problem ($\mathrm{HD}_\gamma^\varepsilon$) admits a unique solution. Moreover, its first-order necessary and sufficient condition is given by*

$$\begin{aligned} (\alpha_1 \oplus \alpha_2 - c)_+ \mathbb{1} + \gamma\varepsilon\alpha_1 &= \gamma\mu_1, \\ (\alpha_1 \oplus \alpha_2 - c)_+^\top \mathbb{1} + \gamma\varepsilon\alpha_2 &= \gamma\mu_2. \end{aligned} \tag{5.1}$$

Proof. It is straightforward to show that the negative of ($\mathrm{HD}_\gamma^\varepsilon$)'s target function is strongly convex. According to [58, Proposition 3.10.8], it therefore has a unique global minimizer, which is at the same time the unique global maximizer of ($\mathrm{HD}_\gamma^\varepsilon$).

Because the mapping $f\colon \mathbb{R} \ni x \mapsto \max\{0, x\}^2 \in \mathbb{R}$ is differentiable and has the derivative $f'(x) = 2\max\{0, x\}$, the term

$$-\frac{1}{2\gamma}\|(\alpha_1 \oplus \alpha_2 - c)_+\|_F^2 = -\frac{1}{2\gamma}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\max\big\{0, \alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2}\big\}^2$$

is differentiable and has the partial derivatives

$$-\frac{1}{\gamma}\sum_{i_2=1}^{n_2}\max\big\{0, \alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2}\big\} \qquad (\text{w.r.t. } \alpha_1^{i_1})$$

and

$$-\frac{1}{\gamma}\sum_{i_1=1}^{n_1}\max\{0,\alpha_1^{i_1}+\alpha_2^{i_2}-c_{i_1,i_2}\}. \qquad \text{(w.r.t. } \alpha_2^{i_2}\text{)}$$

This in turn shows that ($\mathrm{HD}_\gamma^\varepsilon$)'s objective function is differentiable w.r.t. α_1 and α_2. Thus, its unique maximizer can be equally characterized by the first-order conditions

$$\mu_1^{i_1}-\frac{1}{\gamma}\sum_{i_2=1}^{n_2}\left(\alpha_1^{i_1}+\alpha_2^{i_2}-c_{i_1,i_2}\right)_+-\varepsilon\alpha_1^{i_1}=0 \quad \text{for all } i_1=1,\dots,n_1$$

and

$$\mu_2^{i_2}-\frac{1}{\gamma}\sum_{i_1=1}^{n_1}\left(\alpha_1^{i_1}+\alpha_2^{i_2}-c_{i_1,i_2}\right)_+-\varepsilon\alpha_2^{i_2}=0 \quad \text{for all } i_2=1,\dots,n_2,$$

which are equivalent to (5.1). □

Remark 5.2. Note that the marginals μ_1 and μ_2 need not be compatible for the regularized dual problem to admit a unique solution. In fact, μ_1 and μ_2 do not even need to be marginals, since the regularized dual problem allows for vectors of arbitrary sign. Lemma 5.1 therefore implies the existence of the solution operator of ($\mathrm{HD}_\gamma^\varepsilon$),

$$\mathcal{F}_{\gamma,\varepsilon}\colon \mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\to\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}, \quad (\mu_1,\mu_2)\mapsto(\alpha_1,\alpha_2)$$

with (α_1,α_2) being the unique solution to the system in (5.1) w.r.t. μ_1 and μ_2. In particular, the solutions of (5.1) are in general no longer the dual variables to the marginals μ_1 and μ_2, which is why we will define a regularized marginal-to-transport-plan mapping in Subchapter 5.2.

We immediately observe that the operator $\mathcal{F}_{\gamma,\varepsilon}$ is bijective: injectivity follows from the unique solvability of the system in (5.1); surjectivity holds since one can simply evaluate the left-hand side of the equations in (5.1) to compute the corresponding marginals. As a consequence, there exists the inverse operator

$$\mathcal{F}_{\gamma,\varepsilon}^{-1}(\alpha_1,\alpha_2)=\left(\frac{1}{\gamma}(\alpha_1\oplus\alpha_2-c)_+\mathbb{1}+\varepsilon\alpha_1,\ \frac{1}{\gamma}(\alpha_1\oplus\alpha_2-c)_+^\top\mathbb{1}+\varepsilon\alpha_2\right) \tag{5.2}$$

and this inverse operator is continuous. ○

Remark 5.3. In the context of ($\mathrm{HD}_\gamma^\varepsilon$), we further would like to mention the following:

– As was the case with the solution operator $\mathcal{S}_\gamma$ from Subchapter 4.2, one could also include the cost matrix in the formulation of the solution operator from Remark 5.2, i.e., one could consider the operator

$$\tilde{\mathcal{F}}_{\gamma,\varepsilon}\colon \mathbb{R}^{n_1\times n_2}\times\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}\to\mathbb{R}^{n_1}\times\mathbb{R}^{n_2}, \quad (c,\mu_1,\mu_2)\mapsto(\alpha_1,\alpha_2)$$

with being (α_1,α_2) the unique solution of ($\mathrm{HD}_\gamma^\varepsilon$) w.r.t. μ_1, μ_2, and c in order to account for the cost matrix being an optimization variable as well.

However, the focus of this chapter lies on realizing the IP approach for the (regularized) bilevel Hitchcock problem. The cost function served mainly as a tool for the approximation result of Subchapter 4.4 and had (from our viewpoint) no other significant effect on the analysis. Thus, for simplicity, we refrain from including it in the analysis of the present chapter.

– While the regularization approach in ($\mathrm{HD}_\gamma^\varepsilon$) corresponds to a standard Tikhonov regularization of the dual problem (HD_γ), another approach to promote uniqueness of the dual variables would be to consider approximations of the mapping $x \mapsto \max\{0, x\}$ by means of the smooth, strictly increasing, and strictly convex functions

$$f_\varepsilon \colon \mathbb{R} \to \mathbb{R}_{>0}, \quad f_\varepsilon(x) \coloneqq \frac{x}{2} + \frac{1}{2}\sqrt{x^2 + \varepsilon^2}$$

or

$$\mathrm{LSE}_\varepsilon \colon \mathbb{R} \to \mathbb{R}_{>0}, \quad \mathrm{LSE}_\varepsilon(x) \coloneqq \varepsilon \log\bigl(1 + \exp(x/\varepsilon)\bigr),$$

where log refers to the natural logarithm and LSE is short for "log-sum-exp", see e.g. [12, Example 3.1.5], or any other approximation of $x \mapsto \max\{0, x\}$ having the same properties.

Substituting the $(\cdot)_+$ operator in the objective function of (HD_γ) by the element-wise application of one of the above functions would then lead to regularized problems with similar properties as ($\mathrm{HD}_\gamma^\varepsilon$) but smooth first-order optimality systems. The resulting optimality systems would, just like (5.1), contain nonlinear terms but would not, in contrast to (5.1), benefit from the sparsity induced by the $(\cdot)_+$ operator.

We will see below that the induced sparsity is indeed very useful to calculate derivatives of the regularized dual problem.

○

In the remainder of this subchapter, we will analyze the differentiability properties of the solution operator of the regularized dual problem. We start by showing that the operator is Lipschitz continuous and therefore, according to Rademacher's theorem, differentiable almost everywhere on its domain.

Lemma 5.4. *The solution operator $\mathcal{F}_{\gamma,\varepsilon}$ is globally Lipschitz continuous.*

Proof. The inverse operator $\mathcal{F}_{\gamma,\varepsilon}^{-1}$ from (5.2) is continuous and piecewise linear with $n_1 \cdot n_2$ segments, where the slopes are bounded (from below) by a single constant. This implies that its inverse, $\mathcal{F}_{\gamma,\varepsilon}$, is also continuous and piecewise linear and that the slope of each of its segments is bounded (from above) by the inverse of this constant. Thus, $\mathcal{F}_{\gamma,\varepsilon}$ must be Lipschitz continuous. □

Consequently, according to Rademacher's theorem, $\mathcal{F}_{\gamma,\varepsilon}$ is differentiable almost everywhere in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. To be able to make qualitative statements about the points where $\mathcal{F}_{\gamma,\varepsilon}$ is differentiable and moreover to be able to characterize its derivative in these points, we start with the characterization of its directional derivative.

Proposition 5.5. *Let $(\mu_1, \mu_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be an arbitrary point and abbreviate $(\alpha_1, \alpha_2) \coloneqq \mathcal{F}_{\gamma,\varepsilon}(\mu_1, \mu_2)$. Then, the solution operator $\mathcal{F}_{\gamma,\varepsilon}$ is directionally differentiable in each direction $(h_1, h_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and the directional derivative $\mathcal{F}'_{\gamma,\varepsilon}\big((\mu_1, \mu_2); (h_1, h_2)\big)$ is given by the unique point $(\eta_1, \eta_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ that solves the system*

$$\begin{aligned} \max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2)\ \mathbb{1} + \gamma\varepsilon\eta_1 &= \gamma h_1, \\ \max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2)^\top \mathbb{1} + \gamma\varepsilon\eta_2 &= \gamma h_2, \end{aligned} \tag{5.3}$$

where

$$\max'(x; y) = \begin{cases} 0, & \text{if } x < 0, \\ \max\{0, y\}, & \text{if } x = 0, \\ y, & \text{if } x > 0, \end{cases} \tag{5.4}$$

is the directional derivative of the mapping $\mathbb{R} \ni x \mapsto \max\{0, x\} \in \mathbb{R}_+$ at some point $x \in \mathbb{R}$ in the direction $y \in \mathbb{R}$. In (5.3), this directional derivative is understood to be applied entry-wise to the matrices $\alpha_1 \oplus \alpha_2 - c$ and $\eta_1 \oplus \eta_2$, i.e., $\max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2) \in \mathbb{R}^{n_1 \times n_2}$.

Proof. Given $t > 0$, we consider the point $(\alpha_{1,t}, \alpha_{2,t}) \coloneqq \mathcal{F}_{\gamma,\varepsilon}\big((\mu_1, \mu_2) + t(h_1, h_2)\big)$. Then, the difference quotient

$$(\eta_{1,t}, \eta_{2,t}) \coloneqq \Big(\frac{\alpha_{1,t} - \alpha_1}{t}, \frac{\alpha_{2,t} - \alpha_2}{t}\Big) = \frac{\mathcal{F}_{\gamma,\varepsilon}\big((\mu_1, \mu_2) + t(h_1, h_2)\big) - \mathcal{F}_{\gamma,\varepsilon}(\mu_1, \mu_2)}{t}$$

is bounded, since

$$\begin{aligned} &\|(\eta_{1,t}, \eta_{2,t})\|_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} \\ &\quad \leq \frac{L_{\mathcal{F}_{\gamma,\varepsilon}} \|(\mu_1, \mu_2) + t(h_1, h_2) - (\mu_1, \mu_2)\|_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}}}{t} = L_{\mathcal{F}_{\gamma,\varepsilon}} \|(h_1, h_2)\|_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} \end{aligned}$$

by Lemma 5.4. Consequently, if we consider an arbitrary vanishing sequence $t_k \searrow 0$, the sequence $(\eta_{1,t_k}, \eta_{2,t_k})_{k \in \mathbb{N}}$ is bounded and therefore contains some convergent subsequence which we denote by $(\eta_{1,l}, \eta_{2,l})_{l \in \mathbb{N}}$ and which converges to some point $(\bar\eta_1, \bar\eta_2)$.

Subtracting the system (5.1) w.r.t. (α_1, α_2) from the system (5.1) w.r.t. $(\alpha_{1,l}, \alpha_{2,l}) \coloneqq (\alpha_{1,t_l}, \alpha_{2,t_l})$, we observe that the difference quotient satisfies the equations

$$\begin{aligned} \frac{\big((\alpha_{1,l} \oplus \alpha_{2,l} - c)_+ - (\alpha_1 \oplus \alpha_2 - c)_+\big)\ \mathbb{1}}{t_l} + \gamma\varepsilon\eta_{1,l} &= \gamma h_1, \\ \frac{\big((\alpha_{1,l} \oplus \alpha_{2,l} - c)_+ - (\alpha_1 \oplus \alpha_2 - c)_+\big)^\top \mathbb{1}}{t_l} + \gamma\varepsilon\eta_{2,l} &= \gamma h_2, \end{aligned} \tag{5.5}$$

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

Owing to the continuity of the solution operator, see Lemma 5.4, it holds that $(\alpha_{1,l}, \alpha_{2,l}) \to (\alpha_1, \alpha_2)$ and in particular $(\alpha_{1,l}, \alpha_{2,l}) = (\alpha_1, \alpha_2) + t_l(\bar\eta_1, \bar\eta_2) + o(t_l)$. Thus, the Hadamard differentiability of the mapping $x \mapsto \max\{0, x\}$, see Lemma D.7, implies that

$$\frac{(\alpha_{1,l}^{i_1} + \alpha_{2,l}^{i_2} - c_{i_1,i_2})_+ - (\alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2})_+}{t_l} \ \to\ \max'(\alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2}; \bar\eta_1^{i_1} + \bar\eta_2^{i_2})$$

for each $(i_1, i_2) \in \Omega$ as $l \to \infty$. We can therefore pass to the limit in (5.5) to arrive at

$$\begin{aligned} &\max'(\alpha_1 \oplus \alpha_2 - c; \bar{\eta}_1 \oplus \bar{\eta}_2)\ \mathbb{1} + \gamma\varepsilon\bar{\eta}_1 = \gamma h_1, \\ &\max'(\alpha_1 \oplus \alpha_2 - c; \bar{\eta}_1 \oplus \bar{\eta}_2)^\top \mathbb{1} + \gamma\varepsilon\bar{\eta}_2 = \gamma h_2. \end{aligned} \tag{5.6}$$

In the following, we are going to convince ourselves that $(\bar{\eta}_1, \bar{\eta}_2)$ is the only possible solution of the above system. If this is the case, then Lemma D.5 ensures the convergence of the whole sequence $(\eta_{1,t_k}, \eta_{2,t_k}) \to (\bar{\eta}_1, \bar{\eta}_2)$. Moreover, because the sequence $(t_k)_{k\in\mathbb{N}}$ was arbitrary, we then find that

$$\lim_{t\searrow 0} \frac{\mathcal{F}_{\gamma,\varepsilon}\big((\mu_1,\mu_2) + t(h_1,h_2)\big) - \mathcal{F}_{\gamma,\varepsilon}(\mu_1,\mu_2)}{t} = \lim_{t\searrow 0}(\eta_{1,t}, \eta_{2,t}) = (\bar{\eta}_1, \bar{\eta}_2),$$

which proves the claim.

In order to show that the system in (5.6) admits a unique solution, we abbreviate $X \coloneqq \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and define the operator $F\colon X \to X$ by

$$\begin{aligned} (u_1, u_2) \mapsto \Big(&\max'(\alpha_1 \oplus \alpha_2 - c; u_1 \oplus u_2)\ \mathbb{1} + \gamma\varepsilon u_1, \\ &\max'(\alpha_1 \oplus \alpha_2 - c; u_1 \oplus u_2)^\top \mathbb{1} + \gamma\varepsilon u_2\Big). \end{aligned}$$

We are going to prove that F is

(i) strongly monotone, i.e., there exists some $c > 0$ such that

$$(Fu - Fv, u - v)_X \geq c\|u - v\|_X^2 \quad \text{for all } u, v \in X;$$

(ii) coercive, i.e.,

$$\lim_{\|u\|_X \to \infty} \frac{(Fu, u)_X}{\|u\|_X} = \infty;$$

(iii) hemicontinuous, i.e., the function

$$t \mapsto (F(u + tv), w)_X$$

is continuous on $[0, 1]$ for all $u, v, w \in X$.

If (i) – (iii) are established, then the Minty-Browder theorem, see e.g. [67, Satz 1.5], ensures that the equation $Fu = b$ admits a unique solution for any right-hand side $b \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. Choosing $b = \gamma(h_1, h_2)$ then yields the claim.

Ad (i): A paper-and-pencil aided calculation shows that

$$\begin{aligned} &(Fu - Fv, u - v)_X \\ &= \big(\max'(\alpha_1 \oplus \alpha_2 - c; u_1 \oplus u_2) - \max'(\alpha_1 \oplus \alpha_2 - c; v_1 \oplus v_2),\ u_1 \oplus u_2 - v_1 \oplus v_2\big)_F \\ &\quad + \gamma\varepsilon\big(\|u_1 - v_1\|_{\mathbb{R}^{n_1}}^2 + \|u_2 - v_2\|_{\mathbb{R}^{n_2}}^2\big) \\ &\geq \gamma\varepsilon\|u - v\|_X^2, \end{aligned}$$

where the estimate follows from a distinction of cases w.r.t. the sign of $\alpha_1 \oplus \alpha_2 - c$ and $(a_+ - b_+)(a - b) \geq 0$ for all $a, b \in \mathbb{R}$.

Ad (ii): Similar to (i), we find that

$$(Fu, u)_X = \big(\max'(\alpha_1 \oplus \alpha_2 - c; u_1 \oplus u_2),\ u_1 \oplus u_2\big)_F + \gamma\varepsilon\|u\|_X^2$$

$$\geq \gamma\varepsilon\|u\|_X^2$$

and therefore

$$\lim_{\|u\|_X\to\infty} \frac{(Fu,u)_X}{\|u\|_X} \geq \lim_{\|u\|_X\to\infty} \gamma\varepsilon\|u\|_X = \infty.$$

Ad (iii): The mapping : $x \mapsto \max\{0,x\}$ is Lipschitz continuous with Lipschitz constant equal to 1 and is directionally differentiable at every point and in every direction. Thus,

$$\begin{aligned}&|\max'(x;z) - \max'(x;y)|\\ &\quad= \lim_{t\searrow 0} \frac{1}{t} |\max\{0, x+tz\} - \max\{0, x+ty\}| \leq |z-y|\end{aligned}$$

for all $y, z \in \mathbb{R}$, i.e., the mapping $y \mapsto \max'(x;y)$ is Lipschitz continuous with the same Lipschitz constant. Consequently, the mapping $t \mapsto F(u+tv)$ is continuous on all of $\mathbb{R}$. □

Remark 5.6. With the same arguments as in Remark 5.2, we see that the mapping

$$(h_1,h_2) \mapsto \mathcal{F}'_{\gamma,\varepsilon}\big((\mu_1,\mu_2);(h_1,h_2)\big)$$

is bijective: it is injective, because the solution to (5.3) is unique; it is surjective, because any given directional derivative (η_1,η_2) can be realized by some direction (h_1,h_2) that can be computed by simply evaluating the left-hand side of (5.3). ○

In order to characterize the points where $\mathcal{F}_{\gamma,\varepsilon}$ is not only directionally but (totally) differentiable, we need the following definition.

Definition 5.7. Given some point $\mu = (\mu_1,\mu_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with $(\alpha_1,\alpha_2) = \mathcal{F}_{\gamma,\varepsilon}(\mu_1,\mu_2)$, we define the sets

$$\begin{aligned}\Omega_+(\mu) &:= \big\{(i_1,i_2) \in \Omega\colon \alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2} > 0\big\},\\ \Omega_0(\mu) &:= \big\{(i_1,i_2) \in \Omega\colon \alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2} = 0\big\},\\ \Omega_-(\mu) &:= \big\{(i_1,i_2) \in \Omega\colon \alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2} < 0\big\}.\end{aligned}$$

We call the sets $\Omega_+(\mu)$, $\Omega_0(\mu)$, and $\Omega_-(\mu)$ *active set*, *biactive set*, and *inactive set*, respectively. Note that $\Omega = \Omega_+(\mu) \,\dot\cup\, \Omega_0(\mu) \,\dot\cup\, \Omega_-(\mu)$.

The following result characterizes the points at which $\mathcal{F}_{\gamma,\varepsilon}$ is (totally) differentiable.

Proposition 5.8. *The solution operator of the regularized dual problem, $\mathcal{F}_{\gamma,\varepsilon}$, is (totally) differentiable at the point $\mu = (\mu_1,\mu_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ if and only if $\Omega_0(\mu) = \emptyset$.*

Proof. Following [22, p. 30], $\mathcal{F}_{\gamma,\varepsilon}$'s Lipschitz continuity is sufficient for the equivalence of $\mathcal{F}_{\gamma,\varepsilon}$ being differentiable in μ and the mapping $\mathcal{F}'_{\gamma,\varepsilon}(\mu;\,\cdot\,)$ being linear. It therefore suffices to show that $\mathcal{F}'_{\gamma,\varepsilon}(\mu;\,\cdot\,)$ is linear if and only if $\Omega_0 := \Omega_0(\mu) = \emptyset$.

To this end, assume that $\Omega_0 = \emptyset$. We consider arbitrary directions $g, h \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and denote their corresponding directional derivatives by $(\theta_1,\theta_2) :=$

$\mathcal{F}'_{\gamma,\varepsilon}(\mu; g)$ and $(\eta_1, \eta_2) \coloneqq \mathcal{F}'_{\gamma,\varepsilon}(\mu; h)$. According to Proposition 5.5, (θ_1, θ_2) and (η_1, η_2) satisfy the systems

$$\begin{aligned}\max'(\alpha_1 \oplus \alpha_2 - c; \theta_1 \oplus \theta_2)\ \mathbb{1} + \gamma\varepsilon\theta_1 &= \gamma g_1,\\ \max'(\alpha_1 \oplus \alpha_2 - c; \theta_1 \oplus \theta_2)^\top \mathbb{1} + \gamma\varepsilon\theta_2 &= \gamma g_2,\end{aligned}$$

and

$$\begin{aligned}\max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2)\ \mathbb{1} + \gamma\varepsilon\eta_1 &= \gamma h_1,\\ \max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2)^\top \mathbb{1} + \gamma\varepsilon\eta_2 &= \gamma h_2,\end{aligned}$$

respectively. Adding those two systems of equations while multiplying the first system with some $\lambda \in \mathbb{R}$, we arrive at

$$\begin{aligned}\big(\lambda\max'(\alpha_1 \oplus \alpha_2 - c; \theta_1 \oplus \theta_2) + \max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2)\big)\ \mathbb{1}&\\ +\gamma\varepsilon(\lambda\theta_1 + \eta_1) &= \gamma(\lambda g_1 + h_1),\\ \big(\lambda\max'(\alpha_1 \oplus \alpha_2 - c; \theta_1 \oplus \theta_2) + \max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2)\big)^\top \mathbb{1}&\\ +\gamma\varepsilon(\lambda\theta_2 + \eta_2) &= \gamma(\lambda g_2 + h_2).\end{aligned} \tag{5.7}$$

Because Ω is a disjoint union of $\Omega_+(\mu)$, Ω_0, and $\Omega_-(\mu)$ and because by assumption $\Omega_0 = \emptyset$, it holds that

$$\max'(\alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2}; x) = \begin{cases} x, & \text{if } (i_1, i_2) \in \Omega_+(\mu),\\ 0, & \text{if } (i_1, i_2) \in \Omega_-(\mu),\end{cases}$$

i.e., the mapping $\max'(\alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2}; \,\cdot\,)$ is linear for all $(i_1, i_2) \in \Omega$. Thus, we may rewrite the system in (5.7) as

$$\begin{aligned}\max'\big(\alpha_1 \oplus \alpha_2 - c; (\lambda\theta_1 + \eta_1) \oplus (\lambda\theta_2 + \eta_2)\big)\ \mathbb{1} + \gamma\varepsilon(\lambda\theta_1 + \eta_1) &= \gamma(\lambda g_1 + h_1),\\ \max'\big(\alpha_1 \oplus \alpha_2 - c; (\lambda\theta_1 + \eta_1) \oplus (\lambda\theta_2 + \eta_2)\big)^\top \mathbb{1} + \gamma\varepsilon(\lambda\theta_2 + \eta_2) &= \gamma(\lambda g_2 + h_2).\end{aligned}$$

Proposition 5.5 then implies that

$$\mathcal{F}'_{\gamma,\varepsilon}(\mu; \lambda g + h) = (\lambda\theta_1 + \eta_1, \lambda\theta_2 + \eta_2) = \lambda\mathcal{F}'_{\gamma,\varepsilon}(\mu; g) + \mathcal{F}'_{\gamma,\varepsilon}(\mu; h),$$

i.e., linearity of $\mathcal{F}'_{\gamma,\varepsilon}(\mu; \,\cdot\,)$.

To show the opposite implication, let us assume that there exists some $(I_1, I_2) \in \Omega_0$. Let us further assume that $\mathcal{F}'_{\gamma,\varepsilon}(\mu; \,\cdot\,)$ is linear. In this case,

$$\begin{aligned}&\max'\big(\alpha_1 \oplus \alpha_2 - c; (\theta_1 + \eta_1) \oplus (\theta_2 + \eta_2)\big)\ \mathbb{1}\\ &\quad = \big(\max'(\alpha_1 \oplus \alpha_2 - c; \theta_1 \oplus \theta_2) + \max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2)\big)\ \mathbb{1}\end{aligned} \tag{5.8}$$

for all $(\theta_1, \theta_2), (\eta_1, \eta_2) \in \mathrm{Rg}\big(\mathcal{F}'_{\gamma,\varepsilon}(\mu; \,\cdot\,)\big)$. To derive a contradiction, we choose $(\hat{\theta}_1, \hat{\theta}_2), (\hat{\eta}_1, \hat{\eta}_2) \in \mathrm{Rg}\big(\mathcal{F}'_{\gamma,\varepsilon}(\mu; \,\cdot\,)\big)$ in a way that

1. $\hat{\theta}_1^{I_1} + \hat{\theta}_2^{I_2} = 1$ and $\hat{\theta}_1^{I_1} + \hat{\theta}_2^{i_2} = 0$ for all $i_2 \in \Omega_2$ with $i_2 \neq I_2$,
2. $\hat{\eta}_1^{I_1} + \hat{\eta}_2^{I_2} = -1$ and $\hat{\eta}_1^{I_1} + \hat{\eta}_2^{i_2} = 0$ for all $i_2 \in \Omega_2$ with $i_2 \neq I_2$.

The two linear systems 1. and 2. consist of n_2 linearly independent equations with $n_1 + n_2$ unknowns each. Thus both systems have at least one solution in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. In Remark 5.6, we already noted that $\mathcal{F}_{\gamma,\varepsilon}(\mu; \cdot)$ is bijective. Therefore, we can indeed find directions g and h that realize the claimed directional derivatives $(\hat{\theta}_1, \hat{\theta}_2)$ and $(\hat{\eta}_1, \hat{\eta}_2)$.

By construction and because of $(I_1, I_2) \in \Omega_0$, the left-hand side of the I_1-th equation of the system (5.8) evaluates to

$$\begin{aligned}\sum_{i_2=1}^{n_2} \max'\big(\alpha_1^{I_1} + \alpha_2^{i_2} - c_{I_1,i_2}; (\hat{\theta}_1^{I_1} + \hat{\theta}_2^{i_2}) + (\hat{\eta}_1^{I_1} + \hat{\eta}_2^{i_2})\big) \\ = \max\big\{0, (\hat{\theta}_1^{I_1} + \hat{\theta}_2^{I_2}) + (\hat{\eta}_1^{I_1} + \hat{\eta}_2^{I_2})\big\} = 0\end{aligned}$$

whereas the right-hand side of the same equation evaluates to

$$\begin{aligned}\sum_{i_2=1}^{n_2} \max'\big(\alpha_1^{I_1} + \alpha_2^{i_2} - c_{I_1,i_2}; \hat{\theta}_1^{I_1} + \hat{\theta}_2^{i_2}\big) + \max'\big(\alpha_1^{I_1} + \alpha_2^{i_2} - c_{I_1,i_2}; \hat{\eta}_1^{I_1} + \hat{\eta}_2^{i_2}\big) \\ = \max\big\{0, \hat{\theta}_1^{I_1} + \hat{\theta}_2^{I_2}\big\} + \max\big\{0, \hat{\eta}_1^{I_1} + \hat{\eta}_2^{I_2}\big\} = -1,\end{aligned}$$

contradicting (5.8) and therefore the linearity of the mapping $\mathcal{F}'_{\gamma,\varepsilon}(\mu; \cdot)$. It follows that $\mathcal{F}'_{\gamma,\varepsilon}(\mu; \cdot)$ cannot be linear and hence the assertion of the proposition. □

Remark 5.9. One can prove an analogous result in the infinite-dimensional case, i.e., one can show that the regularization of the dual problem of the regularized Kantorovich problem given in [52, Section 2.3] is directionally differentiable at any point and in any direction and we can characterize the points at which the solution operator is Gâteaux differentiable.

However, due to the lack of compactness, the proofs in this case are (unsurprisingly) more complicated, and the analysis of further differentiability properties is beyond the scope of this thesis and is the subject of future research. ○

Now that we have characterized the points at which the solution operator is differentiable, we are now able to calculate the derivative at those points. To this end, we need the following definition:

Definition 5.10. Let $\mathcal{A} \subset \Omega$ be some arbitrary index set. We then define

– the *characteristic matrix* $\chi(\mathcal{A}) \in \mathbb{R}^{n_1 \times n_2}$ of the index set $\mathcal{A}$ by

$$\chi(\mathcal{A})_{i_1,i_2} \coloneqq \begin{cases} 1, & \text{if } (i_1, i_2) \in \mathcal{A}, \\ 0, & \text{else;} \end{cases} \tag{5.9}$$

– the matrix $\mathcal{N}(\mathcal{A}) \in \mathbb{N}_0^{(n_1+n_2)\times(n_1+n_2)}$ corresponding to $\mathcal{A}$ by

$$\mathcal{N}(\mathcal{A}) \coloneqq \begin{pmatrix} \operatorname{diag}\big(\chi(\mathcal{A})\,\mathbb{1}\big) & \chi(\mathcal{A}) \\ \chi(\mathcal{A})^\top & \operatorname{diag}\big(\chi(\mathcal{A})^\top \mathbb{1}\big) \end{pmatrix}. \tag{5.10}$$

Proposition 5.11. *If $\mathcal{F}_{\gamma,\varepsilon}$ is (totally) differentiable at the point $\mu = (\mu_1, \mu_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, then its (total) derivative is given by*

$$\mathcal{F}'_{\gamma,\varepsilon}(\mu) = \gamma\big(\mathcal{N}\big(\Omega_+(\mu)\big) + \gamma\varepsilon E\big)^{-1}.$$

Herein, E denotes the identity matrix of $\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$.

Remark 5.12. Note that, in the formulation of the above proposition, we have tacitly identified the space $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with the space $\mathbb{R}^{n_1+n_2}$. Formally correct, the derivative would be

$$\mathcal{F}'_{\gamma,\varepsilon}(\mu) = \psi^{-1} \circ \gamma\big(\mathcal{N}\big(\Omega_+(\mu)\big) + \gamma\varepsilon E\big)^{-1} \circ \psi,$$

where

$$\psi\colon \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_1+n_2}, \quad (u, v) \mapsto \begin{pmatrix} u \\ v \end{pmatrix},$$

is a linear isometric isomorphism between $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and $\mathbb{R}^{n_1+n_2}$ and where the matrix $\gamma\big(\mathcal{N}\big(\Omega_+(\mu)\big) + \gamma\varepsilon E\big)^{-1}$ is identified with an automorphism on $\mathbb{R}^{n_1+n_2}$. However, here and in the following, we refrain from explicitly including ψ and ψ^{-1} for the sake of simplicity. ○

Proof of Proposition 5.11. We begin with abbreviating $\Omega_+ \coloneqq \Omega_+(\mu)$. Let $h = (h_1, h_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be an arbitrary direction and consider the corresponding directional derivative $(\eta_1, \eta_2) = \mathcal{F}'_{\gamma,\varepsilon}(\mu; h)$. By Proposition 5.5, the directional derivative satisfies

$$\begin{aligned} \max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2)\ \mathbb{1} + \gamma\varepsilon\eta_1 &= \gamma h_1, \\ \max'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2)^\top \mathbb{1} + \gamma\varepsilon\eta_2 &= \gamma h_2, \end{aligned}$$

where again $(\alpha_1, \alpha_2) = \mathcal{F}_{\gamma,\varepsilon}(\mu)$. Applying Proposition 5.8, the above system is equivalent to

$$\begin{aligned} \sum_{i_2\,:\,(i_1,i_2)\in\Omega_+} (\eta_1^{i_1} + \eta_2^{i_2}) + \gamma\varepsilon\eta_1^{i_1} &= \gamma h_1^{i_1} \quad \text{for all } i_1 \in \Omega_1, \\ \sum_{i_1\,:\,(i_1,i_2)\in\Omega_+} (\eta_1^{i_1} + \eta_2^{i_2}) + \gamma\varepsilon\eta_2^{i_2} &= \gamma h_2^{i_2} \quad \text{for all } i_2 \in \Omega_2. \end{aligned}$$

Comparing this with the definition of the matrix $\chi(\Omega_+)$ from (5.9), we can rewrite this equivalently as

$$\begin{aligned} \operatorname{diag}\big(\chi(\Omega_+)\ \mathbb{1}\big)\eta_1 + \chi(\Omega_+)\ \eta_2 + \gamma\varepsilon\eta_1 &= \gamma h_1 \\ \operatorname{diag}\big(\chi(\Omega_+)^\top \mathbb{1}\big)\eta_2 + \chi(\Omega_+)^\top \eta_1 + \gamma\varepsilon\eta_2 &= \gamma h_2. \end{aligned}$$

Plugging in the definition from (5.10) then yields that

$$\big(\mathcal{N}(\Omega_+) + \gamma\varepsilon E\big) \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \gamma \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}. \tag{5.11}$$

The matrix $\mathcal{N}(\Omega_+)$ is by construction nonnegative, symmetric, and diagonally dominant, i.e., it satisfies

$$\mathcal{N}(\Omega_+)_{l,l} \geq \sum_{\substack{k\in\{1,\dots,n_1+n_2\},\\ k\neq l}} \mathcal{N}(\Omega_+)_{l,k} = \sum_{\substack{k\in\{1,\dots,n_1+n_2\},\\ k\neq l}} \mathcal{N}(\Omega_+)_{k,l}$$

for all $l = 1, \dots, n_1 + n_2$. Such matrices are known to be positive semidefinite, see e.g. Gershgorin's circle theorem. Consequently, the matrix $\mathcal{N}(\Omega_+) + \gamma\varepsilon E$ is positive definite, thus invertible, and we can solve equation (5.11) via

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \gamma\big(\mathcal{N}(\Omega_+) + \gamma\varepsilon E\big)^{-1} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}.$$

In light of Remark 5.12, this shows that $\mathcal{F}_{\gamma,\varepsilon}$ is Gâteaux differentiable at μ and that its Gâteaux derivative is $\gamma\big(\mathcal{N}(\Omega_+) + \gamma\varepsilon E\big)^{-1}$. However, since $\mathcal{F}_{\gamma,\varepsilon}$ was assumed to be (totally) differentiable, the Gâteaux derivative and (total) derivative coincide. □

Now that we have precisely characterized the points in which the solution operator is differentiable and have a representation of its derivative we concern ourselves, in a next step, with the points of non-differentiability.

Because $\mathcal{F}_{\gamma,\varepsilon}$ is globally Lipschitz continuous, at each of those points we can find at least a generalized Jacobian in Clarke's sense, see [22, Proposition 2.6.2]. For this reason, we have the following definition.

Definition 5.13 ([74, Definition 2.1] and [22, Definition 2.6.1])**.** Given $m, n \in \mathbb{N}$, let $f\colon \mathbb{R}^m \to \mathbb{R}^n$ be a locally Lipschitz continuous function. By Rademacher's theorem, f is differentiable almost everywhere on $\mathbb{R}^m$. We denote the set of points at which f is differentiable by $\mathcal{D}_f \subset \mathbb{R}^m$.

We call the set

$$\partial_B f(x) \coloneqq \Big\{ \lim_{k\to\infty} f'(x)\colon (x_k)_{k\in\mathbb{N}} \subset \mathcal{D}_f,\ x_k \to x \text{ as } k \to \infty \Big\} \tag{5.12}$$

the *Bouligand subdifferential* of f at the point $x \in \mathbb{R}^m$. It relates to Clarke's *generalized Jacobian* of f at x via the definition

$$\partial f(x) \coloneqq \operatorname{co}\big(\partial_B f(x)\big),$$

where $\operatorname{co}(M)$ denotes the convex hull of some set M. For every point $x \in \mathbb{R}^m$, the Bouligand subdifferential $\partial_B f(x)$ is a nonempty and compact subset of $\mathbb{R}^{n\times m}$, see e.g. [27, Proposition 4.3.1].

In the following, we aim to characterize the Bouligand subdifferential of $\mathcal{F}_{\gamma,\varepsilon}$, the solution operator of the regularized dual problem ($\mathrm{HD}^{\varepsilon}_{\gamma}$). First, however, we need the following lemma which provides a useful property of convergent sequences of $\mathcal{F}_{\gamma,\varepsilon}$'s derivatives.

Lemma 5.14. *Let $\mu = (\mu_1, \mu_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be an arbitrary point and consider a sequence of points $(\mu_{1,k}, \mu_{2,k})_{k\in\mathbb{N}} \subset \mathcal{D}_{\mathcal{F}_{\gamma,\varepsilon}}$ with $\mu_k \coloneqq (\mu_{1,k}, \mu_{2,k}) \to \mu$ as $k \to \infty$. Then $\mathcal{F}'_{\gamma,\varepsilon}(\mu_k) \to G \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ if and only if there exists some $K \in \mathbb{N}$ such that*

1. *$\Omega_+(\mu_k) = \Omega_+(\mu_K)$ and*
2. *$\mathcal{F}'_{\gamma,\varepsilon}(\mu_k) = G = \gamma\big(\mathcal{N}\big(\Omega_+(\mu_K)\big) + \gamma\varepsilon E\big)^{-1}$*

for all $k \geq K$.

Proof. For all $k \in \mathbb{N}$, we abbreviate $\Omega^k_+ \coloneqq \Omega_+(\mu_k)$. In order to show the forward implication, let $\mathcal{F}'_{\gamma,\varepsilon}(\mu_k) \to G$ as $k \to \infty$. By Proposition 5.11, $\mathcal{F}'_{\gamma,\varepsilon}(\mu_k) = \gamma\big(\mathcal{N}(\Omega^k_+) + \gamma\varepsilon E\big)^{-1}$ for each $k \in \mathbb{N}$. By definition, the entries of each of the matrices $\mathcal{N}(\Omega^k_+) \in \mathbb{N}_0^{(n_1+n_2)\times(n_1+n_2)}$ are bounded by

$$0 \leq \big(\mathcal{N}(\Omega^k_+)\big)_{i_1,i_2} \leq \max\{n_1, n_2\} \quad \text{for all } (i_1, i_2) \in \Omega.$$

We can therefore find a convergent subsequence $\big(\mathcal{N}(\Omega_+^{k_l})\big)_{l\in\mathbb{N}} \subset \big(\mathcal{N}(\Omega_+^{k})\big)_{k\in\mathbb{N}}$ and some matrix $H \in \mathbb{N}_0^{(n_1+n_2)\times(n_1+n_2)}$ so that $\mathcal{N}(\Omega_+^{k_l}) \to H$ as $l \to \infty$. Since this is a convergence of integer matrices, there must exist some $L \in \mathbb{N}$ such that $\mathcal{N}(\Omega_+^{k_l}) = H$ for all $l \geq L$.

The definitions in (5.10) and (5.9) then imply that $\chi(\Omega_+^{k_l}) = \chi(\Omega_+^{k_L})$ and in turn $\Omega_+^{k_l} = \Omega_+^{k_L}$, respectively, for all $l \geq L$.

This already yields the desired representation of the limit G, since

$$G = \lim_{k\to\infty} \mathcal{F}'_{\gamma,\varepsilon}(\mu_k) = \lim_{l\to\infty} \mathcal{F}'_{\gamma,\varepsilon}(\mu_{k_l}) = \gamma\big(\mathcal{N}(\Omega_+^{k_L}) + \gamma\varepsilon E\big)^{-1}.$$

The convergence of the entire sequence $\big(\mathcal{N}(\Omega_+^{k})\big)_{k\in\mathbb{N}}$ and followingly the claim of the lemma then follows from the uniqueness of the limit

$$H = \mathcal{N}(\Omega_+^{k_L}) = \gamma G^{-1} - \gamma\varepsilon E$$

and Lemma D.5.

The reverse implication is trivial and therefore omitted. □

In the following theorem we derive a more convenient description of the Bouligand subdifferential at the points where $\mathcal{F}_{\gamma,\varepsilon}$ is not differentiable. However, we first need the following definition that greatly simplifies the notation of that description.

Definition 5.15. Let $\mathcal{A}, \mathcal{B} \subset \Omega$ be given index sets. We say that $\mathcal{A}$ has an *outer structure* w.r.t. $\mathcal{B}$, if

- $\mathcal{A} \subset \mathcal{B}$ and
- there exist vectors $v_1 \in \mathbb{R}^{n_1}$ and $v_2 \in \mathbb{R}^{n_2}$ in a way that $v_1^{i_1} + v_2^{i_2} > 0$ for all $(i_1, i_2) \in \mathcal{A}$ and $v_1^{i_1} + v_2^{i_2} < 0$ for all $(i_1, i_2) \in \mathcal{B} \setminus \mathcal{A}$.

We will often shorten the above notation by simply writing $(v_1 \oplus v_2)_{\mathcal{A}} > 0$ and $(v_1 \oplus v_2)_{\mathcal{B}\setminus\mathcal{A}}$.

Remark 5.16. The number and geometry of sets that have an outer structure w.r.t. some set $\mathcal{B}$ critically depends on the structure of $\mathcal{B}$ itself. Even if $\mathcal{A}$ has an outer structure w.r.t. $\mathcal{B} \subset \mathcal{B}'$, $\mathcal{A}$ generally does not need to have an outer structure w.r.t. $\mathcal{B}'$ because introducing a larger set also imposes more constraints on the set $\mathcal{A}$.

It is indeed nontrivial to determine whether a given set has an outer structure w.r.t. some other set. However, independently of the $\mathcal{B}$, both the empty set and $\mathcal{B}$ itself have an outer structure w.r.t. $\mathcal{B}$ (simply choose (v_1, v_2) to be $(-\mathbb{1}, \mathbb{0})$ and $(\mathbb{1}, \mathbb{0})$, respectively). Moreover, if $\mathcal{A}$ has an outer structure w.r.t. $\mathcal{B}$, then also $\mathcal{B} \setminus \mathcal{A}$ has an outer structure w.r.t. $\mathcal{B}$. ○

Theorem 5.17. *Let $\mu = (\mu_1, \mu_2) \notin \mathcal{D}_{\mathcal{F}_{\gamma,\varepsilon}}$ be a point at which $\mathcal{F}_{\gamma,\varepsilon}$ is not differentiable and abbreviate $\Omega_+ \coloneqq \Omega_+(\mu)$ and $\Omega_0 \coloneqq \Omega_0(\mu)$. Then, the Bouligand subdifferential of $\mathcal{F}_{\gamma,\varepsilon}$ at μ is given by*

$$\begin{aligned}&\partial_B \mathcal{F}_{\gamma,\varepsilon}(\mu)\\ &\quad= \Big\{\gamma\big(\mathcal{N}(\Omega_+ \cup \mathcal{A}) + \gamma\varepsilon E\big)^{-1} \colon \mathcal{A} \text{ has an outer structure w.r.t. } \Omega_0\Big\}.\end{aligned} \tag{5.13}$$

Proof. On the one hand, let $G \in \partial_B \mathcal{F}_{\gamma,\varepsilon}(\mu)$ be arbitrary. Then, there exists a sequence of points $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{D}_{\mathcal{F}_{\gamma,\varepsilon}}$ such that $\mu_k \to \mu$ and $\mathcal{F}'_{\gamma,\varepsilon}(\mu_k) \to G$ as $k \to \infty$. We abbreviate $\Omega_+^k := \Omega_+(\mu_k)$, $\Omega_-^k := \Omega_-(\mu_k)$ and $\Omega_0^k := \Omega_0(\mu_k)$ for all $k \in \mathbb{N}$. By Lemma 5.14, there exists an index $K_1 \in \mathbb{N}$ for which

$$G = \gamma\big(\mathcal{N}\big(\Omega_+^{K_1}\big) + \gamma\varepsilon E\big)^{-1} \quad \text{and} \quad \Omega_+^k = \Omega_+^{K_1}$$

for all $k \geq K_1$. Moreover, we have seen in Lemma 5.4 that $\mathcal{F}_{\gamma,\varepsilon}$ is (Lipschitz-) continuous. Hence,

$$(\alpha_{1,k}, \alpha_{2,k}) := \mathcal{F}_{\gamma,\varepsilon}(\mu_k) \xrightarrow[k\to\infty]{} \mathcal{F}_{\gamma,\varepsilon}(\mu) =: (\alpha_1, \alpha_2)$$

and, owing to the continuity of the $\oplus$-operator,

$$(\alpha_{1,k} \oplus \alpha_{2,k} - c)_{i_1,i_2} \xrightarrow[k\to\infty]{} (\alpha_1 \oplus \alpha_2 - c)_{i_1,i_2} \quad \text{for all } (i_1, i_2) \in \Omega.$$

Since $(\alpha_1 \oplus \alpha_2 - c)_{i_1,i_2} > 0$ for all $(i_1, i_2) \in \Omega_+$ and $(\alpha_1 \oplus \alpha_2 - c)_{i_1,i_2} < 0$ for all $(i_1, i_2) \in \Omega_-$, one can find another index $K_2 \in \mathbb{N}$ such that $\Omega_+ \subset \Omega_+^k$ and $\Omega_- \subset \Omega_-^k$ for all $k \geq K_2$. In particular, if we set $K := \max\{K_1, K_2\}$, then

$$\Omega_+ \subset \Omega_+^K, \quad \Omega_- \subset \Omega_-^K, \quad \text{and} \quad G = \gamma\big(\mathcal{N}(\Omega_+^K) + \gamma\varepsilon E\big)^{-1}. \tag{5.14}$$

Let us abbreviate $\mathcal{A} := \Omega_+^K \setminus \Omega_+$. Because of (5.14) and

$$\Omega_+ \,\dot\cup\, \Omega_0 \,\dot\cup\, \Omega_- = \Omega = \Omega_+^K \,\dot\cup\, \Omega_-^K$$

(remember that μ_K is a point of differentiability and therefore $\Omega_0^K = \emptyset$), we find that $\mathcal{A} \subset \Omega_0$. This together with $\Omega_0 \setminus \mathcal{A} = \Omega_0 \setminus \Omega_+^K \subset \Omega_-^K$ yields that

$$\big((\alpha_{1,K} - \alpha_1) \oplus (\alpha_{2,K} - \alpha_2)\big)_{i_1,i_2} = \big((\alpha_{1,K} \oplus \alpha_{2,K} - c) - (\alpha_1 \oplus \alpha_2 - c)\big)_{i_1,i_2}$$
$$\begin{cases} > 0 & \text{if } (i_1, i_2) \in \mathcal{A}, \\ < 0 & \text{if } (i_1, i_2) \in \Omega_0 \setminus \mathcal{A}, \end{cases}$$

which shows that

$$G = \gamma\big(\mathcal{N}(\Omega_+^K) + \gamma\varepsilon E\big)^{-1} = \gamma\big(\mathcal{N}(\Omega_+ \cup \mathcal{A}) + \gamma\varepsilon E\big)^{-1}$$

is an element of the set on the right-hand side of (5.13) for $\mathcal{A} \subset \Omega_0$ and $(v_1, v_2) = (\alpha_{1,K} - \alpha_1, \alpha_{2,K} - \alpha_2)$.

On the other hand, let $\mathcal{A}$ have an outer structure w.r.t. Ω_0. Then there exist $(v_1, v_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with $(v_1 \oplus v_2)_{\mathcal{A}} > 0$ as well as $(v_1 \oplus v_2)_{\Omega_0\setminus\mathcal{A}} < 0$. We abbreviate $G := \gamma(\mathcal{N}(\Omega_+ \cup \mathcal{A}) + \gamma\varepsilon E)^{-1}$ and show that there exists a sequence of points $(\mu_k)_{k\in\mathbb{N}} \subset \mathcal{D}_{\mathcal{F}_{\gamma,\varepsilon}}$ that satisfies $\mu_k \to \mu$ and $\mathcal{F}'_{\gamma,\varepsilon}(\mu_k) \to G$ as $k \to \infty$, i.e., $G \in \partial_B \mathcal{F}_{\gamma,\varepsilon}(\mu)$. To this end, we define

$$\delta := \frac{1}{2\|v_1 \oplus v_2\|_\infty} \cdot \min_{(I_1,I_2)\in\Omega_+\cup\Omega_-} \big|(\alpha_1 \oplus \alpha_2 - c)_{I_1,I_2}\big| \;\in\; (0, \infty)$$

and consider the sequence of points $(\alpha_{1,k}, \alpha_{2,k})_{k\in\mathbb{N}} \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ defined by

$$(\alpha_{1,k}, \alpha_{2,k}) := (\alpha_1, \alpha_2) + \frac{\delta}{k}(v_1, v_2) \quad \text{for all } k \in \mathbb{N}. \tag{5.15}$$

Obviously,

$$(\alpha_{1,k}, \alpha_{2,k}) \;\to\; (\alpha_1, \alpha_2) = \mathcal{F}_{\gamma,\varepsilon}(\mu)$$

as $k \to \infty$ and, following Remark 5.2, this is also true for the sequence of corresponding marginals:

$$\mu_k \coloneqq \mathcal{F}^{-1}_{\gamma,\varepsilon}(\alpha_{1,k}, \alpha_{2,k}) \;\to\; \mathcal{F}^{-1}_{\gamma,\varepsilon}(\alpha_1, \alpha_2) = \mu$$

as $k \to \infty$.

The construction in (5.15) yields that

$$(\alpha_{1,k} \oplus \alpha_{2,k} - c)_{i_1,i_2} = (\alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2}) + \frac{\delta}{k}(v_1^{i_1} + v_2^{i_2})$$

$$\begin{cases} > (\alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2}) & \\ \quad -\frac{1}{2k}\min_{(I_1,I_2)\in\Omega_+}(\alpha_1 \oplus \alpha_2 - c)_{I_1,I_2} > 0, & \text{if } (i_1,i_2) \in \Omega_+, \\ = \frac{\delta}{k}(v_1^{i_1} + v_2^{i_2}) > 0, & \text{if } (i_1,i_2) \in \mathcal{A}, \\ = \frac{\delta}{k}(v_1^{i_1} + v_2^{i_2}) < 0, & \text{if } (i_1,i_2) \in \Omega_0 \setminus \mathcal{A}, \\ < (\alpha_1^{i_1} + \alpha_2^{i_2} - c_{i_1,i_2}) & \\ \quad +\frac{1}{2k}\min_{(I_1,I_2)\in\Omega_-} -(\alpha_1 \oplus \alpha_2 - c)_{I_1,I_2} < 0, & \text{if } (i_1,i_2) \in \Omega_-, \end{cases}$$

i.e.,

$$(\alpha_{1,k} \oplus \alpha_{2,k} - c)_{i_1,i_2} \begin{cases} > 0, & \text{if } (i_1,i_2) \in \Omega_+ \cup \mathcal{A}, \\ < 0, & \text{if } (i_1,i_2) \in \Omega_- \cup (\Omega_0 \setminus \mathcal{A}), \end{cases} \tag{5.16}$$

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

Because of (5.16), we find that $\Omega_+^k \coloneqq \Omega_+(\mu_k) = \Omega_+ \cup \mathcal{A}$ and $\Omega_-^k \coloneqq \Omega_-(\mu_k) = \Omega_- \cup (\Omega_0 \setminus \mathcal{A})$ independently of k. Also,

$$\Omega_+^k \cup \Omega_-^k = \Omega_+ \cup \Omega_0 \cup \Omega_- = \Omega,$$

i.e., $\Omega_0(\mu_k) = \emptyset$. Thus, Proposition 5.8 and Proposition 5.11 imply that

$$\begin{aligned} &\lim_{k\to\infty} \mathcal{F}_{\gamma,\varepsilon}(\mu_k) \\ &\quad = \lim_{k\to\infty} \gamma\big(\mathcal{N}(\Omega_+^k) + \gamma\varepsilon E\big)^{-1} = \gamma\big(\mathcal{N}(\Omega_+ \cup \mathcal{A}) + \gamma\varepsilon E\big)^{-1} = G. \end{aligned}$$

Comparing this with the definition of the Bouligand subdifferential shows that $G \in \partial_B \mathcal{F}_{\gamma,\varepsilon}(\mu)$ as claimed. □

Remark 5.18. In light of Theorem 5.17, we want to mention the following:

- We can always obtain at least two elements of $\partial_B \mathcal{F}_{\gamma,\varepsilon}$ by choosing the sets $\mathcal{A}_1 = \Omega_0$ and $\mathcal{A}_2 = \emptyset$, see Remark 5.16.
- The theorem does not provide a description of $\partial_B \mathcal{F}_{\gamma,\varepsilon}(\mu)$ for $\mu \in \mathcal{D}_{\mathcal{F}_{\gamma,\varepsilon}}$, i.e., for points at which $\mathcal{F}_{\gamma,\varepsilon}$ is differentiable. Even though, in this case, $\Omega_0 = \emptyset$ (by Proposition 5.8) implies that $\mathcal{A} = \emptyset$ is the only subset of Ω_0 and therefore that the right-hand side of (5.13) reduces to

$$\big\{\gamma\big(\mathcal{N}(\Omega_+) + \gamma\varepsilon E\big)^{-1}\big\} = \{\mathcal{F}'_{\gamma,\varepsilon}(\mu)\},$$

we only find that

$$\{\mathcal{F}'_{\gamma,\varepsilon}(\mu)\} \subset \partial_B(\mu), \tag{5.17}$$

since, in general, a stronger notion of differentiability, e.g. strict differentiability or continuous differentiability, see [27, Proposition 4.3.4] or [74, Proposition 2.2], respectively, is required for the Bouligand subdifferential to be a singleton.

In fact, in [27, Exercise 4.3.3], the authors give an example of a one-dimensional function that is differentiable but not strictly (and therefore not continuously) differentiable at the point 0, and which has a non-singleton Bouligand subdifferential at this point.

Nevertheless, the inclusion from (5.17) guarantees that we can even at differentiable points find an element of the Bouligand subdifferential by computing the corresponding derivative. This will prove useful in Chapter 6, where a nonsmooth optimization algorithm based on the Bouligand subdifferential will be applied.

○

In the following subchapter, we take the IP approach one step further by concatenating the solution operator of the regularized dual problem with a map that is motivated by the first-order system (4.2). This way we obtain a regularized marginal-to-transport-plan mapping for which we can compute subgradients in a convenient way.

5.2 The Regularized Marginal-to-Transport-Plan Mapping

To progress with the IP approach, we are going to use the results from the previous subchapter to obtain a mapping from the marginals to a unique optimal transport plan that entails certain differentiability properties.

For this purpose, given $\gamma > 0$ and some cost matrix $c \in \mathbb{R}^{n_1 \times n_2}$, let us consider the *dual-variable-to-transport-plan mapping*

$$\mathcal{P}_\gamma \colon \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_1 \times n_2}, \quad (v_1, v_2) \mapsto \frac{1}{\gamma}(v_1 \oplus v_2 - c)_+.$$

For any given point $(v_1, v_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, Theorem 4.9 guarantees that the transport plan defined by $\pi \coloneqq \mathcal{P}_\gamma(v_1, v_2)$ uniquely solves the regularized Hitchcock problem (H_γ) w.r.t. the marginals $\pi\mathbb{1}$ and $\pi^\top\mathbb{1}$ as well as the cost c. This, in turn, implies that v_1 and v_2 indeed are the dual variables to π, justifying the name of the above mapping.

Given the mapping $\mathcal{P}_\gamma$, we now define the *regularized marginal-to-transport-plan mapping* by $\mathcal{S}_{\gamma,\varepsilon} \coloneqq \mathcal{P}_\gamma \circ \mathcal{F}_{\gamma,\varepsilon}$, i.e.,

$$\mathcal{S}_{\gamma,\varepsilon} \colon \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_1 \times n_2}, \quad (\mu_1, \mu_2) \mapsto \frac{1}{\gamma}(\alpha_1 \oplus \alpha_2 - c)_+,$$

where $(\alpha_1, \alpha_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ denotes the solution to the regularized dual problem in ($\mathrm{HD}^\varepsilon_\gamma$). The symbol we have chosen for the regularized marginal-to-transport-plan mapping already hints at the fact that we want to treat this mapping as the

solution operator we need to realize the IP approach for the bilevel Hitchcock problem. The principal idea of this construction is the following:

As we indicated at the beginning of Chapter 5, we would like to replace the Hitchcock problem in the constraints of (BH) by a solution operator with certain differentiability properties. Unfortunately, the explicit description of the optimal regularized transport plan from Theorem 4.9 depends on the (nonunique) dual variables, which can lead to problems when trying to compute derivatives.

For this reason, we make a detour via the regularization of the dual problem given in Subchapter 5.1. We have studied its differentiability properties to the extent that we can deduce the same differentiability properties of the regularized marginal-to-transport plan mapping in the present subchapter.

Moreover, the authors of [52] proved that a standard Tikhonov regularization approach can be used to approximate solutions of the regularized Kantorovich problem sufficiently well. Since we adopted the same approach for the regularization of the finite-dimensional dual problem in Subchapter (5.1), we can assume that when the regularization vanishes, the regularized transport plans can also be approximated to an arbitrary precision.

We therefore expect that, owing to the Tikhonov regularization, the regularized marginal-to-transport-plan mapping not only behaves well numerically, but also allows the use of standard nonsmooth optimization techniques to compute the regularized transport plans. More on this topic can be found in Chapter 6.

The next theorem shows that the regularized marginals-on-transport-plan mapping inherits all relevant properties of the solution operator of the regularized dual problem from Subchapter 5.1. First, however, we need the following definition.

Definition 5.19. Let some subset $\mathcal{A} \subset \Omega$ be given. We then define the *mask* by

$$\mathcal{H}(\mathcal{A})\colon \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}, \quad M \mapsto \left(\begin{cases} M_{i_1,i_2}, & \text{if } (i_1,i_2) \in \mathcal{A}, \\ 0, & \text{else.} \end{cases} \right)_{(i_1,i_2)\in\Omega}$$

Simply put, $\mathcal{H}(\mathcal{A})$ manipulates a matrix by setting those entries whose indices belong to $\Omega \setminus \mathcal{A}$ to 0. It corresponds to entrywise multiplication with the characteristic matrix $\chi(\mathcal{A})$ from (5.9), i.e.,

$$\mathcal{H}(\mathcal{A})(M) = \left(\chi(\mathcal{A})_{i_1,i_2} M_{i_1,i_2}\right)_{(i_1,i_2)\in\Omega}.$$

Theorem 5.20. *The regularized marginal-to-transport-plan mapping has the following properties:*

1. *$\mathcal{S}_{\gamma,\varepsilon}$ is (globally) Lipschitz continuous.*
2. *$\mathcal{S}_{\gamma,\varepsilon}$ is differentiable almost everywhere on $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. We denote the corresponding set of differentiable points by $\mathcal{D}_{\mathcal{S}_{\gamma,\varepsilon}}$.*
3. *$\mathcal{S}_{\gamma,\varepsilon}$ is Hadamard differentiable and its directional derivative at some point $\mu = (\mu_1, \mu_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ in the direction $h = (h_1, h_2) \in \mathbb{R}^m \times \mathbb{R}^{n_2}$ is given by*

$$\mathcal{S}'_{\gamma,\varepsilon}(\mu; h) = \frac{1}{\gamma} \max{}'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2) \in \mathbb{R}^{n_1\times n_2},$$

where $(\alpha_1, \alpha_2) = \mathcal{F}_{\gamma,\varepsilon}(\mu)$ and $(\eta_1, \eta_2) = \mathcal{F}'_{\gamma,\varepsilon}(\mu; h)$.

4. *$\mathcal{S}_{\gamma,\varepsilon}$ is differentiable at the point $\mu \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ if and only if $\Omega_0(\mu) = \emptyset$, in particular, $\mathcal{D}_{\mathcal{S}_{\gamma,\varepsilon}} = \mathcal{D}_{\mathcal{F}_{\gamma,\varepsilon}}$.*

5. *If $\mathcal{S}_{\gamma,\varepsilon}$ is (totally) differentiable at some point $\mu \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, then its (total) derivative is given by*

$$\mathcal{S}'_{\gamma,\varepsilon}(\mu) = \mathcal{H}\big(\Omega_+(\mu)\big) \circ \oplus \circ \big(\mathcal{N}\big(\Omega_+(\mu)\big) + \gamma\varepsilon E\big)^{-1},$$

where the matrix $\big(\mathcal{N}\big(\Omega_+(\mu)\big) + \gamma\varepsilon E\big)^{-1}$ again is understood as a linear automorphism on $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, see Remark 5.12.

6. *The Bouligand subdifferential of $\mathcal{S}_{\gamma,\varepsilon}$ at some point $\mu \notin \mathcal{D}_{\mathcal{S}_{\gamma,\varepsilon}}$ with $\Omega_+ \coloneqq \Omega_+(\mu)$ and $\Omega_0 \coloneqq \Omega_0(\mu)$ is given by*

$$\partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu) = \left\{ \mathcal{H}(\Omega_+ \cup \mathcal{A}) \circ \oplus \circ \big(\mathcal{N}(\Omega_+ \cup \mathcal{A}) + \gamma\varepsilon E\big)^{-1} : \begin{array}{c} \mathcal{A} \text{ has an outer} \\ \text{structure w.r.t. } \Omega_0 \end{array} \right\}.$$

In particular, if $G \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu)$ then there exists some $\tilde{G} = \gamma\big(\mathcal{N}(\Omega_+ \cup \mathcal{A}) + \gamma\varepsilon E\big)^{-1} \in \partial_B \mathcal{F}_{\gamma,\varepsilon}(\mu)$ such that

$$G = \frac{1}{\gamma} \mathcal{H}(\Omega_+ \cup \mathcal{A}) \circ \oplus \circ \tilde{G},$$

where $\mathcal{A}$ is the subset of Ω_0 that realizes $\tilde{G}$.

Proof. Ad *1.*: Let $(v_1, v_2), (w_1, w_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be arbitrary. Because the mapping $(\,\cdot\,)_+ \colon \mathbb{R} \ni x \mapsto \max\{0, x\} \in \mathbb{R}_+$ is (globally) Lipschitz continuous with Lipschitz constant equal to 1, it holds that

$$|(v_1^{i_1} + v_2^{i_2} - c_{i_1,i_2})_+ - (w_1^{i_1} + w_2^{i_2} - c_{i_1,i_2})_+| \leq |v_1^{i_1} - w_1^{i_1} + v_2^{i_2} - w_2^{i_2}|$$

for all $(i_1, i_2) \in \Omega$. Therefore and because $(a_+ - b_+)^2 \leq (a - b)^2$ for all $a, b \in \mathbb{R}$,

$$\begin{aligned} \|\mathcal{P}_\gamma(v_1, v_2) - \mathcal{P}_\gamma(w_1, w_2)\|_F^2 &= \frac{1}{\gamma^2} \|(v_1 \oplus v_2 - c)_+ - (w_1 \oplus w_2 - c)_+\|_F^2 \\ &\leq \frac{1}{\gamma^2} \|(v_1 - w_1) \oplus (v_2 - w_2)\|_F^2 \\ &\leq \frac{2\max\{n_1, n_2\}}{\gamma^2} \big(\|v_1 - w_1\|_{\mathbb{R}^{n_1}}^2 + \|v_2 - w_2\|_{\mathbb{R}^{n_2}}^2\big) \\ &= L_{\mathcal{P}_\gamma}^2 \|(v_1, v_2) - (w_1, w_2)\|_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}}^2, \end{aligned}$$

see Remark 4.1. Thus, $\mathcal{P}_\gamma$ is (globally) Lipschitz continuous with Lipschitz constant $L_{\mathcal{P}_\gamma} = \gamma^{-1}\sqrt{2\max\{n_1, n_2\}} > 0$. Being the composition of (globally) Lipschitz continuous mappings, see Lemma 5.4, the mapping $\mathcal{S}_{\gamma,\varepsilon}$ itself is (globally) Lipschitz continuous.

Ad *2.*: Follows from *1.* and Rademacher's theorem.

Ad *3.*: Given arbitrary points $v = (v_1, v_2)$ and directions $h = (h_1, h_2)$, we compute $\mathcal{P}_\gamma$'s directional derivative by

$$\big(\mathcal{P}'_\gamma(v; h)\big)_{i_1,i_2}$$

$$= \frac{1}{\gamma} \lim_{t \searrow 0} \frac{\big((v_1^{i_1} + t h_1^{i_1}) + (v_2^{i_2} + t h_2^{i_2}) - c_{i_1,i_2}\big)_+ - (v_1^{i_1} + v_2^{i_2} - c_{i_1,i_2})_+}{t}$$
$$= \frac{1}{\gamma} \max{}'\big(v_1^{i_1} + v_2^{i_2} - c_{i_1,i_2}; h_1^{i_1} + h_2^{i_2}\big)$$

for all $(i_1, i_2) \in \Omega$, i.e.,

$$\mathcal{P}'_\gamma(v; h) = \frac{1}{\gamma} \max{}'(v_1 \oplus v_2 - c; h_1 \oplus h_2).$$

Both $\mathcal{P}_\gamma$ and $\mathcal{F}_{\gamma,\varepsilon}$ are (globally) Lipschitz continuous and directionally differentiable everywhere and in every direction. Thus, according to Lemma D.7, both $\mathcal{P}_\gamma$ and $\mathcal{F}_{\gamma,\varepsilon}$ are Hadamard differentiable. We can therefore use the chain rule for Hadamard differentiable mappings from [70, Proposition 3.6] to conclude that $\mathcal{S}_{\gamma,\varepsilon}$ itself is Hadamard differentiable and has the directional derivative

$$\begin{aligned}\mathcal{S}'_{\gamma,\varepsilon}(\mu; h) &= \mathcal{P}'_\gamma\big(\mathcal{F}_{\gamma,\varepsilon}(\mu); \mathcal{F}'_{\gamma,\varepsilon}(\mu; h)\big) \\ &= \frac{1}{\gamma} \max{}'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2),\end{aligned}$$

as claimed.

Ad *4.*: Following the same rationale as in the proof of Proposition 5.8, it is sufficient to show the equivalence

$$\Omega_0(\mu) = \emptyset \iff h \mapsto \mathcal{S}'_{\gamma,\varepsilon}(\mu; h) \text{ is linear.}$$

On the one hand, if $\Omega_0(\mu) = \emptyset$, then (5.4) implies that

$$\begin{aligned}\mathcal{S}'_{\gamma,\varepsilon}(\mu; h) &= \frac{1}{\gamma} \max{}'(\alpha_1 \oplus \alpha_2 - c; \eta_1 \oplus \eta_2) \\ &= \frac{1}{\gamma} \left(\begin{cases} \eta_1^{i_1} + \eta_2^{i_2} & \text{if } (i_1, i_2) \in \Omega_+, \\ 0 & \text{if } (i_1, i_2) \in \Omega_-, \end{cases} \right)_{(i_1,i_2) \in \Omega}\end{aligned}$$

for all directions $h \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with $(\eta_1, \eta_2) = \mathcal{F}'_{\gamma,\varepsilon}(\mu; h)$. Because the mapping $h \mapsto \mathcal{F}'_{\gamma,\varepsilon}(\mu; h)$ is linear, see the proof of Proposition 5.8, the mapping $h \mapsto \mathcal{S}'_{\gamma,\varepsilon}(\mu; h)$ is linear w.r.t. to h.

On the other hand, assume that $h \mapsto \mathcal{S}'_{\gamma,\varepsilon}(\mu; h)$ is linear and let $h \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be arbitrary. By linearity,

$$\begin{aligned}0 &= \mathcal{S}'_{\gamma,\varepsilon}(\mu; h) + \mathcal{S}'_{\gamma,\varepsilon}(\mu; -h) \\ &= \frac{1}{\gamma} \begin{cases} (\eta_1^{i_1} + \eta_2^{i_2}) + (\theta_1^{i_1} + \theta_2^{i_2}) & \text{if } (i_1, i_2) \in \Omega_+(\mu), \\ \max\{0, \eta_1^{i_1} + \eta_2^{i_2}\} + \max\{0, \theta_1^{i_1} + \theta_2^{i_2}\} & \text{if } (i_1, i_2) \in \Omega_0(\mu), \\ 0 & \text{if } (i_1, i_2) \in \Omega_-(\mu), \end{cases}\end{aligned}$$

where $(\eta_1, \eta_2) = \mathcal{F}'_{\gamma,\varepsilon}(\mu; h)$ and $(\theta_1, \theta_2) = \mathcal{F}'_{\gamma,\varepsilon}(\mu; -h)$. This implies that

$$\eta_1^{i_1} + \eta_2^{i_2} \leq 0 \quad \text{and} \quad \theta_1^{i_1} + \theta_2^{i_2} \leq 0 \tag{5.18}$$

for all $(i_1, i_2) \in \Omega_0(\mu)$. However, it is easy to construct a direction $\tilde{h}$ whose corresponding directional derivative $(\tilde{\eta}_1, \tilde{\eta}_2) = \mathcal{F}'_{\gamma,\varepsilon}(\mu; h)$ satisfies $\tilde{\eta}_1^{i_1} + \tilde{\eta}_2^{i_2} = 1$, see

e.g. the proof of Proposition 5.8. This directly contradicts (5.18). Consequently, there can be no $(i_1, i_2) \in \Omega_0(\mu)$ and therefore $\Omega_0(\mu) = \emptyset$.

Ad *5.*: This property is more or less just a corollary to Proposition 5.11. Because of *4.* and Proposition 5.8, $\mathcal{S}_{\gamma,\varepsilon}$ is differentiable in μ if and only if $\mathcal{F}_{\gamma,\varepsilon}$ is differentiable in μ. Therefore, given some arbitrary $h \in \mathbb{R}^{n_1 \times n_2}$ and $(\eta_1, \eta_2) = \mathcal{F}'_{\gamma,\varepsilon}(\mu; h)$, we rewrite

$$\eta_1 \oplus \eta_2 = \oplus\big(\mathcal{F}'_{\gamma,\varepsilon}(\mu; h)\big) = \big(\oplus \circ \mathcal{F}'_{\gamma,\varepsilon}(\mu)\big)(h).$$

Taking into account that $\Omega_0(\mu) = \emptyset$, we derive from (5.4) and Definition 5.19 that

$$\mathcal{S}'_{\gamma,\varepsilon}(\mu; h) = \frac{1}{\gamma}\big(\mathcal{H}\big(\Omega_+(\mu)\big) \circ \oplus \circ \mathcal{F}'_{\gamma,\varepsilon}(\mu)\big)(h),$$

which is linear and bounded w.r.t. h. Repeating the same reasoning as in the end of the proof of Proposition 5.11, this shows that

$$\begin{aligned}\mathcal{S}'_{\gamma,\varepsilon}(\mu) &= \frac{1}{\gamma}\mathcal{H}\big(\Omega_+(\mu)\big) \circ \oplus \circ \mathcal{F}'_{\gamma,\varepsilon}(\mu) \\ &= \mathcal{H}\big(\Omega_+(\mu)\big) \circ \oplus \circ \big(\mathcal{N}\big(\Omega_+(\mu)\big) + \gamma\varepsilon E\big)^{-1}.\end{aligned}$$

Ad *6.*: "$\subset$": Let $G \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu)$ be given. By definition, there exists a sequence of differentiable points $(\mu_k)_{k\in\mathbb{N}}$ with $\mu_k \to \mu$ and $\mathcal{S}'_{\gamma,\varepsilon}(\mu_k) \to G$ as $k \to \infty$.

Analogous to the proof of Lemma 5.14, one can find both a subsequence $(k_l)_{l\in\mathbb{N}}$ and an index $L \in \mathbb{N}$ such that $\Omega_+(\mu_{k_l}) = \Omega_+(\mu_{k_L})$ for all $l \geq L$. Hence,

$$\begin{aligned}G = \lim_{k\to\infty} \mathcal{S}'_{\gamma,\varepsilon}(\mu_k) &= \lim_{l\to\infty} \mathcal{S}'_{\gamma,\varepsilon}(\mu_{k_l}) \\ &= \lim_{l\to\infty} \mathcal{H}\big(\Omega_+(\mu_{k_l})\big) \circ \oplus \circ \big(\mathcal{N}\big(\Omega_+(\mu_{k_l})\big) + \gamma\varepsilon\big)^{-1} \\ &= \mathcal{H}\big(\Omega_+(\mu_{k_L})\big) \circ \oplus \circ \big(\mathcal{N}\big(\Omega_+(\mu_{k_L})\big) + \gamma\varepsilon\big)^{-1}\end{aligned}$$

The rest of the proof can be taken almost word for word from the first part of the proof of Theorem 5.17.

"$\supset$": Conversely, if $\mathcal{A}$ has an outer structure w.r.t. Ω_0, then the second part of the proof of Theorem 5.17 reveals the existence of a sequence of differentiable points $(\mu_k)_{k\in\mathbb{N}}$ which satisfy $\Omega_+(\mu_k) = \Omega_+ \cup \mathcal{A}$ for all $k \in \mathbb{N}$ and $\mu_k \to \mu$ as $k \to \infty$. Thus,

$$\begin{aligned}\partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu) \ni \lim_{k\to\infty} \mathcal{S}'_{\gamma,\varepsilon}(\mu_k) &= \lim_{k\to\infty} \mathcal{H}\big(\Omega_+(\mu_k)\big) \circ \oplus \circ \big(\mathcal{N}\big(\Omega_+(\mu_k)\big) + \gamma\varepsilon\big)^{-1} \\ &= \mathcal{H}\big(\Omega_+ \cup \mathcal{A}\big) \circ \oplus \circ \big(\mathcal{N}\big(\Omega_+ \cup \mathcal{A}\big) + \gamma\varepsilon\big)^{-1}.\end{aligned}$$

□

Remark 5.21. As in the case of the solution operator of the regularized dual problem (see Remark 5.18), we note the following:

– We can always find at least two elements of $\partial_B \mathcal{S}_{\gamma,\varepsilon}$ by choosing the sets $\mathcal{A}_1 = \emptyset$ and $\mathcal{A}_2 = \Omega_0$.

– Theorem 5.20 does not yield a description of $\mathcal{S}_{\gamma,\varepsilon}$'s Bouligand subdifferential at points where it is differentiable. However, the derivative is always contained in the Bouligand subdifferential.

○

To summarize the findings of the current subchapter, Theorem 5.20 ensures that the marginal-to-transport-plan mapping is suitable for the IP approach in the sense of the beginning of Chapter 5. In particular,

– $\mathcal{S}_{\gamma,\varepsilon}$ is single-valued by construction;

– $\mathcal{S}_{\gamma,\varepsilon}$ satisfies a (fairly distinct) notion of differentiability: at almost every point it allows to compute a (total) derivative, and where it does not, we have at least a manageable characterization of its Bouligand subdifferential at hand.

Consequently, all that remains is to replace the lower-level problem of (BH) with the regularized marginal-to-transport-plan mapping we have just analyzed, i.e., apply the IP approach to the bilevel Hitchcock problem. This will be the topic of the next subchapter.

5.3 Application of the Implicit Programming Approach to the Regularized Bilevel Hitchcock Problem

In this last subchapter of Chapter 5, we finally want to apply the implicit programming approach, which we discussed at the beginning of Chapter 5, to the case of the bilevel Hitchcock problem.

To this end, let us recall the bilevel Hitchcock problem

$$\begin{aligned} \inf_{\pi,\mu_1} \quad & \mathcal{J}(\pi,\mu_1) \\ \text{s.t.} \quad & \mu_1 \in \mathbb{R}^{n_1},\ \mu_1 \geq 0,\ \mathbb{1}^\top \mu_1 = 1, \\ & \pi \in \arg\min\big\{(\theta, c_\mathrm{d})_F \colon \theta \in \mathbb{R}^{n_1 \times n_2},\ \theta \geq 0,\ \theta\mathbb{1} = \mu_1,\ \theta^\top \mathbb{1} = \mu_2^\mathrm{d}\big\}. \end{aligned} \tag{BH}$$

In the above, as usual, $\mathcal{J} \colon \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1} \to \mathbb{R}$ is a given lower-semicontinuous and bounded target function, $\mu_2^\mathrm{d} \in \mathbb{R}^{n_2}$, with $\mu_2^\mathrm{d} \geq 0$ and $\mathbb{1}^\top \mu_2^\mathrm{d} = 1$, is some fixed target marginal, and $c_\mathrm{d} \in \mathbb{R}^{n_1,n_2}$ is a cost matrix describing the cost of transportation, see Subchapter 4.1.

We then apply the IP approach to the bilevel problem (BH) by considering, for given regularization parameters $\gamma, \varepsilon > 0$, the *twice regularized bilevel Hitchcock problem*

$$\begin{aligned} \inf_{\pi,\mu_1} \quad & \mathcal{J}(\pi,\mu_1) \\ \text{s.t.} \quad & \mu_1 \in \mathbb{R}^{n_1},\ \mu_1 \geq 0,\ \mathbb{1}^\top \mu_1 = 1, \\ & \pi = \mathcal{S}_{\gamma,\varepsilon}(\mu_1, \mu_2^\mathrm{d}). \end{aligned} \tag{BH$_\gamma^\varepsilon$}$$

This problem is similar to (BH$_\gamma$), the quadratic regularization of (BH) from Chapter 4, but differs in two major aspects:

1. The operator replacing the lower-level Hitchcock problem is no longer the solution operator of (H_γ) but the regularized marginal-to-transport-plan mapping w.r.t. the cost matrix c_d from Subchapter 5.2.

2. We no longer consider the cost function as an optimization variable. We have already briefly commented on this in Subchapter 4.2, but we repeat the arguments once again and, on this occasion, add an additional one:

   – The only purpose for which we introduced the cost matrix as an optimization variable in the first place was that it allowed for the explicit construction of a recovery sequence in Subchapter 4.5. Now that we are, in this chapter, mainly interested in the implementation of the IP Approach, there is no use for it anymore.

   – Omitting the cost function greatly simplifies the notation not only in this section but also in Chapter 6.

   – Especially with respect to the implementation of the IP approach in Chapter 6, we benefit from neglecting the cost function, since this means that none of the optimization variables are subject to the curse of dimensionality anymore. Assuming a certain structure of the cost matrix c_d, e.g. $(c_\mathrm{d})_{i_1,i_2} = |i_1 - i_2|^\rho$ for some $\rho \geq 1$, the entries of the matrix $\frac{1}{\gamma}(\alpha_1 \oplus \alpha_2 - c)_+$ can be computed and stored efficiently, allowing the algorithm to potentially handle large problems without much effort.

If we replace the optimization variable π in ($\mathrm{BH}^\varepsilon_\gamma$) with the regularized marginal-to-transport-plan mapping $\mathcal{S}_{\gamma,\varepsilon}$, we obtain the *reduced bilevel problem*

$$\begin{aligned} \inf_{\mu_1} \quad & f_{\gamma,\varepsilon}(\mu_1) \coloneqq \mathcal{J}\big(\mathcal{S}_{\gamma,\varepsilon}(\mu_1, \mu_2^\mathrm{d}), \mu_1\big) \\ \text{s.t.} \quad & \mu_1 \in \mathbb{R}^{n_1},\ \mu_1 \geq 0,\ \mathbb{1}^\top \mu_1 = 1. \end{aligned} \tag{$\mathrm{RB}^\varepsilon_\gamma$}$$

The above problem is no longer a bilevel problem, but instead a nonsmooth optimization problem with linear constraints. Depending on the properties of $\mathcal{J}$, we can then attempt to solve this problem using nonsmooth optimization techniques such as subgradient descent, bundle methods, or gradient sampling methods, see e.g. [5], [4], [14] and the references therein.

In the next chapter, we propose a nonsmooth trust region method for the solution of ($\mathrm{RB}^\varepsilon_\gamma$) that is based on the Clarke subdifferential. To this end, we must first convince ourselves that, assuming that $\mathcal{J}$ is sufficient smooth, $f_{\gamma,\varepsilon}$ is indeed Lipschitz continuous and that we can compute Clarke subgradients at each point.

Proposition 5.22. *Let $\mathcal{J} \in C^1(\mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1})$ be continuously differentiable and denote its derivatives w.r.t. the first variable by $\nabla_\pi \mathcal{J}$ and its derivative w.r.t. the second variable by $\nabla_{\mu_1} \mathcal{J}$, i.e.,*

$$\mathcal{J}'(\pi, \mu_1) = \big(\nabla_\pi \mathcal{J}(\pi, \mu_1), \nabla_{\mu_1} \mathcal{J}(\pi, \mu_1)\big).$$

Then, $f_{\gamma,\varepsilon}$ is locally Lipschitz continuous and differentiable almost everywhere on $\mathbb{R}^{n_1}$. Moreover, for any $\mu_1 \in \mathbb{R}^{n_1}$ and any $\mathcal{A}$ that has an outer structure w.r.t. $\Omega_0(\mu_1, \mu_2^\mathrm{d})$, an element of the Clarke subdifferential of $f_{\gamma,\varepsilon}$ at μ_1 is given by

$$g_1 \coloneqq p_1 + \nabla_{\mu_1} \mathcal{J}\big(\mathcal{S}_{\gamma,\varepsilon}(\mu_1, \mu_2^\mathrm{d}), \mu_1\big) \in \partial f_{\gamma,\varepsilon}(\mu_1),$$

where $(p_1, p_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ *is the unique solution of the linear system*

$$\left(\mathcal{N}\left(\Omega_+(\mu_1, \mu_2^{\mathrm{d}}) \cup \mathcal{A}\right) + \gamma\varepsilon E\right) \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} M\mathbb{1} \\ M^\top \mathbb{1} \end{pmatrix}$$

with

$$M \coloneqq \mathcal{H}\left(\Omega_+(\mu_1, \mu_2^{\mathrm{d}}) \cup \mathcal{A}\right) \nabla_\pi \mathcal{J}\left(\mathcal{S}_{\gamma,\varepsilon}(\mu_1, \mu_2^{\mathrm{d}}), \mu_1\right).$$

Proof. According to [22, Corollary 2.2.1], $\mathcal{J}$ is strictly differentiable as well as locally Lipschitz continuous at every point and its strict derivative and total derivative coincide. This, together with $\mathcal{S}_{\gamma,\varepsilon}$'s Lipschitz continuity from Subchapter 5.2 and Rademacher's theorem yields the first claim.

To prove the second claim, we consider the operator

$$G_{\gamma,\varepsilon} \colon \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1}, \quad (\mu_1, \mu_2) \mapsto \left(\mathcal{S}_{\gamma,\varepsilon}(\mu_1, \mu_2), \mu_1\right).$$

$\mathcal{J}$'s (Clarke) subdifferential satisfies

$$\partial \mathcal{J}(\pi, \mu_1) = \left\{\mathcal{J}'(\pi, \mu_1)\right\} = \left\{\left(\nabla_\pi \mathcal{J}(\pi, \mu_1), \nabla_{\mu_1} \mathcal{J}(\pi, \mu_1)\right)\right\},$$

see e.g. [22, Proposition 2.2.4], and $G_{\gamma,\varepsilon}$'s (Clarke) subdifferential is given by

$$\partial G_{\gamma,\varepsilon}(\mu_1, \mu_2) = \left(\partial \mathcal{S}_{\gamma,\varepsilon}(\mu_1, \mu_2), (E_{n_1 \times n_1}, \mathbb{0}_{n_1 \times n_2})\right),$$

where $E_{n_1 \times n_1}$ denotes the identity matrix of $\mathbb{R}^{n_1 \times n_1}$ and $\mathbb{0}_{n_1 \times n_2}$ denotes the zero matrix of $\mathbb{R}^{n_1 \times n_2}$. Therefore, for all $\mu = (\mu_1, \mu_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, the chain rule for subdifferentials, see e.g. [22, Theorem 2.6.6], implies the set-valued equation

$$\begin{aligned} &\partial\left(\mathcal{J} \circ G_{\gamma,\varepsilon}\right)(\mu) \\ &\quad = \partial \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right) \circ \partial G_{\gamma,\varepsilon}(\mu) \\ &\quad = \left(\nabla_\pi \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right), \nabla_{\mu_1} \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right)\right) \circ \left(\partial \mathcal{S}_{\gamma,\varepsilon}(\mu), (E_{n_1 \times n_1}, \mathbb{0}_{n_1 \times n_2})\right) \\ &\quad = \nabla_\pi \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right) \partial \mathcal{S}_{\gamma,\varepsilon}(\mu) + \left(\nabla_{\mu_1} \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right)^\top, \mathbb{0}^\top\right). \end{aligned} \tag{5.19}$$

Now, consider some arbitrary $G \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu) \subset \partial \mathcal{S}_{\gamma,\varepsilon}(\mu)$. If we interpret $\nabla_\pi \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right)$ to be a linear operator from $\mathbb{R}^{n_1 \times n_2}$ to $\mathbb{R}$ and G to be a linear operator from $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ to $\mathbb{R}^{n_1 \times n_2}$, their composition is a linear operator from $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ to $\mathbb{R}$ and we can test it with some $(u, v) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ to obtain that

$$\begin{aligned} \nabla_\pi \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right) G(u, v) &= \left(\nabla_\pi \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right), G(u, v)\right)_F \\ &= \left(G^* \nabla_\pi \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right), (u, v)\right)_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}}, \end{aligned}$$

where G^* is the adjoint to G, which takes a linear operator from $\mathbb{R}^{n_1 \times n_2}$ to $\mathbb{R}$ (i.e., a matrix) and turns it into a linear operator on $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ (i.e., a pair of column vectors). Therefore, (5.19) implies that for each $G \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu)$ an element of $\partial(\mathcal{J} \circ G_{\gamma,\varepsilon})(\mu)$ is given

$$g \coloneqq G^* \nabla_\pi \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right) + \left(\nabla_{\mu_1} \mathcal{J}\left(G_{\gamma,\varepsilon}(\mu)\right)^\top, \mathbb{0}^\top\right) \tag{5.20}$$

Let G from above be fixed and abbreviate $\Omega_+ := \Omega_+(\mu)$. Furthermore, let $\mathcal{A} \subset \Omega_0(\mu)$ be the set that realizes G, i.e.,

$$G = \mathcal{H}(\Omega_+ \cup \mathcal{A}) \circ \oplus \circ \big(\mathcal{N}(\Omega_+ \cup \mathcal{A}) + \gamma\varepsilon E\big)^{-1},$$

see Theorem 5.20. For the computation of the subgradient g we shall now find a representation of the adjoint of G.

If $A \in L(Y, Z)$ and $B \in L(X, Y)$ are arbitrary linear operators between the Banach spaces X, Y, and Z, then

$$\begin{aligned}\langle (A \circ B)^* z^*, x\rangle_{X^*,X} &= \langle z^*, (A \circ B)x\rangle_{Z^*,Z} \\ &= \langle A^* z^*, Bx\rangle_{Y^*,Y} = \langle (B^* \circ A^*)z^*, x\rangle_{X^*,X}\end{aligned}$$

for all $x \in X$ and $z^* \in Z^*$ and therefore $(A \circ B)^* = B^* \circ A^*$. Applying this to G, yields that

$$G^* = \Big(\big(\mathcal{N}\big(\Omega_+(\mu) \cup \mathcal{A}\big) + \gamma\varepsilon E\big)^{-1}\Big)^* \circ \oplus^* \circ \mathcal{H}\big(\Omega_+(\mu) \cup \mathcal{A}\big)^*.$$

Taking a look at the definition of $\mathcal{N}$ in (5.10), we observe that the matrix $\mathcal{N}(\Omega_+(\mu) \cup \mathcal{A}) + \gamma\varepsilon E$ is symmetric. Consequently, its inverse is symmetric and we obtain that

$$\begin{aligned}&\Big(\big(\mathcal{N}\big(\Omega_+(\mu) \cup \mathcal{A}\big) + \gamma\varepsilon E\big)^{-1}\Big)^* \\ &\quad = \Big(\big(\mathcal{N}\big(\Omega_+(\mu) \cup \mathcal{A}\big) + \gamma\varepsilon E\big)^{-1}\Big)^\top = \big(\mathcal{N}\big(\Omega_+(\mu) \cup \mathcal{A}\big) + \gamma\varepsilon E\big)^{-1}.\end{aligned}$$

The adjoint of the $\oplus$-operator, which we already identified in Remark 4.8, is given by $\oplus^*(\theta) = (\theta\mathbb{1}, \theta^\top) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ for all $\theta \in \mathbb{R}^{n_1 \times n_2}$. Moreover, it is easy to see that the mask $\mathcal{H}(\Omega_+ \cup \mathcal{A})$ from Definition 5.19 is self-adjoint (as a operator on matrices): for any two matrices $\theta, \zeta \in \mathbb{R}^{n_1 \times n_2}$, it holds that

$$\big(\mathcal{H}(\Omega_+ \cup \mathcal{A})\theta, \zeta\big)_F = \sum_{(i_1,i_2) \in \Omega_+ \cup \mathcal{A}} \theta_{i_1,i_2}\zeta_{i_1,i_2} = \big(\theta, \mathcal{H}(\Omega_+ \cup \mathcal{A})\zeta\big)_F.$$

Altogether, this implies that we can compute the subgradient from (5.20) by

$$g = \big(\mathcal{N}\big(\Omega_+(\mu) \cup \mathcal{A}\big) + \gamma\varepsilon E\big)^{-1}(M\mathbb{1}, M^\top\mathbb{1}) + \big(\nabla_{\mu_1}\mathcal{J}\big(G_{\gamma,\varepsilon}(\mu)\big)^\top, \mathbb{0}^\top\big),$$

where

$$M := \mathcal{H}\big(\Omega_+(\mu) \cup \mathcal{A}\big)\,\nabla_\pi \mathcal{J}\big(G_{\gamma,\varepsilon}(\mu)\big) \in \mathbb{R}^{n_1 \times n_2}$$

and the inverse of $\mathcal{N}\big(\Omega_+(\mu) \cup \mathcal{A}\big) + \gamma\varepsilon E$ is again understood to be a linear automorphism on $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, see Remark 5.12.

The assertion then follows from the restriction of $\mathcal{J} \circ G_{\gamma,\varepsilon}$ to the set $\mathbb{R}^{n_1} \times \{\mu_2^{\mathrm{d}}\}$ so that g_1 corresponds to the first component of the subgradient $g \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. □

In the next and last chapter of this thesis, we will first set up a non-smooth trust region method for solving the reduced bilevel problem and then apply it to the transportation identification problem, which is a special case of (BH).

Chapter 6

Implementation of the Implicit Programming Approach for the Regularized Bilevel Problem

In the last chapter of this thesis, we propose an algorithm for the solution of the reduced bilevel problems ($\mathrm{RB}^{\varepsilon}_{\gamma}$). Having found a description of a subset of the Clarke subdifferential of the reduced target function $f_{\gamma,\varepsilon}$ in the previous chapter, we can use this to apply a nonsmooth optimization algorithm.

6.1 A Nonsmooth Trust Region Method for the Solution of Constrained Problems

In [21], the authors propose a nonsmooth trust region method for the solution of general *nonsmooth optimization problems*,

$$\inf_{x\in\mathbb{R}^n} \; f(x), \tag{NP}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$, for $n \in \mathbb{N}$, is a nonsmooth but locally Lipschitz continuous target function. Being locally Lipschitz continuous, f bears (Clarke) subgradients at every point $x \in \mathbb{R}^n$, see e.g. [22, Proposition 2.1.2], which can be used in conjunction with an approximation of f's Hessian to set up an ordinary trust region method with quadratic subproblems.

However, to avoid convergence to nonsmooth nonstationary points, the trust region method relies on the construction of a *model function* $\phi\colon \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}^n$. In some sense, this model function is designed to provide first-order information in a neighborhood of the current iterate and also to ensure stationarity of cluster points of the algorithm, see [21, Assumption 2.4]. If the ordinary trust region method converges to a nonsmooth point and if the trust region radius degenerates, the authors use the model function ϕ to define a modified trust region

subproblem that includes the neighborhood information of the model function to show either that the current iterate is stationary or to escape the sphere of influence of the current iterate.

Although there are various strategies to construct such a function ϕ, and the construction also may depend on the differentiability properties of the objective function f, constructions of ϕ at some given point typically involve (subsets of) the Clarke subdifferential at that point or an approximation of the Clarke subdifferential in a neighborhood of that point, see e.g. [21, Section 3] and the references therein.

Details on the implementation and other important aspects that would be too technical at this point can be found in the referenced paper. We only provide the proposed method for reference:

Algorithm 6.1 (Nonsmooth Trust Region Algorithm)**.**

1: *Initialization:* Choose constants

$$\Delta_{\min} > 0, \quad 0 < \eta_1 < \eta_2 < 1, \quad 0 < \beta_1 < 1 < \beta_2, \quad 0 < \nu \leq 1,$$

an initial value $x_0 \in \mathbb{R}^n$, and an initial trust region radius $\Delta_0 > \Delta_{\min}$. Set $k \leftarrow 0$.

2: **for** $k = 0, 1, 2, \ldots$ **do**

3: hoose a subgradient $g_k \in \partial f(x_k)$ and a symmetric matrix $H_k \in \mathbb{R}^{n\times n}_{\text{sym}}$.

4: **if** $\|g_k\|_{\mathbb{R}^n} = 0$ **then**

5: *Stop:* x_k is (Clarke) stationary, i.e., $\mathbb{0} \in \partial f(x_k)$.

6: **else**

7: **if** $\Delta_k \geq \Delta_{\min}$ **then**

8: Compute an (inexact) solution d_k of the *trust region subproblem*

$$\begin{aligned} \inf_d \quad & q_k(d) \coloneqq f(x_k) + (g_k, d)_{\mathbb{R}^n} + \tfrac{1}{2}\, d^\top H_k d \\ \text{s.t.} \quad & \|d\|_{\mathbb{R}^n} \leq \Delta_k, \end{aligned} \tag{6.1}$$

that satisfies the *generalized Cauchy decrease condition*

$$f(x_k) - q_k(d_k) \geq \frac{\nu}{2}\|g_k\|_{\mathbb{R}^n} \min\Big\{\Delta_k, \frac{\|g_k\|_{\mathbb{R}^n}}{\|H_k\|_{\mathbb{R}^{n\times n}}}\Big\}. \tag{6.2}$$

9: Compute the *quality indicator*

$$\rho_k \coloneqq \frac{f(x_k) - f(x_k + d_k)}{f(x_k) - q_k(d_k)}.$$

10: **else**

11: Compute an (inexact) solution d_k of the *modified trust region subproblem*

$$\begin{aligned} \inf_d \quad & \tilde{q}_k(d) \coloneqq f(x_k) + \phi(x_k, \Delta_k; d) + \tfrac{1}{2}\, d^\top H_k d \\ \text{s.t.} \quad & \|d\|_{\mathbb{R}^n} \leq \Delta_k, \end{aligned} \tag{6.3}$$

that satisfies the *modified generalized Cauchy decrease condition*

$$f(x_k) - \tilde{q}_k(d_k) \geq \frac{\nu}{2}\psi(x_k, \Delta_k) \min\Big\{\Delta_k, \frac{\psi(x_k, \Delta_k)}{\|H_k\|_{\mathbb{R}^{n\times n}}}\Big\}, \tag{6.4}$$

where

$$\psi(x_k, \Delta_k) \coloneqq -\min_{\|d\|_{\mathbb{R}^n} \leq 1} \phi(x, \Delta; d) \geq 0.$$

12: Compute the *modified quality indicator*

$$\rho_k \coloneqq \begin{cases} \dfrac{f(x_k) - f(x_k + d_k)}{f(x_k) - \tilde{q}_k(d_k)}, & \text{if } \psi(x_k, \Delta_k) > \|g_k\|_{\mathbb{R}^n} \Delta_k, \\ 0, & \text{if } \psi(x_k, \Delta_k) \leq \|g_k\|_{\mathbb{R}^n} \Delta_k. \end{cases}$$

13: **end if**

14: *Update:* Set

$$x_{k+1} \coloneqq \begin{cases} x_k, & \text{if } \rho_k \leq \eta_1 \\ x_k + d_k, & \text{otherwise,} \end{cases}$$

and

$$\Delta_{k+1} \coloneqq \begin{cases} \beta_1 \Delta_k, & \text{if } \rho_k \leq \eta_1, \\ \max\{\Delta_{\min}, \Delta_k\}, & \text{if } \eta_1 < \rho_k \leq \eta_2, \\ \max\{\Delta_{\min}, \beta_2 \Delta_k\}, & \text{if } \rho_k > \eta_2. \end{cases}$$

Set $k \leftarrow k + 1$.

15: **end if**

16: **end for**

Essentially, there are two reasons that prevent us from applying Algorithm 6.1 unchanged to the reduced bilevel problem ($\mathrm{RB}_\gamma^\varepsilon$):

1. Unlike the reduced bilevel problem ($\mathrm{RB}_\gamma^\varepsilon$), the nonsmooth problem (NP) is unconstrained and Algorithm 6.1 does not consider (linear) constraints. We therefore need to make sure that the nonsmooth trust region method respects the given constraints. This will not only affect the formulations of the trust region subproblems in Step 8 and Step 11 of the algorithm but also on the Cauchy decrease conditions in (6.2) and (6.4).

   Another approach would be to turn the regularized bilevel problem into a nonsmooth unconstrained problem by adding a penalization term to the target functional, which ensures that in the limit the marginal μ_1 satisfies the linear constraints. This, however, would require the introduction of another regularization term and regularization parameter, so we rather choose to modify the algorithm as described above.

2. Being an unconstrained optimization problem, a local minimum of (NP) must satisfy $\mathbb{0} \in \partial f(x)$, see e.g. [22, Proposition 2.3.2]. Therefore, it makes sense to choose this as a stationarity criterion in Step 5 of the above algorithm.

   In the case of ($\mathrm{RB}_\gamma^\varepsilon$), however, this stationarity must not be satisfied. We therefore have to find and incorporate a notion of stationarity that respects the constraints of the problem in ($\mathrm{RB}_\gamma^\varepsilon$). This will not only affect the termination criterion in Step 4 of the algorithm but also the Cauchy decrease conditions in (6.2) and (6.4).

In the following, we will be intentionally sparing with details on the changes we make to the algorithm and will only explain the most necessary points. This is mainly because the convergence analysis with the changes we make is pretty much along the lines of the convergence analysis from the original paper and, moreover, providing detailed proofs is beyond the scope of this thesis and subject to future research.

We begin with the discussion of the second of above's points and consider a different notion of stationarity.

Definition 6.2. Consider the *constrained nonsmooth optimization problem*

$$\begin{aligned} \inf_{x} \quad & f(x) \\ \text{s.t.} \quad & x \in \mathcal{C}, \end{aligned} \tag{CNP}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$, for $n \in \mathbb{N}$, is a locally Lipschitz continuous target function and $\mathcal{C} \subset \mathbb{R}^n$ is a closed convex set.

We call a point $\bar{x} \in \mathbb{R}^n$ *first-order stationary* for the constrained optimization problem (CNP), if it satisfies the *(generalized) variational inequality*

$$f^\circ(\bar{x}; z - \bar{x}) \geq 0 \quad \text{for all } z \in \mathcal{C}.$$

In the above variational inequality, $f^\circ(u; v)$ denotes the generalized directional derivative of f at $u \in \mathbb{R}^n$ in the direction $v \in \mathbb{R}^n$, see e.g. [22, Section 2.1].

That the stationarity condition from the above definition is a natural choice for our purposes, is shown in the following lemma.

Lemma 6.3. *Let $x^* \in \mathcal{C}$ be a local minimum of (CNP). Then, x^* is first-order stationary in the sense of Definition 6.2.*

Proof. Let $z \in \mathcal{C}$ be arbitrary. Then, for $t \in (0, 1)$ sufficiently small, the point $x^* + t(z - x^*) \in \mathcal{C}$ is included in the neighborhood of local optimality of x^*. Consequently,

$$\begin{aligned} 0 &\leq \liminf_{t \searrow 0} \frac{f\big(x^* + t(z - x^*)\big) - f(x^*)}{t} \\ &\leq \limsup_{\substack{y \to x^* \\ t \searrow 0}} \frac{f\big(y + t(z - x^*)\big) - f(y)}{t} = f^\circ(x^*; z - x^*), \end{aligned}$$

where the last equality is just the definition of the generalized directional derivative. □

Remark 6.4. 1. That the above notion of stationarity is a reasonable generalization of the first-order stationarity considered in [21], can be seen as follows. Being locally Lipschitz continuous, f is differentiable almost everywhere on $\mathbb{R}^n$. If we assume that $\bar{x}$ is a point of strict differentiability, see [22, p. 30], then the gradient of f at $\bar{x}$ exists and

$$\nabla f(\bar{x})^\top (z - \bar{x}) = f^\circ(\bar{x}; z - \bar{x}) \geq 0 \quad \text{for all } z \in \mathcal{C},$$

which is just the well-known *variational inequality* of nonlinear optimization, see e.g. [34, p. 13]. Moreover, if $\mathbb{0} \in \partial f(\bar{x})$, then by the very definition of the Clarke subdifferential it holds that

$$f^\circ(\bar{x}; z - \bar{x}) \geq (\mathbb{0}, z - \bar{x})_{\mathbb{R}^n} = 0 \quad \text{for all } z \in \mathcal{C},$$

i.e., vanishing subgradients are sufficient for first-order stationarity in the sense of Definition 6.2.

2. The reduced bilevel problem ($\mathrm{RB}^{\varepsilon}_{\gamma}$) from Subchapter 5.3 is a problem of the form (CNP): its target function $f_{\gamma,\varepsilon}$ is locally Lipschitz continuous, see Proposition 5.22, and ($\mathrm{RB}^{\varepsilon}_{\gamma}$)'s feasible set

$$\mathcal{C}_{\mathrm{RB}} \coloneqq \left\{\xi \in \mathbb{R}^{n_1} : \xi \geq 0,\ \mathbb{1}^{\top}\xi = 1\right\}$$

is a closed and convex subset of $\mathbb{R}^{n_1}$. Note that $\mathcal{C}_{\mathrm{RB}}$ is just the standard simplex of $\mathbb{R}^{n_1}$.

○

Along with the new notion of stationarity, we also need to define measures with which we can measure the degree of stationarity at a given point.

Definition 6.5. In the setting of (CNP), let $\bar{x} \in \mathcal{C}$, $\bar{g} \in \partial f(\bar{x})$, and $R > 0$ be given. We define a *stationarity measure* by

$$\theta_R(\bar{x}, \bar{g}) \coloneqq -\min_{\bar{x}+d\in\mathcal{C},\, \|d\|\leq R} (\bar{g}, d) \geq 0. \tag{6.5}$$

Moreover, if we are given a model function $\phi\colon \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}^n$ in the sense of [21, Assumption 2.4] and some trust region radius $\bar{\Delta} > 0$, we define a *modified stationarity measure* by

$$\psi_R(\bar{x}, \bar{\Delta}) \coloneqq -\min_{\bar{x}+d\in\mathcal{C},\, \|d\|\leq R} \phi(\bar{x}, \bar{\Delta}; d) \geq 0. \tag{6.6}$$

Remark 6.6. Definition 6.5 gives rise to several remarks:

1. The stationarity measures from (6.5) and (6.6) are (obvious) generalizations of the stationarity measures that were used in [21] to the setting of (CNP). If we consider $R = 1$ and $\mathcal{C} = \mathbb{R}^n$, we obtain that

$$\theta_R(\bar{x}, \bar{g}) = -\left(\bar{g}, \frac{-\bar{g}}{\|\bar{g}\|}\right) = \|\bar{g}\| \quad \text{and} \quad \psi_R(\bar{x}, \bar{\Delta}) = -\min_{\|d\|\leq 1} \phi(\bar{x}, \bar{\Delta}; d),$$

which are just the stationarity measures used in [21].

2. The reason we included the radius R in the definition of the stationarity measures lies in the structure of the feasible set of ($\mathrm{RB}^{\varepsilon}_{\gamma}$). Consider some arbitrary point $\bar{\mu}_1$ in the standard simplex. If $d \in \mathbb{R}^{n_1}$ is given such that $\bar{\mu}_1 + d \in \mathcal{C}_{\mathrm{RB}}$, then

$$1 = \mathbb{1}^{\top}(\bar{\mu}_1 + d) = \mathbb{1}^{\top}\bar{\mu}_1 + \mathbb{1}^{\top}d \quad \text{and} \quad \bar{\mu}_1 + d \in [\mathbb{0}, \mathbb{1}]$$

or equivalently

$$\mathbb{1}^{\top}d = 0 \quad \text{and} \quad d \in [-\bar{\mu}_1, \mathbb{1} - \bar{\mu}_1] \subset [-\mathbb{1}, \mathbb{1}],$$

where we understand the inclusions $d \in [a, b]$ for $a, b \in \mathbb{R}^{n_1}$ elementwise, i.e., $d_i \in [a_i, b_i]$ for all $i = 1, \ldots, n_1$. This shows that the linear constraints of ($\mathrm{RB}^{\varepsilon}_{\gamma}$) already imply that $\|d\|_{\infty} \leq 1$ and therefore $\|d\|_{\mathbb{R}^{n_1}} \leq \sqrt{n_1}\|d\|_{\infty}$, with the constant on the right-hand side arising from the equality of norms on $\mathbb{R}^{n_1}$. Consequently, if we set $\bar{R} \coloneqq \sqrt{n_1}$, then the (nonlinear) norm

constraints in the definition of (6.5) and (6.6) are superfluous and we obtain that

$$\theta_{\bar{R}}(\bar{\mu}_1, \bar{g}) = -\min_{\bar{\mu}_1 + d \in \mathcal{C}_{\mathrm{RB}}} (\bar{g}, d) \quad \text{and} \quad \psi_{\bar{R}}(\bar{\mu}_1, \bar{\Delta}) = -\min_{\bar{\mu}_1 + d \in \mathcal{C}_{\mathrm{RB}}} \phi(\bar{\mu}_1, \bar{\Delta}; d).$$

Thus, the calculation of the stationarity measure (which is necessary after each successful iteration and for every modified iteration) reduces to solving a linear problem, which benefits both the numerical implementation and the performance of the trust region method.

However, we must be careful to take the radius R into account in the further implementation of the algorithm, see e.g. the modified Cauchy decrease conditions below.

3. Obviously,

$$\theta_R(\bar{x}, \bar{g}) = -\min_{\bar{x} + d \in \mathcal{C},\, \|d\| \leq R} (\bar{g}, d) \geq -(\bar{g}, \mathbb{0}) = 0$$

and the same estimate for the modified stationarity measure follows from the presupposed positive homogeneity of the model function ϕ, see [21, Assumption 2.4].

4. Assume that the stationarity measure from (6.5) vanishes, i.e., $\theta_R(\bar{x}, \bar{g}) = 0$. Then, by the definition of the Clarke subdifferential it holds that

$$0 = \min_{\bar{x} + d \in \mathcal{C},\, \|d\| \leq R} (\bar{g}, d) \leq \inf_{\bar{x} + d \in \mathcal{C},\, \|d\| \leq R} f^\circ(\bar{x}; d).$$

Now, for arbitrary $z \in \mathcal{C}$, we find that $z_t \coloneqq (1-t)\bar{x} + tz \in \mathcal{C}$ and that $\|z_t - \bar{x}\| = t\|z - \bar{x}\| \leq R$ for all $t \in (0, 1)$ small enough. Consequently,

$$0 \leq \frac{1}{t} f^\circ(\bar{x}; z_t - \bar{x}) = f^\circ(\bar{x}; z - \bar{x})$$

because of the positive homogeneity of the generalized directional derivative, see e.g. [22, Proposition 2.1.1]. This shows that $\theta_R(\bar{x}, \bar{g}) = 0$ indeed implies that $\bar{x}$ is first-order stationary in the sense of Definition 6.2, providing another rationale for the usefulness of the stationarity measure from (6.5).

However, we cannot as easily show the same property for the modified stationarity measure from (6.6) as it critically depends on the construction of ϕ. We therefore assume the corresponding property to be given, see [21, Assumption 2.4].

○

To ensure that, during the iteration, we do not violate the feasibility of the sequence of iterates $(x_k)_{k \in \mathbb{N}_0} \subset \mathcal{C}$, it must hold in each iteration $k \in \mathbb{N}_0$ that $x_{k+1} \coloneqq x_k + d_k \in \mathcal{C}$, which effectively imposes a constraint on the descent direction d_k.

We have already seen this constraint in the definition of the stationarity measures and we also include it in the quadratic subproblems, i.e., we consider the *constrained trust region subproblem*

$$\begin{aligned} \inf_d \quad & q_k(d) \\ \text{s.t.} \quad & x_k + d \in \mathcal{C},\ \|d\|_{\mathbb{R}^n} \leq \Delta_k, \end{aligned} \tag{Q_k}$$

with q_k being defined as in (6.1), and the *modified constrained trust region subproblem*

$$\begin{aligned} \inf_d \quad & \tilde{q}_k(d) \\ \text{s.t.} \quad & x_k + d \in \mathcal{C}, \ \|d\|_{\mathbb{R}^n} \leq \Delta_k, \end{aligned} \tag{$\tilde{\text{Q}}_k$}$$

with $\tilde{q}_k$ being defined as in (6.3). By construction any (inexact) solution d_k of (Q_k) or ($\tilde{\text{Q}}_k$) guarantees that the feasibility of x_{k+1} is retained.

Since we have, compared to (6.1) and (6.3), restricted the feasible set of the trust region subproblems in (Q_k) and ($\tilde{\text{Q}}_k$), this of course affects the descent directions d_k and $\tilde{d}_k$, respectively, and in turn the expected reduction of the target value at the next iterate. As a consequence, we might not be able to find an (inexact) solution of (Q_k) or ($\tilde{\text{Q}}_k$) that satisfies the Cauchy decrease conditions from (6.2) or (6.4), respectively.

For this reason we apply the following changes in the spirit of [25, Part III] to the generalized Cauchy decrease conditions used in Algorithm 6.1:

- In the case (Q_k), we consider the *constrained Cauchy decrease condition*

$$f(x_k) - q_k(d_k) \geq \frac{\nu}{2R}\theta_R(x_k, g_k) \min\Big\{R, \Delta_k, \frac{\theta_R(x_k, g_k)}{R\|H_k\|_{\mathbb{R}^{n\times n}}}\Big\}. \tag{6.7}$$

- In the case of ($\tilde{\text{Q}}_k$), we consider the *modified constrained Cauchy decrease condition*

$$f(x_k) - \tilde{q}_k(\tilde{d}_k) \geq \frac{\nu}{2R}\psi_R(x_k, \Delta_k) \min\Big\{R, \Delta_k, \frac{\psi_R(x_k, \Delta_k)}{R\|H_k\|_{\mathbb{R}^{n\times n}}}\Big\}. \tag{6.8}$$

In the context of the constrained Cauchy decrease conditions from (6.7) and (6.8), we can prove a result similar to that from [21, Lemma 2.8] which, like the latter, forms the basis for the convergence analysis of the trust region method:

Lemma 6.7. *Let $x_k \in \mathcal{C}$, $g_k \in \partial f(x_k)$, $H_k \in \mathbb{R}^{n\times n}$, and $\Delta_k > 0$ be given. Denote the global minimizers of (Q_k) and ($\tilde{\text{Q}}_k$) by d_k^* and $\tilde{d}_k^*$, respectively. Then, d_k^* and $\tilde{d}_k^*$ satisfy (6.7) and (6.8), respectively, for every $\nu \leq 1$.*

Proof. The following proof is an adaption of the proof of [64, Lemma 3.2] to our setting. We only present the proof for $\tilde{d}_k^*$ and (6.8), since the proof for d_k^* and (6.7) is completely analogous.

To begin with, we consider the minimization problem associated with the stationarity measure, i.e.,

$$\min_{x_k + d \in \mathcal{C},\, \|d\| \leq R} \phi(x_k, \Delta_k; d) = -\psi_R(x_k, \Delta_k).$$

Owing to the presupposed lower semicontinuity of ϕ w.r.t. d, see Assumption 6.9 below, and the compactness of the feasible set, this problem admits at least one global minimizer which we denote by $\tilde{d}_k$. The optimality of $\tilde{d}_k^*$ for ($\tilde{\text{Q}}_k$) implies that

$$f(x_k) - \tilde{q}_k(\tilde{d}_k^*) \geq f(x_k) - \tilde{q}_k(d) \geq -\phi(x_k, \Delta_k; d) - \frac{1}{2}\|H_k\|_{\mathbb{R}^{n\times n}}\|d\|_{\mathbb{R}^n}^2 \tag{6.9}$$

for all $d \in \mathbb{R}^n$ with $x_k + d \in \mathcal{C}$ and $\|d\|_{\mathbb{R}^n} \leq \Delta_k$. On the one hand, if

$$\psi_R(x_k, \Delta_k) = -\phi(x_k, \Delta_k; \tilde{d}_k) \geq \|\tilde{d}_k\|_{\mathbb{R}^n}^2 \|H_k\|_{\mathbb{R}^{n\times n}} \min\{1, \Delta_k R^{-1}\},$$

then we set $d := \min\{1, \Delta_k R^{-1}\}\tilde{d}_k \in (\mathcal{C} - x_k) \cap \overline{B(0;\Delta_k)}$, insert this into (6.9), and use the positive homogeneity of $d \mapsto \phi(x_k, \Delta_k; d)$ to obtain that

$$\begin{aligned} f(x_k) - \tilde{q}_k(\tilde{d}_k^*) &\\ &\geq -\min\{1, \Delta_k R^{-1}\}\phi(x_k, \Delta_k; \tilde{d}_k) - \frac{1}{2}\big(\min\{1, \Delta_k R^{-1}\}\big)^2 \|H_k\|_{\mathbb{R}^{n\times n}} \|\tilde{d}_k\|^2 \\ &\geq \frac{1}{2}\psi_R(x_k, \Delta_k)\min\{1, \Delta_k R^{-1}\} = \frac{1}{2R}\psi_R(x_k, \Delta_k)\min\{R, \Delta_k\}. \end{aligned}$$

On the other hand, if

$$\psi_R(x_k, \Delta_k) = -\phi(x_k, \Delta_k; \bar{d}_k^*) < \|\tilde{d}_k\|_{\mathbb{R}^n}^2 \|H_k\|_{\mathbb{R}^{n\times n}} \min\{1, \Delta_k R^{-1}\},$$

we insert $d := -\phi(x_k, \Delta_k; \bar{d}_k^*)\|\tilde{d}_k\|_{\mathbb{R}^n}^{-2}\|H_k\|_{\mathbb{R}^{n\times n}}^{-1}\tilde{d}_k \in (\mathcal{C} - x_k) \cap \overline{B(0;\Delta_k)}$ into (6.9), which yields that

$$\begin{aligned} f(x_k) - \tilde{q}_k(\tilde{d}_k^*) &\geq \frac{\phi(x_k, \Delta_k; \tilde{d}_k)^2}{\|\tilde{d}_k\|_{\mathbb{R}^n}^2 \|H_k\|_{\mathbb{R}^{n\times n}}} - \frac{1}{2}\|H_k\|_{\mathbb{R}^{n\times n}} \frac{\phi(x_k, \Delta_k; \tilde{d}_k)^2}{\|\tilde{d}_k\|_{\mathbb{R}^n}^4 \|H_k\|_{\mathbb{R}^{n\times n}}^2}\|\tilde{d}_k\|_{\mathbb{R}^n}^2 \\ &= \frac{1}{2}\frac{\psi_R(x_k, \Delta_k)^2}{\|\tilde{d}_k\|_{\mathbb{R}^n}^2 \|H_k\|_{\mathbb{R}^{n\times n}}} \geq \frac{1}{2}\frac{\psi_R(x_k, \Delta_k)^2}{R^2\|H_k\|_{\mathbb{R}^{n\times n}}}. \end{aligned}$$

To summarize,

$$f(x_k) - \tilde{q}_k(\tilde{d}_k^*) \geq \frac{1}{2R}\psi_R(x_k, \Delta_k)\min\Big\{R, \Delta_k, \frac{\psi_R(x_k, \Delta_k)}{R\|H_k\|}\Big\},$$

which is exactly the modified constrained Cauchy-decrease condition from (6.8) with $\nu = 1$. Because the right-hand side is nonnegative, the estimate holds for every other value of $\nu \leq 1$. □

Remark 6.8. Lemma 6.7 is meaningful for three reasons. First, it shows that the constrained Cauchy decrease conditions that we defined in (6.7) and (6.8) are compatible with the constrained trust region subproblems (Q_k) and ($\tilde{Q}_k$), respectively, in the sense that the former give reasonable estimates of the descent of the objective function that we can achieve with the directions we get from the latter.

Second, it lays the foundation for the convergence analysis of the trust region algorithm from Algorithm 6.10. As mentioned earlier, this convergence analysis is in large part parallel to the convergence analysis in [21] and is therefore omitted.

Third, the proof provides us with a descent direction that satisfies the modified constrained Cauchy decrease condition. Thus, instead of solving the modified trust region subproblem directly (which, depending on the choice of the model function ϕ, may not be possible at all, see subsection 6.2), we can resort to the vector d from the proof as an inexact solution. ○

One last change we need to make to Algorithm 6.1 in order to apply it to the constrained nonsmooth problem (CNP) concerns the calculation of the modified quality indicator ρ_k. In Step 12 of Algorithm 6.1, we observe that ρ_k is computed in dependence of the ratio of the subgradient's norm and the stationarity measure of the modified subproblem. As we have already seen in

Remark 6.6, θ_R is the obvious generalization of $\|g_k\|_{\mathbb{R}^n}$ to the case of (CNP). If this is taken into account when calculating the modified quality indicator, one obtains the definition

$$\rho_k := \begin{cases} \dfrac{f(x_k) - f(x_k + d_k)}{f(x_k) - \tilde{q}_k(d_k)}, & \text{if } \psi_R(x_k, \Delta_k) > \theta_R(x_k, g_k)\Delta_k, \\ 0, & \text{if } \psi_R(x_k, \Delta_k) \leq \theta_R(x_k, g_k)\Delta_k. \end{cases}$$

We conclude this subchapter by presenting the method that arises when introducing all of the previously mentioned modifications to the method from Algorithm 6.1. First, however, we need to specify the assumptions on the model function ϕ, which

- are the obvious generalizations of the assumptions on the model function made in [21, Assumption 2.4];
- form the basis for the convergence analysis (the latter of which we do not present in this thesis, as already announced).

Assumption 6.9. We assume that we are given a model function $\phi\colon \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}^n$ with the following properties

1. for every $(x, \Delta) \in \mathbb{R}^n \times \mathbb{R}_+$, the mapping $d \mapsto \phi(x, \Delta; d)$ is positively homogeneous and lower semicontinuous;
2. given $x \in \mathcal{C}$, $\Delta > 0$, and $R > 0$, the stationarity measure ψ_R has the following property: if there is a sequence $(x_k, \Delta_k)_{k\in\mathbb{N}} \subset \mathcal{C} \times \mathbb{R}_+$ such that
$$x_k \to x, \quad \Delta_k \to 0, \quad \text{and} \quad \psi_R(x_k, \Delta_k) \to 0,$$
then x is first-order stationary for the problem (CNP);
3. if there is a sequence $(x_k, \Delta_k)_{k\in\mathbb{N}} \subset \mathcal{C} \times \mathbb{R}_+$ such that
$$x_k \to x, \quad \Delta_k \to 0, \quad \text{and} \quad \lim_{k\to\infty} \psi_R(x_k, \Delta_k) > 0,$$
then
$$\limsup_{k\to\infty} \sup_{\substack{x+d\in\mathcal{C}, \\ d\in\overline{B}(0;\Delta_k)}} \frac{f(x_k + d) - f(x_k) - \phi(x_k, \Delta_k; d)}{\Delta_k} \leq 0.$$

Given the above properties of the model function, we consider the following nonsmooth trust region method for the solution of problems of the type (CNP):

Algorithm 6.10 (Constrained Nonsmooth Trust Region Algorithm)**.**

1: *Initialization:* Choose constants
$$R, \Delta_{\min} > 0, \quad 0 < \eta_1 < \eta_2 < 1, \quad 0 < \beta_1 < 1 < \beta_2, \quad 0 < \nu \leq 1,$$
an initial value $x_0 \in \mathbb{R}^n$, and an initial trust region radius $\Delta_0 > \Delta_{\min}$. Set $k \leftarrow 0$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Choose a subgradient $g_k \in \partial f(x_k)$ and a symmetric matrix $H_k \in \mathbb{R}^{n\times n}_{\text{sym}}$.
4: **if** $\theta_R(x_k, g_k) = 0$ **then**

Stop: x_k is first-order stationary in the sense of Definition 6.2.
else
  if $\Delta_k \geq \Delta_{\min}$ **then**
    Compute an (inexact) solution d_k of the constrained trust region subproblem

$$\begin{aligned} \inf_d \quad & q_k(d) = f(x_k) + (g_k, d)_{\mathbb{R}^n} + \tfrac{1}{2} d^\top H_k d \\ \text{s.t.} \quad & x_k + d \in \mathcal{C},\ \|d\|_{\mathbb{R}^n} \leq \Delta_k \end{aligned} \tag{Q_k}$$

    that satisfies the constrained Cauchy decrease condition

$$f(x_k) - q_k(d_k) \geq \frac{\nu}{2R} \theta_R(x_k, g_k) \min\Big\{R, \Delta_k, \frac{\theta_R(x_k, g_k)}{R\|H_k\|_{\mathbb{R}^{n\times n}}}\Big\}.$$

    Compute the quality indicator

$$\rho_k := \frac{f(x_k) - f(x_k + d_k)}{f(x_k) - q_k(d_k)}.$$

  else
    Compute an (inexact) solution $\tilde{d}_k$ of the modified constrained trust region subproblem

$$\begin{aligned} \inf_d \quad & \tilde{q}_k(d) = f(x_k) + \phi(x_k, \Delta_k; d) + \tfrac{1}{2}\, d^\top H_k d \\ \text{s.t.} \quad & x_k + d \in \mathcal{C},\ \|d\|_{\mathbb{R}^n} \leq \Delta_k, \end{aligned} \tag{$\tilde{\mathrm{Q}}_k$}$$

    that satisfies the modified constrained Cauchy decrease condition

$$f(x_k) - \tilde{q}_k(\tilde{d}_k) \geq \frac{\nu}{2R} \psi_R(x_k, \Delta_k) \min\Big\{R, \Delta_k, \frac{\psi_R(x_k, \Delta_k)}{R\|H_k\|_{\mathbb{R}^{n\times n}}}\Big\}.$$

    Compute the modified quality indicator

$$\rho_k := \begin{cases} \dfrac{f(x_k) - f(x_k + d_k)}{f(x_k) - \tilde{q}_k(d_k)}, & \text{if } \psi_R(x_k, \Delta_k) > \theta_R(x_k, g_k)\Delta_k, \\ 0, & \text{if } \psi_R(x_k, \Delta_k) \leq \theta_R(x_k, g_k)\Delta_k. \end{cases}$$

  end if
  Update: Set

$$x_{k+1} := \begin{cases} x_k, & \text{if } \rho_k \leq \eta_1 \\ x_k + d_k, & \text{if } \rho_k > \eta_1, \end{cases}$$

  and

$$\Delta_{k+1} := \begin{cases} \beta_1 \Delta_k, & \text{if } \rho_k \leq \eta_1, \\ \max\{\Delta_{\min}, \Delta_k\}, & \text{if } \eta_1 < \rho_k \leq \eta_2, \\ \max\{\Delta_{\min}, \beta_2 \Delta_k\}, & \text{if } \rho_k > \eta_2. \end{cases}$$

  Set $k \leftarrow k + 1$.
  end if
end for

Now that we have defined a method for solving constrained nonsmooth problems, we of course want to apply it (successfully) to the reduced bilevel problem. This will be the topic of the next subchapter and includes, in particular, the choice of a model function which serves as the basis for the modified trust region subproblems.

6.2 Construction of a Model Function for the Reduced Bilevel Problem

Before we can actually present the results of the application of the constrained nonsmooth trust region method from Algorithm 6.10 to the reduced bilevel problem ($\mathrm{RB}_\gamma^\varepsilon$), we must first contemplate the choice of a model function in this particular case.

If we look closely at the method described in Algorithm 6.10, we observe that the modified trust region subproblem, and hence the (in advance) chosen model function, come into play after the trust region radius degenerates because the method makes no (significant) progress. This can, among other reasons, be caused by

- a bad choice of the current subgradient: if the current iterate's subdifferential contains the zero element, then the current iterate is already first-order stationary, see Remark 6.4. However, if the subdifferential additionally contains other elements, it is in general not guaranteed that in Step 3 of Algorithm 6.10 the zero element is chosen to be the current iterate's subgradient which would then terminate the iteration in Step 5. This would then lead to the stationarity of the current iteration not being detected, leading to a degeneration of the trust region radius.
- insufficient neighborhood information: in [21, Lemma 3.4], the authors show that an unfavorable combination of parameters can cause the trust region method from Algorithm 6.1 to converge to a nonstationary and nonoptimal point, even in the case of a piecewise affine and convex objective function and having information about the entire Clarke subdifferential at each point. The authors attribute this behavior to a lack of information about the subgradients in a neighborhood of each current iterate.

A possible solution to both of the above (and potentially other) problems is to include information about the (Clarke) subdifferential of adjacent points of the current iterate in the calculation of both the descent direction and the stationarity measure.

For this very reason, in the case of the reduced bilevel problem ($\mathrm{RB}_\gamma^\varepsilon$), we define the collective Bouligand subdifferential that unifies all Bouligand subdifferentials of the regularized marginal-to-transport-plan mapping in a ball around a given point.

Definition 6.11. Given $\gamma, \varepsilon > 0$, some point $\mu \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, and some radius $\Delta > 0$, we define the *collective Bouligand subdifferential* of the regularized marginal-to-transport-plan mapping $\mathcal{S}_{\gamma,\varepsilon}$ by

$$\mathcal{G}(\mu, \Delta) := \bigcup_{\xi \in \overline{B(\mu;\Delta)}} \partial_B \mathcal{S}_{\gamma,\varepsilon}(\xi).$$

By construction, $\partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu) \subset \mathcal{G}(\mu, \Delta)$.

The collective Bouligand subdifferential can easily become huge if not uncountable and it is currently unclear whether it is possible to obtain a computable description of its elements. We can, however, show the following approximation result:

Lemma 6.12. *Let $\mu \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be arbitrary. If $(\mu_k, \Delta_k) \to (\mu, 0)$, then*

$$\sup_{G \in \mathcal{G}(\mu_k, \Delta_k)} \inf_{H \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu)} \|G - H\|_{L(\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}, \mathbb{R}^{n_1 \times n_2})} \;\to\; 0$$

as $k \to \infty$.

Proof. To begin with, we abbreviate $(\alpha_1, \alpha_2) := \mathcal{S}_{\gamma,\varepsilon}(\mu)$. For all $\epsilon > 0$, if we choose $k \in \mathbb{N}$ large enough, then

$$\|\xi - \mu\|_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} \leq \Delta_k + \|\mu_k - \mu\|_{\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} < \epsilon$$

for all $\xi \in \overline{B(\mu_k; \Delta_k)}$. The solution operator of the regularized dual problem, $\mathcal{F}_{\gamma,\varepsilon}$, is (Lipschitz) continuous, see Lemma 5.4. Thus, if $(i_1, i_2) \in \Omega_+(\mu)$ or $(i_1, i_2) \in \Omega_-(\mu)$, then choosing k large enough always yields that

$$(\tilde{\alpha}_1 \oplus \tilde{\alpha}_2 - c)_{i_1,i_2} > 0 \quad \text{or} \quad (\tilde{\alpha}_1 \oplus \tilde{\alpha}_2 - c)_{i_1,i_2} < 0,$$

respectively, for all $(\tilde{\alpha}_1, \tilde{\alpha}_2) = \mathcal{S}_{\gamma,\varepsilon}(\xi)$ with $\xi \in \overline{B(\mu_k; \Delta_k)}$. Consequently, we can find some $K \in \mathbb{N}$ such that, for all $k \geq K$,

$$\Omega_+(\mu) \subset \Omega_+(\xi), \quad \Omega_-(\mu) \subset \Omega_-(\xi), \quad \text{and} \quad \Omega_0(\xi) \subset \Omega_0(\mu) \tag{6.10}$$

for all $\xi \in \overline{B(\mu_k; \Delta_k)}$. Note that the last inclusion in (6.10) directly follows from the former ones and the disjointness of the sets $\Omega_+(\xi)$, $\Omega_0(\xi)$, and $\Omega_-(\xi)$. Similarly,

$$\Omega_+(\xi) \setminus \Omega_+(\mu) \subset \Omega_0(\mu), \quad \Omega_-(\xi) \setminus \Omega_-(\mu) \subset \Omega_0(\mu) \tag{6.11}$$

for each such k and ξ.

Now, let $k \geq K$ be fixed and consider an arbitrary point $\xi \in \overline{B(\mu_k; \Delta_k)}$ with $(\tilde{\alpha}_1, \tilde{\alpha}_2) = \mathcal{S}_{\gamma,\varepsilon}(\xi)$ and an arbitrary subgradient $\tilde{G} \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\xi)$. If we manage to show that $\tilde{G} \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu)$, then the definition of $\mathcal{G}(\mu_k, \Delta_k)$ would yield that $\mathcal{G}(\mu_k, \Delta_k) \subset \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu)$ and thus

$$\sup_{G \in \mathcal{G}(\mu_k, \Delta_k)} \inf_{H \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu)} \|G - H\|_{L(\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}, \mathbb{R}^{n_1 \times n_2})} = 0$$

for all $k \geq K$, which would prove the claim.

In order to show that indeed $\tilde{G} \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu)$, we first note that, by Theorem 5.20, there exist $\tilde{\mathcal{A}} \subset \Omega_0(\xi)$ and $(\tilde{v}_1, \tilde{v}_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with

$$(\tilde{v}_1 \oplus \tilde{v}_2)_{\tilde{\mathcal{A}}} > 0 \quad \text{and} \quad (\tilde{v}_1 \oplus \tilde{v}_2)_{\Omega_0(\xi) \setminus \tilde{\mathcal{A}}} < 0$$

such that

$$\tilde{G} = \mathcal{H}\big(\Omega_+(\xi) \cup \tilde{\mathcal{A}}\big) \circ \oplus \circ \big(\mathcal{N}\big(\Omega_+(\xi) \cup \tilde{\mathcal{A}}\big) + \gamma\varepsilon E\big)^{-1}.$$

Because of (6.10) and (6.11), we may rewrite

$$\Omega_+(\xi) \cup \tilde{\mathcal{A}} = \Omega_+(\mu) \cup \mathcal{A},$$

with

$$\mathcal{A} := \big(\Omega_+(\xi) \setminus \Omega_+(\mu)\big) \cup \tilde{\mathcal{A}} \;\subset\; \Omega_0(\mu).$$

We now set

$$\lambda \coloneqq \frac{1}{2} \cdot \frac{\min_{(I_1,I_2)\in\Omega_+(\xi)\cup\Omega_-(\xi)} |(\tilde{\alpha}_1 \oplus \tilde{\alpha}_2 - c)_{I_1,I_2}|}{\max\{1, \max_{(I_1,I_2)\in\Omega_+(\xi)\cup\Omega_-(\xi)} |(\tilde{v}_1 \oplus \tilde{v}_2)_{I_1,I_2}|\}} \in (0,\infty)$$

to define $(v_1, v_2) \coloneqq (\tilde{\alpha}_1 - \alpha_1 + \lambda\tilde{v}_1, \tilde{\alpha}_2 - \alpha_2 + \lambda\tilde{v}_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. By construction,

$$v_1 \oplus v_2 = (\tilde{\alpha}_1 \oplus \tilde{\alpha}_2 - c) - (\alpha_1 \oplus \alpha_2 - c) + \lambda(\tilde{v}_1 \oplus \tilde{v}_2).$$

From this reformulation, we can deduce on the one hand, again applying (6.11), that

$$\begin{aligned}(v_1 \oplus v_2)_{\Omega_+(\xi)\setminus\Omega_+(\mu)} &= (\tilde{\alpha}_1 \oplus \tilde{\alpha}_2 - c)_{\Omega_+(\xi)\setminus\Omega_+(\mu)} + \lambda(\tilde{v}_1 \oplus \tilde{v}_2)_{\Omega_+(\xi)\setminus\Omega_+(\mu)} \\ &\geq (\tilde{\alpha}_1 \oplus \tilde{\alpha}_2 - c)_{\Omega_+(\xi)\setminus\Omega_+(\mu)} \\ &\quad - \frac{1}{2} \min_{(I_1,I_2)\in\Omega_+(\xi)\cup\Omega_-(\xi)} |(\tilde{\alpha}_1 \oplus \tilde{\alpha}_2 - c)_{I_1,I_2}| > 0\end{aligned}$$

and on the other hand (remember that $\tilde{\mathcal{A}} \subset \Omega_0(\xi) \subset \Omega_0(\mu)$) that

$$(v_1 \oplus v_2)_{\tilde{\mathcal{A}}} = \lambda(\bar{v}_1 \oplus \tilde{v}_2)_{\tilde{\mathcal{A}}} > 0.$$

Hence, to summarize, $(v_1 \oplus v_2)_{\mathcal{A}} > 0$. Analogously,

$$(v_1 \oplus v_2)_{\Omega_0(\mu)\cap\Omega_-(\xi)} < 0 \quad \text{and} \quad (v_1 \oplus v_2)_{\Omega_0(\xi)\setminus\tilde{\mathcal{A}}} < 0.$$

Because of $\Omega_+(\mu) \cap \Omega_0(\mu) = \emptyset$, one finds that

$$\begin{aligned}\Omega_0(\mu) \setminus \mathcal{A} &= \Omega_0(\mu) \setminus \big((\Omega_+(\xi) \setminus \Omega_+(\mu)) \cup \tilde{\mathcal{A}}\big) \\ &= \Omega_0(\mu) \setminus \big(\Omega_+(\xi) \cup \tilde{\mathcal{A}}\big) = \big(\Omega_0(\mu) \cap \Omega_-(\xi)\big) \cup \big(\Omega_0(\xi) \setminus \tilde{\mathcal{A}}\big)\end{aligned}$$

Hence,

$$(v_1 \oplus v_2)_{\Omega_0(\mu)\setminus\mathcal{A}} = (v_1 \oplus v_2)_{(\Omega_0(\mu)\cap\Omega_-(\xi))\cup(\Omega_0(\xi)\setminus\tilde{\mathcal{A}})} < 0$$

and consequently

$$\tilde{G} = \mathcal{H}\big(\Omega_+(\mu) \cup \mathcal{A}\big) \circ \oplus \circ \big(\mathcal{N}\big(\Omega_+(\mu) \cup \mathcal{A}\big) + \gamma\varepsilon E\big)^{-1} \in \partial_B \mathcal{S}_{\gamma,\varepsilon}(\mu),$$

see Theorem 5.20. Given our previous considerations, this concludes the proof. □

We have thus shown that the collective Bouligand subdifferential $\mathcal{G}(\mu, \Delta)$ satisfies the assumption on the approximation of the Bouligand subdifferential from [21, Assumption 4.1]. For this reason, it seems reasonable to adopt the construction of the model function in the cited paper, and we therefore define the model function we are going to utilize in the case of the reduced bilevel problem ($\mathrm{RB}^{\varepsilon}_{\gamma}$) by

$$\phi(\mu_1, \Delta; d) \coloneqq \sup_{G\in\mathcal{G}((\mu_1,\mu_2^{\mathrm{d}}),\,\Delta)} (p_G + \nabla_{\mu_1}\mathcal{J}(\pi, \mu_1), d)_{\mathbb{R}^{n_1}}, \tag{6.12}$$

where $\pi \coloneqq \mathcal{S}_{\gamma,\varepsilon}(\mu_1, \mu_2^{\mathrm{d}})$ and, for any $G \in \mathcal{G}(\mu, \Delta)$, the vector p_G (which we occasionally call *adjoint state*) is given by the first component of $G^* \nabla_\pi \mathcal{J}(\pi, \mu_1) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, see Proposition 5.22.

Attention. Whether our particular choice of the model function in (6.12) satisfies the properties from Assumption 6.9 is the subject of **ongoing research**. Therefore, from this point on, we cannot make any qualified predictions about the convergence of the sequence of iterates (or subsequences thereof) generated by the method from Algorithm 6.1.
Nevertheless, since the authors of [21] have demonstrated the convergence of their method for a very similar model function in a comparable scenario and since the numerical results we present in the upcoming subchapter tend to point in the right direction, we nevertheless use the model function from (6.12) to be finally able to numerically test the results of this thesis.

Finally, we have everything at hand to implement the constrained nonsmooth trust region method from Algorithm 6.10 and to apply it to the reduced bilevel problem ($\mathrm{RB}_\gamma^\varepsilon$). Details of the implementation and the discussion of the numerical results are the subject of the next and at the same time last subchapter.

6.3 A Transportation Identification Problem to Test the Constrained Nonsmooth Trust Region Method

In the last subchapter of this thesis, we present a certain instance of the bilevel Hitchcock problem (BH), which is intended to serve as a test problem for the constrained nonsmooth trust region method from Algorithm 6.10.

Suppose there is an (unknown) source marginal $\mu_1^* \in \mathbb{R}^{n_1}$ and a (known) target marginal $\mu_2^{\mathrm{d}} \in \mathbb{R}^{n_2}$, both of which satisfy $\mu_1^*, \mu_2^{\mathrm{d}} \geq 0$ as well as $\mathbb{1}^\top \mu_1^* = \mathbb{1}^\top \mu_2^{\mathrm{d}} = 1$, and that the cost of transportation between the domains Ω_1 and Ω_2 is given by a cost matrix $c_{\mathrm{d}} \in \mathbb{R}^{n_1 \times n_2}$. By Lemma 4.2, there exists (at least one) optimal transport plan π^* which describes the transportation between μ_1^* and μ_2^{d}. Let us assume that we can (possibly inaccurately) observe both the source marginal μ_1^* and the optimal transportation plan π^*, however, restricted to subdomains $D_1 \subset \Omega_1$ and $D \subset \Omega$, respectively. Denote the observations made on the subdomains by μ_1^{d} and π_{d}.

Then, the *tracking-type target function*, which is given by the function

$$\mathcal{J} \colon \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_1} \to \mathbb{R}_+, \quad \mathcal{J}(\pi, \mu_1) \coloneqq \frac{1}{2} \|\pi - \pi_{\mathrm{d}}\|_D^2 + \frac{\lambda}{2} \left\|\mu_1 - \mu_1^{\mathrm{d}}\right\|_{D_1}^2,$$

with weighting parameter $\lambda \geq 0$ and

$$\|M\|_D \coloneqq \sqrt{\sum_{(i_1,i_2) \in D} M_{i_1,i_2}^2} \quad \text{as well as} \quad \|v\|_{D_1} \coloneqq \sqrt{\sum_{i_1 \in D_1} v_{i_1}^2}$$

for all matrices $M \in \mathbb{R}^{n_1 \times n_2}$ as well as all vectors $v \in \mathbb{R}^{n_1}$, respectively, measures the distance between a point (π, μ_1) and the observed point $(\pi_{\mathrm{d}}, \mu_1^{\mathrm{d}})$. Obviously,

$\mathcal{J}$ is continuous w.r.t. π and μ_1 and thus satisfies the assumptions on the objective function of the bilevel Hitchcock problem (BH). Inserting this target function into the bilevel problem then yields the *transportation identification problem*

$$\begin{aligned} \inf_{\pi,\mu_1} \quad & \tfrac{1}{2}\|\pi - \pi_{\mathrm{d}}\|_D^2 + \tfrac{\lambda}{2}\big\|\mu_1 - \mu_1^{\mathrm{d}}\big\|_{D_1}^2 \\ \text{s.t.} \quad & \mu_1 \in \mathbb{R}^{n_1},\ \mu_1 \geq 0,\ \mathbb{1}^\top \mu_1 = 1, \\ & \pi \in \arg\min\big\{(\theta, c_{\mathrm{d}})_F \colon \theta \in \mathbb{R}^{n_1 \times n_2},\ \theta \geq 0,\ \theta\mathbb{1} = \mu_1,\ \theta^\top \mathbb{1} = \mu_2^{\mathrm{d}}\big\}, \end{aligned} \tag{TIP}$$

which seeks to find a source marginal μ_1 and an optimal transport plan π, transporting μ_1 onto μ_2^{d}, in a way that the distance between (μ_1, π) and the observed variables $(\pi_{\mathrm{d}}, \mu_1^{\mathrm{d}})$ is minimized. In other words, by solving (TIP), we try to reconstruct a transportation process where we know the target marginal and the cost function, but can only partially observe the source marginal and the transport plan.

The benefits of this type of problem are obvious: if we consider a weighting parameter $\lambda > 0$, the observation domains $D_1 = \Omega_1$ and $D = \Omega$, as well as the observations $\mu_1^{\mathrm{d}} = \mu_1^*$ and $\pi_{\mathrm{d}} = \pi^*$, then the unique solution of (TIP) is given by the point (π^*, μ_1^*) which realizes the target value $\mathcal{J}(\pi^*, \mu_1^*) = 0$. By choosing μ_1^* and π^* in advance, this allows us to evaluate the performance of the method from Algorithm 6.10 by means of a nontrivial bilevel problem whose solution is already known.

If we, however, choose proper subsets $D_1 \subsetneq \Omega_1$ and $D \subsetneq \Omega$ or add an error ϵ to the observation, i.e., if we consider $\mu_1^{\mathrm{d}} = \mu_1^* + \epsilon$ and $\pi_{\mathrm{d}} = \pi^* + \epsilon$, this allows us to introduce incomplete information or uncertainty to the transportation identification problem (TIP).

More advanced problems for testing both the constrained nonsmooth trust region method from Algorithm 6.10 and the results of Part II will be the subject of future research and publications.

A quick calculation shows that the tracking-type target function $\mathcal{J}$ is continuously differentiable as has the derivatives

$$\nabla_\pi \mathcal{J}(\pi, \mu_1) = \mathcal{H}(D)(\pi - \pi_{\mathrm{d}}) \quad \text{and} \quad \nabla_{\mu_1} \mathcal{J}(\pi, \mu_1) = \lambda\, \vec{\mathcal{H}}(D_1)(\mu_1 - \mu_1^{\mathrm{d}}).$$

In the above, $\mathcal{H}(D)$ is the mask from Definition 5.19 and $\vec{\mathcal{H}}(D_1)$ is an analogously defined operator operating on vectors instead of matrices, i.e.,

$$\vec{\mathcal{H}}(D_1) \colon \mathbb{R}^{n_1} \to \mathbb{R}^{n_1}, \quad v \mapsto \left(\begin{cases} v_{i_1}, & \text{if } i_1 \in D_1, \\ 0, & \text{else}, \end{cases} \right)_{i_1 \in \Omega_1}$$

for any subset $D_1 \subset \Omega_1$.

We follow the approach from Subchapter 5.3 and replace, given a pair of regularization parameters $\gamma, \varepsilon > 0$, the lower-level Hitchcock problem in the formulation of (TIP) by the regularized marginal-to-transport-plan mapping from Subchapter 5.2 to arrive at the problem

$$\begin{aligned} \inf_{\mu_1} \quad & f_{\gamma,\varepsilon}(\mu_1) \\ \text{s.t.} \quad & \mu_1 \in \mathcal{C}_{\mathrm{RB}} = \big\{\xi \in \mathbb{R}^{n_1} \colon \xi \geq 0,\ \mathbb{1}^\top \xi = 1\big\}, \end{aligned} \tag{TIP$_\gamma^\varepsilon$}$$

which is a problem of the form ($\mathrm{RB}_\gamma^\varepsilon$) with target function

$$f_{\gamma,\varepsilon}(\mu_1) = \frac{1}{2}\left\|\mathcal{S}_{\gamma,\varepsilon}(\mu_1, \mu_2^{\mathrm{d}}) - \pi_{\mathrm{d}}\right\|_D^2 + \frac{\lambda}{2}\left\|\mu_1 - \mu_1^{\mathrm{d}}\right\|_{D_1}^2$$

and thus treatable with the constrained nonsmooth trust region method which we described in Algorithm 6.10.

In the following, we will first briefly discuss important details of the implementation of the trust region method and then present the results of a number of different tests that we carried out in the setting of the transportation identification problem ($\mathrm{TIP}_\gamma^\varepsilon$).

6.3.1 Details on the Implementation of the Constrained Nonsmooth Trust Region Method

We implement the constrained nonsmooth trust region method from Algorithm 6.10 in MATLAB® R2023a[1]. In the following, we explain details of the implementation that are not immediately apparent from the description of the algorithm or that require further explanation.

Attention. Since it was, within the scope of this thesis, our primary goal to obtain first numerical results to validate the proposed trust region method and the results obtained in the previous parts, the implementation strategies presented here should be taken with a grain of salt.
It still subject of **ongoing research** how to optimally choose the subgradients in the modified and nonmodified case and how to reliably find solutions to the constrained trust region subproblems that not only satisfy the constrained Cauchy decrease conditions but also realize a substantial reduction of the target function and the stationarity measures.
Therefore, our implementation approaches should not be considered as sophisticated and reliable strategies, but merely as **heuristics**.

Step 3: *Choice of the subgradient and the symmetric matrix.*
Proposition 5.22 provides instructions on how to compute (Clarke) subgradients for a given iterate $\mu_{1,k}$. For a fixed pair of regularization parameters $\gamma, \varepsilon > 0$, we first apply a standard semismooth Newton method, see e.g. [52, Section 3.2], to compute the associated transport plan $\pi_k = \mathcal{S}_{\gamma,\varepsilon}(\mu_{1,k}, \mu_2^{\mathrm{d}})$ w.r.t. the cost matrix c_{d}.

In every iteration k, we choose the set $\mathcal{A}_k = \emptyset$, which has an outer structure w.r.t. the biactive set $\Omega_0(\mu_{1,k}, \mu_2^{\mathrm{d}})$, see Remark 5.16. With this choice, we construct the matrix

$$\begin{aligned} M_k &= \mathcal{H}\big(\Omega_+(\mu_{1,k}, \mu_2^{\mathrm{d}}) \cup \mathcal{A}_k\big)\nabla_\pi \mathcal{J}(\pi_k, \mu_{1,k}) \\ &= \mathcal{H}\big(\Omega_+(\mu_{1,k}, \mu_2^{\mathrm{d}})\big)\nabla_\pi \mathcal{J}(\pi_k, \mu_{1,k}) = \mathcal{H}\big(\Omega_+(\mu_{1,k}, \mu_2^{\mathrm{d}}) \cap D\big)(\pi_k - \pi_{\mathrm{d}}), \end{aligned}$$

then use the MATLAB® function "mldivide" to compute the unique solution

[1] MATLAB is a registered trademark of The MathWorks, Inc. See mathworks.com/trademarks for a list of additional trademarks.

$(p_{1,k}, p_{2,k}) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ to the linear system

$$\left(\mathcal{N}\big(\Omega_+(\mu_{1,k}, \mu_2^{\mathrm{d}})\big) + \gamma\varepsilon E\right) \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} \mathbb{1} M_k \\ \mathbb{1}^\top M_k \end{pmatrix}$$

and receive a subgradient for the current iteration by setting

$$g_k = p_{1,k} + \nabla_{\mu_1} \mathcal{J}(\pi_k, \mu_1) = p_{1,k} + \lambda \vec{\mathcal{H}}(D_1)(\mu_{1,k} - \mu_1^{\mathrm{d}}). \tag{6.13}$$

We feel the urge to note that there may be a more sophisticated way to select the set $\mathcal{A}_k$. However, this is subject to future research and will not be discussed in the scope of this thesis.

In each iteration k, we choose the matrix H_k to be an approximation of the Hessian of $f_{\gamma,\varepsilon}$, which we compute via a BFGS update using the subgradient g_k and the initial matrix $H_0 = E_{n_1 \times n_1}$, the identity matrix of $\mathbb{R}^{n_1 \times n_1}$. Even though we have no theoretical evidence that this approach increases the numerical performance, we reset the BFGS update to the identity matrix in every 10th successful iteration or when the norm of the matrix outgrows a certain threshold.

Step 4: Calculation of the stationarity measure.
We fix the parameter $R = \sqrt{n_1}$. This way, the calculation of the stationarity measure θ_R reduces to finding a solution $\bar{d}_k$ to the linear program

$$\begin{aligned} \min_d \quad & (g_k, d)_{\mathbb{R}^{n_1}} \\ \text{s.t.} \quad & d \in \mathbb{R}^{n_1},\ \mu_{1,k} + d \in \mathcal{C}_{\mathrm{RB}}, \end{aligned}$$

which we efficiently solve by employing the MATLAB® function "linprog".

Of course, after computing the stationarity measure, we do not test whether it is exactly zero, but whether it is close to zero, i.e., we choose a tolerance $0 < \mathrm{TOL} \ll 1$ and evaluate the expression $\theta_R(\mu_{1,k}, g_k) \leq \mathrm{TOL}$ in order to obtain a numerically meaningful termination criterion.

Step 8: Computation of an inexact solution to (Q_k).
The strategy we describe below for finding an inexact solution to the constrained trust region subproblem (Q_k) was inspired by the projected gradient methods that are discussed in [25, Chapter 12].

We consider both the linearized subproblem

$$\begin{aligned} \inf_d \quad & f(\mu_{1,k}) + (g_k, d)_{\mathbb{R}^{n_1}} \\ \text{s.t.} \quad & d \in \mathbb{R}^{n_1},\ \mu_{1,k} + d \in \mathcal{C}_{\mathrm{RB}}, \end{aligned}$$

whose solution is given by $\bar{d}_k$ from Step 4 above, and the "classical" trust region subproblem

$$\begin{aligned} \inf_d \quad & q_k(d) = f(\mu_{1,k}) + (g_k, d)_{\mathbb{R}^{n_1}} + \tfrac{1}{2} d^\top H_k d \\ \text{s.t.} \quad & d \in \mathbb{R}^{n_1},\ \|d\|_{\mathbb{R}^{n_1}} \leq \Delta_k, \end{aligned}$$

which does not include the linear constraints of (Q_k) and for which we compute an (inexact) solution $\hat{d}_k$ via the well-known dogleg method.

We then compute the convex combination between $\bar{d}_k$ and the metric projection[2] of $\hat{d}_k$ onto the shifted standard simplex $\mathcal{C}_{\mathrm{RB}} - \mu_{1,k}$ that minimizes q_k.

[2] To calculate the metric projection onto the linear constraints, we use the method described in [20] and the MATLAB® function "projsplx" provided by the authors of that paper.

By construction, this convex combination is feasible for (Q_k) and finding it is a quadratic problem, which can be solved with a simple case-by-case analysis.

Even though, at the current time, we have no theoretical evidence to support this claim, the descent direction we obtain from this approach seems to always satisfy the constrained Cauchy decrease condition (6.7) with $\nu = 1$.

Step 11: *Calculation of the modified stationarity measure.*
With the choice of the parameter R from above, the computation of the modified stationarity measure reduces to solving the problem

$$\begin{aligned} \min_{d} \quad & \phi(\mu_{1,k}, \Delta_k; d) \\ \text{s.t.} \quad & d \in \mathbb{R}^{n_1},\ \mu_{1,k} + d \in \mathcal{C}_{\mathrm{RB}}, \end{aligned}$$

which is, because of

$$\phi(\mu_{1,k}, \Delta_k; d) = \sup_{G \in \mathcal{G}((\mu_{1,k}, \mu_2^{\mathrm{d}}), \Delta_k)} \big(p_G + \lambda \vec{\mathcal{H}}(D_1)(\mu_{1,k} - \mu_1^{\mathrm{d}}), d\big)_{\mathbb{R}^{n_1}},$$

see its definition in (6.12), equivalent to the problem

$$\begin{aligned} \min_{\xi, d} \quad & \xi \\ \text{s.t.} \quad & \xi \in \mathbb{R},\ d \in \mathbb{R}^{n_1},\ \mu_{1,k} + d \in \mathcal{C}_{\mathrm{RB}}, \\ & \big(p_G + \lambda \vec{\mathcal{H}}(D_1)(\mu_{1,k} - \mu_1^{\mathrm{d}}), d\big)_{\mathbb{R}^{n_1}} \leq \xi \quad \text{for all } G \in \mathcal{G}\big((\mu_{1,k}, \mu_2^{\mathrm{d}}), \Delta_k\big), \end{aligned}$$

which is a problem with linear objective function and (possibly uncountably many) linear inequality constraints. Instead of solving the above problem exactly (which, depending on the structure of $\mathcal{G}$, might not even be possible), we compute a solution of the linear approximating problem

$$\begin{aligned} \min_{\xi, d} \quad & \xi \\ \text{s.t.} \quad & \xi \in \mathbb{R},\ d \in \mathbb{R}^{n_1},\ \mu_{1,k} + d \in \mathcal{C}_{\mathrm{RB}}, \\ & \big(p_G + \lambda \vec{\mathcal{H}}(D_1)(\mu_{1,k} - \mu_1^{\mathrm{d}}), d\big)_{\mathbb{R}^{n_1}} \leq \xi \quad \text{for all } G \in \hat{\mathcal{G}}, \end{aligned} \tag{6.14}$$

where $\hat{\mathcal{G}} \subset \mathcal{G}\big((\mu_{1,k}, \mu_2^{\mathrm{d}}), \Delta_k\big)$ is an approximation that contains up to $10(n_1+n_2)$ different subgradients of the collective Bouligand subdifferential.

To construct the approximation $\hat{\mathcal{G}}$, we explore the $n_1 + n_2$ sphere around the center point $(\mu_{1,k}, \mu_2^{\mathrm{d}})$ with radius Δ_k in every (positive and negative) unit direction of $\mathbb{R}^{n_1+n_2}$. For each of these $2(n_1 + n_2)$ points on the sphere around $(\mu_{1,k}, \mu_2^{\mathrm{d}})$, we calculate (if possible) multiple unique subgradients, see the instructions given in Proposition 5.22.

Although there are many more points on the sphere and inside the ball that could be generated in a similar way, we limit ourselves to the points generated as described above, since each point requires the calculation of the corresponding regularized transport plan, i.e., in particular the application of the semismooth Newton method from Step 3. The effort to estimate the modified stationarity measure with the approximation of the collective Bouligand subdifferential described above is already a multiple of that of the unmodified one and each additional point contributes to extending the runtime of the method.

Again, we efficiently solve the approximating linear problem from (6.14) by using the MATLAB® function "linprog" and we denote the solution by $\tilde{d}_k$.

Even though this is not obvious from the description of the method in Algorithm 6.10, just as in [21, Remark 2.13], we implement a modified termination criterion by evaluating $\psi_R(\mu_{1,k}, \Delta_k) \leq \text{TOL}$.

Step 11: Computation of an inexact solution to $(\tilde{\text{Q}}_k)$.
Similar to step 4, we do not attempt to solve the modified subproblem exactly (which may not be possible due to the structure of $\mathcal{G}$), but instead use the direction $\tilde{d}_k$ from the approximation of the stationarity measure ψ_R above and scale it to obtain an admissible descent direction. To do this, we stick to the proof of Lemma 6.7 and construct the inexact solution of $(\tilde{\text{Q}}_k)$ the same way as the direction d in the mentioned proof.

Despite the fact that latter is based on the exact solution of the modified stationarity measure and we only rely on an approximate solution, in our numerical tests, we virtually always obtain descent directions that satisfy the modified Cauchy descent condition with $\nu = 1$.

Step 14: Update of the variables.
If the trust region method from Algorithm 6.10 consecutively generates a large number of successful steps, the trust region radius can, according to the update rule in Step 14, grow dramatically. Therefore, if the iteration approaches a first-order stationary point, it usually requires a large number of null steps until the trust region radius adjusts to the neighborhood of local stationarity and until the solution of the constrained subproblem (Q_k) realizes a sufficient reduction of the objective function again.

To limit the number of (unnecessary) null steps, we therefore define an upper bound for the trust region radius by setting $\Delta_{\max} = \sqrt{n_1}$ and considering the modified update rule

$$\Delta_{k+1} := \begin{cases} \beta_1 \Delta_k, & \text{if } \rho_k \leq \eta_1, \\ \max\{\Delta_{\min}, \Delta_k\}, & \text{if } \eta_1 < \rho_k \leq \eta_2, \\ \min\big\{\max\{\Delta_{\min}, \beta_2 \Delta_k\}, \Delta_{\max}\big\}, & \text{if } \rho_k > \eta_2. \end{cases} \tag{6.15}$$

Since the structure of the linear constraints in (Q_k) and $(\tilde{\text{Q}}_k)$ implies an upper bound on the norm of the search direction anyway, this modified update rule also makes sense from a theoretical point of view.

6.3.2 Presentation & Discussion of the Numerical Results

We now present the results of two different numerical examples with which we test our implementation of the constrained nonsmooth trust region method from Algorithm 6.10.

For both of the numerical examples, we set $R = \Delta_{\max} = \sqrt{n_1}$ to simplify the calculation of the stationarity measures and to reduce the number of unsuccessful iterations, see Section 6.3.1. We choose the remaining parameters of the trust region method according to Table 6.1. For the initial values, we always choose

$$\mu_{1,0} = n_1^{-1} \mathbb{1}, \quad \Delta_0 = 1, \quad \text{and} \quad H_0 = E_{n_1 \times n_1},$$

the latter of which being the identity matrix of $\mathbb{R}^{n_1 \times n_1}$.

Transportation identification on the entire domain

$\Delta_{\min}$	η_1	η_2	β_1	β_2	ν
10^{-6}	0.1	0.9	0.5	1.5	1

Table 6.1: Choice of the parameters for our test of the constrained nonsmooth trust region method from Algorithm 6.10.

For our first example, we set $n_1 = n_2 = 20$ and consider the cost matrix that is given by $(c_{\mathrm{d}})_{i_1,i_2} = |i_1 - i_2|^2$ for all $(i_1, i_2) \in \Omega$. Using the MATLAB® function "sprand", we (pseudo-)randomly generate a source marginal $\mu_1^* \in \mathbb{R}^{20}$ and a target marginal $\mu_2^{\mathrm{d}} \in \mathbb{R}^{20}$, both of which are to a large extent sparse (roughly 75% of their entries are equal to 0). Applying the MATLAB® function "linprog", we then calculate the (unique) optimal transport plan π^* between the marginals μ_1^* and μ_2^{d} w.r.t. the cost c_{d}.

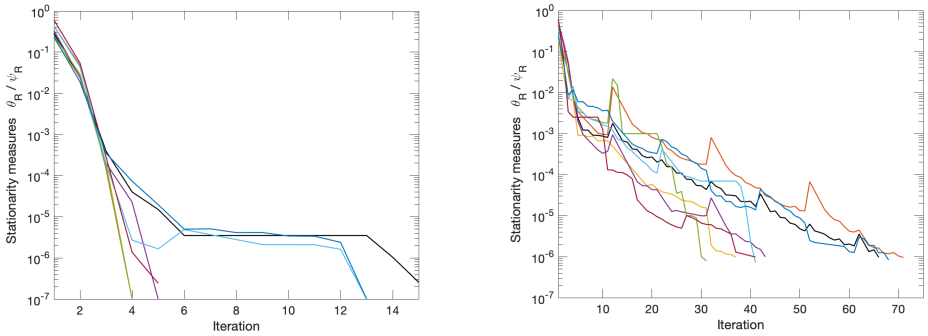
We moreover set the weight $\lambda = 1$ and consider the observation domains $D_1 = \Omega_1$ and $D = \Omega$ as well as the observations $\mu_1^{\mathrm{d}} = \mu_1^*$ and $\pi_{\mathrm{d}} = \pi^*$. This setup corresponds to the attempt of reconstructing the source marginal μ_1^* and the optimal transport plan π^* from exact observations on the entire domain. Although this optimization problem is a (from an analytical point of view) trivial exercise, it is well suited as a first test for the trust region method. We generate eight independent instances of this problem, i.e., eight different tuples $(\mu_1^*, \mu_2^{\mathrm{d}}, \pi^*)$, and combine them with different choices of regularization parameters $\gamma = \varepsilon$, ranging between 10^0 and 10^{-9}.

To begin with, we would like to point out that graphs that have the same color across the pictures below correspond to the same instance. Therefore, if we, for example, refer to the "red instance", we actually refer to the instance that corresponds to the red graphs in the pictures.

Figure 6.1 shows the history of the stationarity measure for these eight instances for different choices of regularization parameters $\gamma = \varepsilon$. We observe that, for the majority of instances, after about 30 to 70 iterations (including successful steps as well as null steps), stationarity is achieved within a tolerance of less than 10^{-6}. It is noteworthy that the method failed to achieve this tolerance in four cases (black instance for $\gamma = \varepsilon = 10^{-3}$; red and dark blue instance for $\gamma = \varepsilon = 10^{-4}$; purple instance for $\gamma = \varepsilon = 10^{-7}$). Plateaus with constant values of the stationarity measures correspond to periods of unsuccessful iterations.

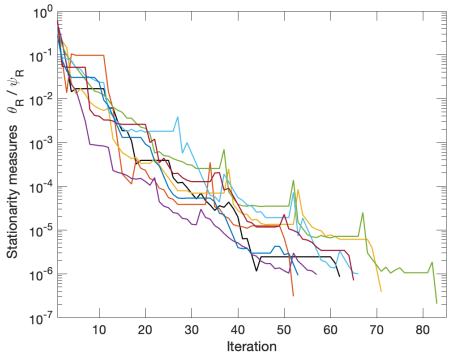
Figure 6.2 shows the history of target function values, i.e., the sum of the squared residuals of the method's output $\bar{\mu}_1$ and the corresponding transport plan $\bar{\pi} = \mathcal{S}_{\gamma,\varepsilon}(\bar{\mu}_1, \mu_2^{\mathrm{d}})$. The target function is bounded from below by 0 and its optimal value of 0 can only be realized by the unique point μ_1^*. With a conservative choice of regularization parameters, see Figure 6.2a–c, the reduction of the objective function and thereby the quality of the approximation of μ_1^* and π^* is rather poor. However, for smaller choices of regularization parameters, see Figure 6.2d–f, we observe that (after a few globalization steps) the method achieves a significant reduction of the stationarity measure in a few steps before taking a large number of iterations to drive the stationarity measure towards zero. For $\gamma = \varepsilon = 10^{-6}$, Figure 6.2f shows that each instance is solved after at most 30 iterations with a (squared) residual of less than 10^{-6}. For several instances, this accuracy is even below 10^{-7} if not close to 10^{-8}.

A comparison of the final iteration number of the eight instances across all

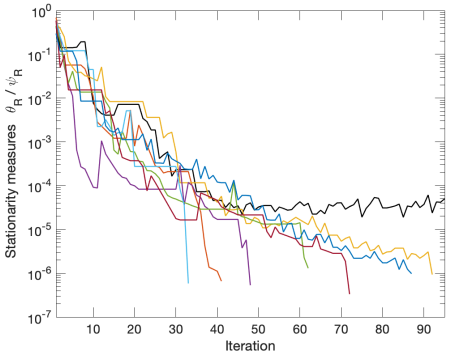


(a) $\gamma = \varepsilon = 10^0$

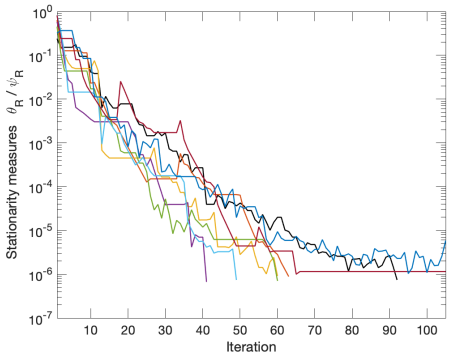
(b) $\gamma = \varepsilon = 10^{-1}$



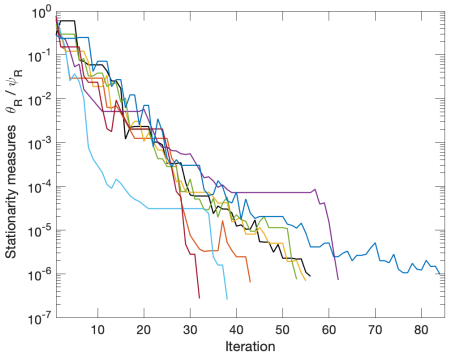
(c) $\gamma = \varepsilon = 10^{-2}$



(d) $\gamma = \varepsilon = 10^{-3}$

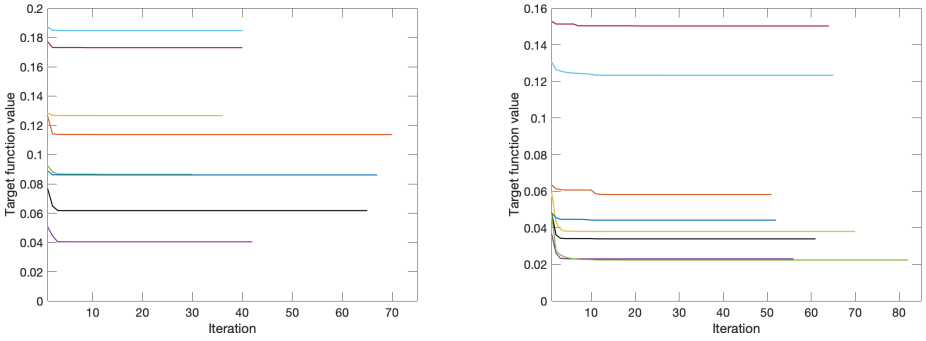


(e) $\gamma = \varepsilon = 10^{-4}$



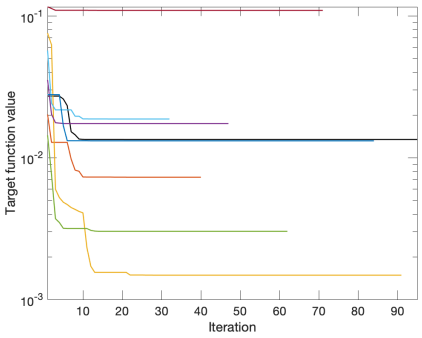
(f) $\gamma = \varepsilon = 10^{-5}$

Figure 6.1: Stationarity plots of the eight instances of the transport identification problem with exact observations on the entire domain, given different values of regularization parameters. Graphs that touch the "Iteration"-axis, see picture (a), correspond to a final stationarity measure of exactly 0.

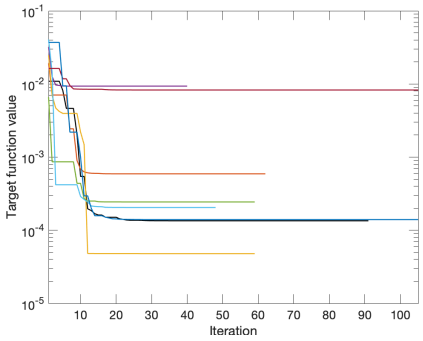


(a) $\gamma = \varepsilon = 10^{-1}$

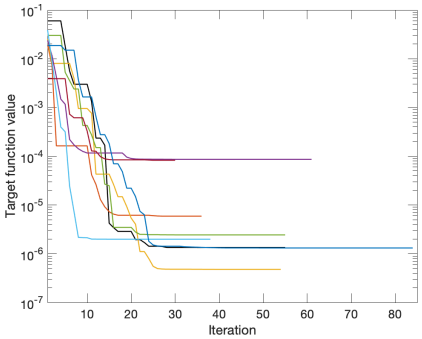
(b) $\gamma = \varepsilon = 10^{-2}$



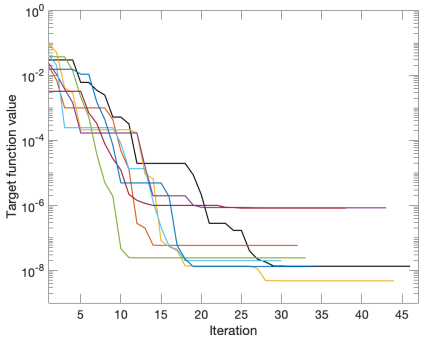
(c) $\gamma = \varepsilon = 10^{-3}$



(d) $\gamma = \varepsilon = 10^{-4}$



(e) $\gamma = \varepsilon = 10^{-5}$



(f) $\gamma = \varepsilon = 10^{-6}$

Figure 6.2: Target function plots of the eight instances from Figure 6.1 for different values of the regularization parameters.

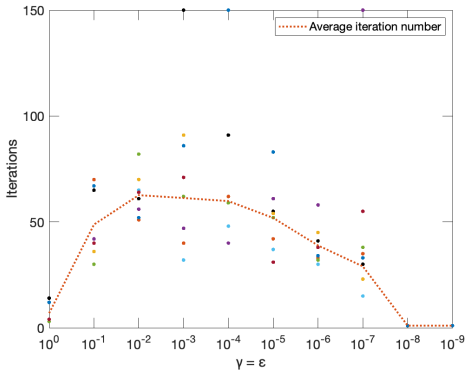


Figure 6.3: Comparison of the final iteration number of the eight instances for different values of the regularization parameters. The termination tolerance was chosen to be 10^{-6}. A value of 150 indicates that the method failed to converge in that particular test. The dotted line corresponds to the average iteration number with the non-convergent cases excluded.

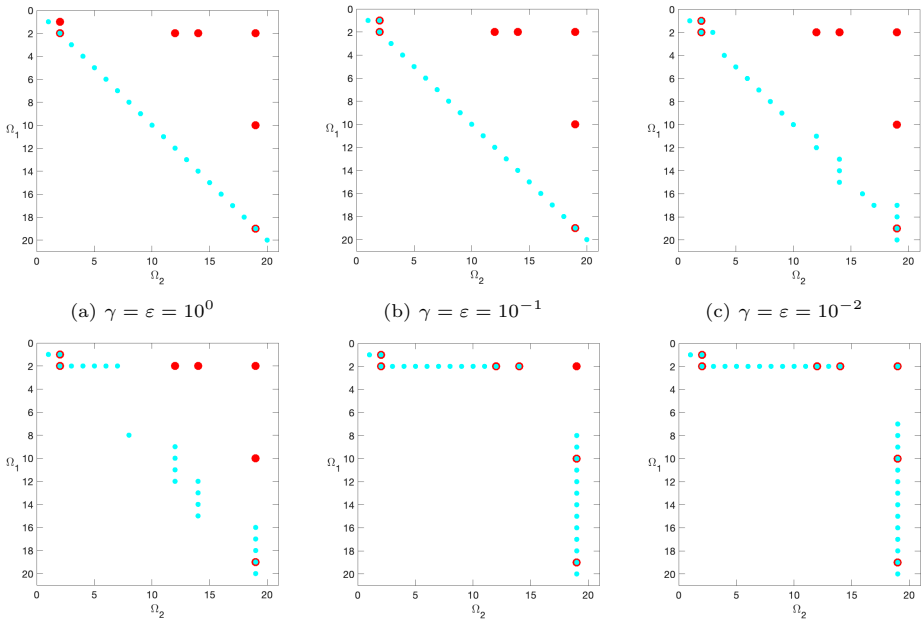
choices of the regularization parameters $\gamma = \varepsilon$ is shown in Figure 6.3. We observe that the average iteration number (dotted line) initially rises when decreasing the regularization parameters $\gamma = \varepsilon$ from 10^0 to 10^{-2}, but subsequently falls when decreasing the regularization parameters further beyond 10^{-2}. For values smaller than 10^{-7}, the method terminates in the very first iteration, as the matrices used during the semismooth Newton method become singular and the method returns infeasible solutions. It can be seen that the number of iteration steps is subject to a large variance and that some instances (e.g. the dark blue and the purple one) tend to require more iterations than other instances (e.g. the light blue one) across the majority of the tests.

Figure 6.4 shows, for a single instance, the evolution of the sparsity pattern of the transport plan $\bar{\pi}$ associated to the output $\bar{\mu}_1$ of the method. The pictures show that the sparsity pattern of $\bar{\pi}$, which is induced by the $(\,\cdot\,)_+$-operator in the definition of the regularized marginal-to-transport-plan mapping, see Subchapter 5.2, in a sense "converges" to the sparsity pattern of the optimal transport plan π^*. This approximation of the sparsity pattern is, however, only an outer approximation, meaning that the sparsity pattern of the approximation $\bar{\pi}$ always contains points that do not belong to the sparsity pattern of π^*.

To conclude this numerical example, we would like to briefly discuss the usage of the modified trust region subproblem and the corresponding stationarity measure. Table 6.2 shows the cases (in parenthesis) in which the modified stationarity measure or the modified subproblem was used. Excluding the cases, where the method failed to converge, we count a total of 29 of modified iterations. Among all of the executed tests, we essentially observed four distinctive behaviors. The following list is sorted according to the frequency of occurrence of those cases.

1. The modified subproblem is not used during the iteration: this case accounts for the majority of tests carried out.

2. The modified subproblem and the modified stationarity measure were only used a few times during the iteration and each time led to a significant reduction of the stationarity measure: we observed this behavior, for ex-



(a) $\gamma = \varepsilon = 10^{0}$ (b) $\gamma = \varepsilon = 10^{-1}$ (c) $\gamma = \varepsilon = 10^{-2}$

(d) $\gamma = \varepsilon = 10^{-3}$ (e) $\gamma = \varepsilon = 10^{-4}$ (f) $\gamma = \varepsilon = 10^{-5}$

Figure 6.4: Comparison of the sparsity pattern of the optimal transport plan π^* (red points) with the sparsity pattern of the transport plan $\bar{\pi} = \mathcal{S}_{\gamma,\varepsilon}(\bar{\mu}_1, \mu_2^{\mathrm{d}})$ (light blue points) corresponding to the method's output $\bar{\mu}_1$ of the trust region method for different values of regularization parameters. The data presented in this figure corresponds to the purple instance. For regularization parameters smaller than 10^{-5}, the sparsity pattern remains unchanged.

instance \ $\gamma=\varepsilon$	10^{0}	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
black	14	65	61(1)	∞	91(2)	55	41	30
orange	3	70	51	40	62	42	33(1)	35(1)
yellow	3	36	70	91	59(1)	54	45	23
purple	4	42	56	47	40	61	58	$\infty(\infty)$
green	3	30	82	62	59	52	32	38(1)
light blue	12	40	65	32	48	37	30	15
red	4	40	64	71	$\infty(\infty)$	31	38(1)	55(14)
dark blue	12	67	52	86(4)	∞	83(3)	34	33

Table 6.2: Summary of the final iteration numbers as presented in Figure 6.3. Values in parenthesis show the number of iterations in which either the modified stationarity measure was computed (and the iteration stopped subsequently) or the modified subproblem was solved. A value of ∞ indicates that the method failed to converge in this particular case.

ample, in the test of the black instance for $\gamma = \varepsilon = 10^{-2}$ or $\gamma = \varepsilon = 10^{-4}$.

3. Once the trust region radius falls below the threshold value $\Delta_{\min}$, the modified subproblem exclusively generates null steps and the trust region radius degenerates: this case occurred twice in the executed tests (red instance for $\gamma = \varepsilon = 10^{-4}$ and purple instance for $\gamma = \varepsilon = 10^{-7}$, the former of which can be inspected in Figure 6.1e).

The behavior described in the first two cases is very favorable. Each nonmodified iteration is inexpensive and requires, besides a number of matrix-vector multiplications and distinctions of cases, only the solution of one nonlinear and two linear systems of equations as well as the solution of a linear program. In contrast, each iteration of the modified subproblem requires a multiple of the effort of a nonmodified iteration, because

- for every point $\xi \in \overline{B(\mu_{1,k}; \Delta_k)}$ that we choose from the ball around the current iterate, we need to compute the corresponding transport plan $\pi_\xi = \mathcal{S}_{\gamma,\varepsilon}(\xi, \mu_2^{\mathrm{d}})$, which requires the application of the semismooth Newton method; depending on the size of the biactive set, i.e., the number of elements in $\Omega_0(\xi, \mu_2^{\mathrm{d}})$, this occurs up to $2(n_1 + n_2)$ times per modified iterate;

- the computation of each subgradient requires to solve a (sparse but high-dimensional) linear system.

In particular, the behavior of the second case is exactly what we had hoped to achieve with the choice of the model function ϕ in Subsection 6.2: if possible, the modified model should not be used, but if it cannot be avoided due to the problem structure, then it should only contribute a few iterations (being, in some sense, a "safeguard").

The behavior described in the last case is obviously not ideal. However, it is at this stage not clear whether it is provoked by an inadequate approximation of the collective Bouligand subdifferential in Step 11, by a too inaccurate solution of the modified subproblem ($\tilde{\mathrm{Q}}_k$), or by an unfavorable choice of the model function ϕ. It could also be attributed to an impractical choice of the termination tolerance, because the graphs in Figure 6.2 show that there is usually no significant reduction of the target function after the 30th iteration, even if the iteration is continued. It is the subject of future research, how this particular behavior of the method can be avoided.

Transportation identification on a part of the domain

For our second example, we set $n_1 = n_2 = 50$ and consider the same cost matrix as before, i.e., $(c_{\mathrm{d}})_{i_1,i_2} = |i_1 - i_2|^2$ for all $(i_1, i_2) \in \Omega$. Just as in the previous test, we generate a test instance by pseudorandomly drawing a source marginal μ_1^* as well as a target marginal μ_2^{d} and by calculating the corresponding (unique) optimal transport plan π^* afterwards. Using this instance, we now investigate the effect of different choices of observation domains on the result of the trust region method. In all of the subsequent tests, we choose the regularization parameters $\gamma = \varepsilon = 10^{-6}$ in order to achieve a close approximation of the optimal solution. We set the stationarity tolerance to 10^{-4} and, initially, fix the weighting parameter $\lambda = 1$.

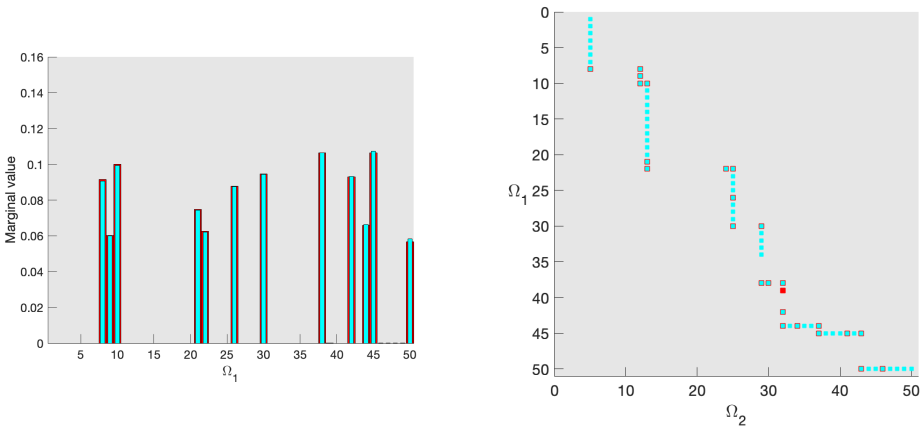


Figure 6.5: Left-hand picture: Comparison of the source marginal μ_1^* (red bars) with the trust region method's output $\bar{\mu}_1$ (blue bars). Right-hand picture: Comparison of the sparsity pattern of the optimal transport plan π^* (red squares) with the sparsity pattern of the calculated transport plan $\bar{\pi} = \mathcal{S}_{\gamma,\varepsilon}(\bar{\mu}_1, \mu_2^{\mathrm{d}})$ (blue squares).
Note that while the value of the 38th entry of μ_1^* is in the order of 10^{-5} and therefore barely visible in the left-hand picture, one can see the corresponding nonzero entry in the right-hand picture.

As a reference, we first solve the transportation identification on the entire domain, i.e., we consider the observation domains $D_1 = \Omega_1$ and $D = \Omega$. The trust region method terminated after 58 iterations (of which 33 were successful) with an objective value (sum of squared residuals) of about $5 \cdot 10^{-8}$. A comparison of the method's output $\bar{\mu}_1$ with the source marginal μ_1^* as well as a comparison of the sparsity patterns of $\bar{\pi} = \mathcal{S}_{\gamma,\varepsilon}(\bar{\mu}_1, \mu_2^{\mathrm{d}})$ and π^* can be found in Figure 6.5. The left-hand picture shows that the source marginal μ_1^* was indeed approximated very well with the element-wise deviation of $\bar{\mu}_1$ from μ_1^* being less than $2 \cdot 10^{-3}$. The right-hand picture shows that the sparsity pattern of π^* was recovered to a large extend.

To test the trust region method on proper subsets of the domains, we consider observation domains that result from gradually "punching" certain points out of the domains Ω_1 and Ω. To be more precise, we define the observation domain $D_1^1 \subset \Omega_1$ to be the equal to Ω_1, but with every fifth index removed, i.e.,

$$D_1^1 \coloneqq \Omega_1 \setminus \{5, 10, \ldots, 50\}.$$

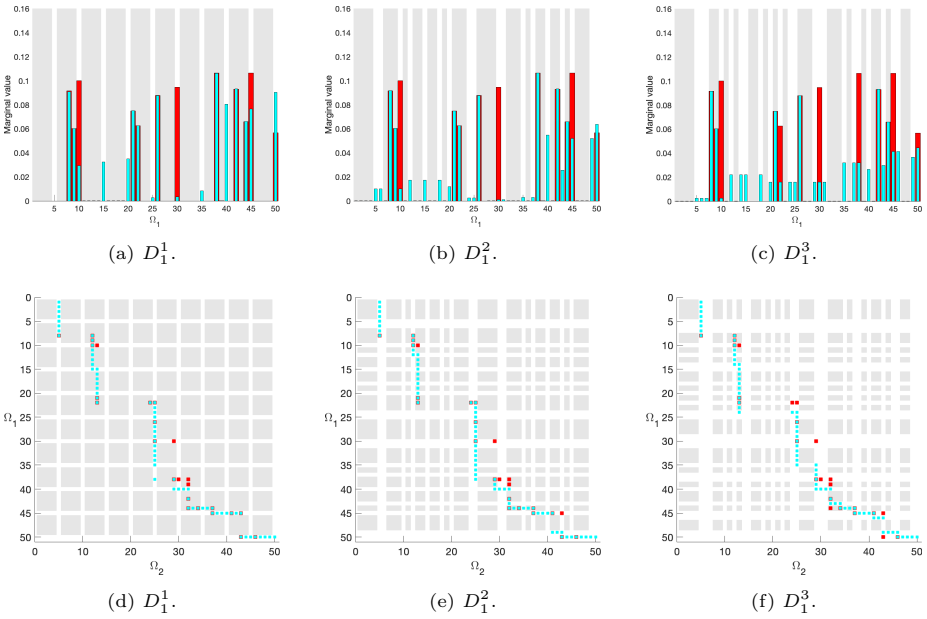
We construct the observation domain $D^1 \subset \Omega$ similarly, but instead of removing single indices, we remove every fifth row and every fifth column, i.e.,

$$D^1 \coloneqq \Omega \setminus \big\{\{5\} \times \Omega_2, \Omega_1 \times \{5\}, \{10\} \times \Omega_2, \Omega_1 \times \{10\}, \ldots\big\}.$$

We then construct D_1^2 by removing every fifth index of D_1^1 and construct D^2 by removing every fifth row and every fifth column from D^1. Repeating this construction over and over again, we obtain sequences of observation domains

$$\Omega_1 \supset D_1^1 \supset D_1^2 \supset \ldots \quad \text{and} \quad \Omega \supset D^1 \supset D^2 \supset \ldots.$$

With each repetition, the observation domains become more sparse while a certain number of points is retained. For each repetition $k = 1, 2, \ldots$, we define the observations to be $\mu_1^{\mathrm{d}} \coloneqq \mu_1^*|_{D_1^k}$ and $\pi^{\mathrm{d}} \coloneqq \pi^*|_{D^k}$.



(a) D_1^1. (b) D_1^2. (c) D_1^3.

(d) D_1^1. (e) D_1^2. (f) D_1^3.

Figure 6.6: (a) – (c): Comparison of μ_1^* (red bars) with $\bar{\mu}_1$ (blue bars) for different instances of the observation domains D_1^k (gray bars). (d) – (f): Comparison of the sparsity pattern of π^* (red squares) with the sparsity pattern of $\bar{\pi}$ (blue squares) for different instances of the observation domain D^k (gray squares).

In Figure 6.6, we see both the shape of the resulting observation domains for the first three repetitions as well as the results of the trust region method when applying it to the given data. We observe, and this is exactly what we would have expected, that the method's output $\bar{\mu}_1$ approximates the source marginal μ_1^* still very well, but only for the indices that lie in the observation domain. On the complement of the observation domain, the difference between μ_1^* and $\bar{\mu}_1$ can be quite large, see e.g. $i_1 = 30$ in Figure 6.6a or $i_1 = 22$ in Figure 6.6b and Figure 6.6c. Although most of the nonzero entries of $\bar{\pi}$ lie in the complement of D^1, D^2 and D^3, see figures 6.6d – 6.6f, yet the transport plan $\bar{\pi}$ cannot completely avoid the observation domains, since some points of μ_1^* lie in all shown observation domains and $\bar{\mu}_1$ closely approximates μ_1^* at these points (remember that the mapping of the marginals onto the regularized transport plan is continuous).

Table 6.3 shows some relevant data that we collected after the trust region method terminated. As expected, the realized target function value falls as the observation domains become more sparse whereas the residual, i.e., $\frac{1}{2}\|\bar{\pi} - \pi^*\|_\Omega + \frac{\lambda}{2}\|\bar{\mu}_1 - \mu_1^*\|_{\Omega_1}$, rises due to the loss of information about the source marginal μ_1^* and the optimal transport plan π^*. At the transition from D_1^4 & D^4 to D_1^5 & D^5, the target function value increases again. We attribute this to the subgradients becoming increasingly sparse and the stationarity tolerance therefore being reached earlier, in this case after only 6 iterations.

To conclude this experiment, we choose the observation domain D_1 to be the last third of Ω_1 and the observation domain D to be a diagonal strip shifted to the higher end of Ω_1. Figure 6.7 shows the effect of different choices of the weighting parameter λ on the reconstruction of μ_1^* and π^* and Table 6.4 shows

Observ. dom.	Stat. measure	Target val.	Iter. count	Resid.
D_1^1 & D^1	$7.4710 \cdot 10^{-5}$	$5.2670 \cdot 10^{-4}$	65	$2.0537 \cdot 10^{-2}$
D_1^2 & D^2	$6.1495 \cdot 10^{-5}$	$5.2617 \cdot 10^{-4}$	45	$2.3373 \cdot 10^{-2}$
D_1^3 & D^3	$7.0698 \cdot 10^{-5}$	$5.9076 \cdot 10^{-6}$	16	$2.9190 \cdot 10^{-2}$
D_1^4 & D^4	$5.6165 \cdot 10^{-5}$	$5.9499 \cdot 10^{-8}$	24	$3.8291 \cdot 10^{-2}$
D_1^5 & D^5	$1.5872 \cdot 10^{-5}$	$4.3354 \cdot 10^{-3}$	6	$6.8572 \cdot 10^{-2}$

Table 6.3: End of iteration data for the first pairs of observation domains. Residual refers to the sum of the squared residual of $\bar{\mu}_1$ to μ_1^* and $\bar{\pi}$ to π^*, i.e., $\mathcal{J}(\bar{\pi}, \bar{\mu}_1)$ for $D = \Omega$ and $D_1 = \Omega_1$.

the corresponding output data.

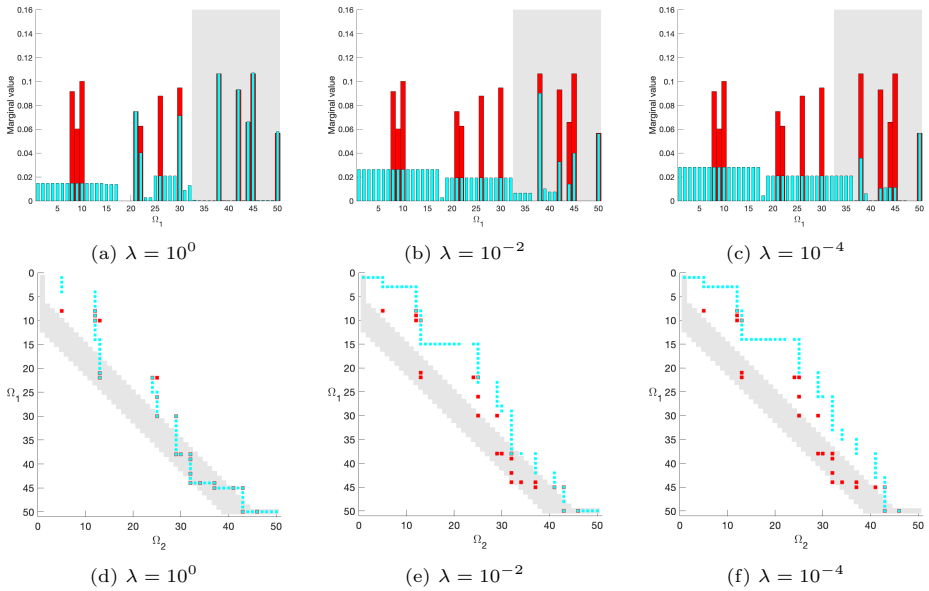
While for $\lambda = 1$ (see Figure 6.7a) the reconstruction of μ_1^* (on D_1) is very accurate, this changes when the influence of the weighting parameter on the objective function is reduced, see Figure 6.7b & Figure 6.7c. In the complement of the observation domain D_1, we observe little to no reconstruction of the source marginal.

As the weighting parameter that controls the influence of μ_1 on the target function decreases, the influence of π on the target function increases. We would therefore expect the accuracy with which μ_1^* is approximated to decrease and the accuracy of the approximation of π^* to increase. However, we make the following, rather counterintuitive observation: For $\lambda = 1$, the approximation of both μ_1^* and π^* appears to be very good, but as the weighting parameter λ decreases, the quality of both approximations decreases.

A possible explanation for this behavior would be that, for $\lambda = 1$, the proximity of $\bar{\mu}_1$ to μ_1^* (see Figure 6.7d) together with the (Lipschitz) continuity of the marginal-to-transport-plan mapping $\mathcal{S}_{\gamma,\varepsilon}$ force the corresponding transport plan $\bar{\pi}$ to be close to π^*, leading to good approximations and small target values. Reducing the weighting parameter λ, however, significantly reduces the quality of the approximation of μ_1^* on the observation domain, thus (again by continuity of the marginal-to-transport-plan mapping) leading to a poor approximation of π^*.

Further, reducing the weighting parameter λ significantly reduces the number of iterations required to achieve a certain stationarity tolerance, see Table 6.4. This is, however, not a surprise given the construction of the subgradients from (6.13): in the calculation of g_k, the derivative of $\mathcal{J}$ w.r.t. μ_1 is weighted with lambda and this weighting is directly transferred to the calculation of the (positively homogeneous) stationarity measure, which results in reaching stationarity after fewer iterations. This certainly contributes to the poor approximation of both π^* and μ_1^* as discussed above.

Moreover, when performing tests with instances other than the one we have just discussed or with observation domains other than those shown in the figures, we encountered another type of behavior of the trust region method that we did not encounter in the previous numerical example: during the iteration, the trust region method periodically switches between the nonmodified and the modified subproblems and stationarity measures, while realizing only an insignificant decrease of both the target function and the stationarity measures. Although this behavior is not inconsistent with a possible convergence result (a small decrease of the objective function was achieved in each step), it is, similar to



(a) $\lambda = 10^0$ (b) $\lambda = 10^{-2}$ (c) $\lambda = 10^{-4}$

(d) $\lambda = 10^0$ (e) $\lambda = 10^{-2}$ (f) $\lambda = 10^{-4}$

Figure 6.7: Top row: Comparison of μ_1^* (red bars) with $\bar{\mu}_1$ (blue bars) for different values of the weighting parameter. Bottom row: Comparison of the sparsity pattern of π^* (red squares) with the sparsity pattern of $\bar{\pi}$ (blue squares) for different values of the weighting parameter. The gray areas represent the domains $D_1 \subset \Omega_1$ and $D \subset \Omega$.

λ	Stat. measure	Target val.	Iter. count	Resid.
10^0	$7.0115 \cdot 10^{-5}$	$5.2670 \cdot 10^{-6}$	68	$2.2509 \cdot 10^{-2}$
10^{-1}	$4.7459 \cdot 10^{-5}$	$1.3738 \cdot 10^{-2}$	54	$5.9141 \cdot 10^{-2}$
10^{-2}	$8.0479 \cdot 10^{-5}$	$1.3291 \cdot 10^{-2}$	60	$5.8373 \cdot 10^{-2}$
10^{-3}	$9.4643 \cdot 10^{-5}$	$1.3772 \cdot 10^{-2}$	4	$6.8351 \cdot 10^{-2}$
10^{-4}	$1.0745 \cdot 10^{-5}$	$1.3760 \cdot 10^{-2}$	4	$6.8500 \cdot 10^{-2}$
10^{-5}	$1.0181 \cdot 10^{-5}$	$1.3759 \cdot 10^{-2}$	4	$6.8516 \cdot 10^{-2}$

Table 6.4: End of iteration data for different values of the weighting parameter λ. Residual refers to the sum of the squared residual of $\bar{\mu}_1$ to μ_1^* and $\bar{\pi}$ to π^*, i.e., $\mathcal{J}(\bar{\pi}, \bar{\mu}_1)$ for $D = \Omega$ and $D_1 = \Omega_1$.

the third case mentioned in the previous experiment, far from ideal.

Discussion & final remarks

The two numerical experiments show that the constrained nonsmooth trust region method from Algorithm 6.10 produces accurate and reasonable results when it is applied in the context of the transportation identification problem (TIP). In most tests, the modified subproblem and the modified stationarity measure are not needed to obtain results with high accuracy (in terms of squared residuals or stationarity). In many cases, the modified subproblem was only invoked a handful of times to achieve a stationarity of below than 10^{-6}.

However, we cannot deny that the presence of test cases where the modified subproblem and the modified stationarity measure were heavily used and where the trust region method did not converge within our limit of 200 iterations (and probably never would have done). Up to this point of time, we cannot say what

exactly triggers these cases and how we can circumvent them to make the trust region method more robust.

What we can say with certainty, however, is that our implementation leaves room for improvement and further research:

- The descent directions that we obtained in our experiments virtually always satisfied the corresponding constrained Cauchy decrease conditions in (6.7) and (6.8). While this may be sufficient to prove convergence results, this is not enough to guarantee a certain rate of convergence. We do not know how we can (ideally with little effort) obtain solutions to the subproblems that guarantee to realize a substantial reduction of the objective function and the stationarity measure.

- In our implementation, we have not accounted for mass matrices, i.e., we have not considered the effect of different mesh sizes on the individual steps of the iteration, so that the performance of our implementation of the trust region method strongly depends on the size of the marginals. In particular, the value of the stationarity measure tends to increase as the number of variables increases, and it becomes less likely that the method converges if the stationarity tolerance is fixed to e.g. 10^{-6}. For this reason, we had to significantly increase the stationarity tolerance in the second experiment to ensure convergence of the method.

- A more sophisticated description of the Bouligand subdifferential could pave the way for a computable approximation of the subdifferential around the current iterate, see e.g. [21, Section 5.2], where this was realized for an optimal control problem which is constrained by a variational inequality. A better description would yield the advantage to not be dependent of a (possibly bad) heuristic to compute a large enough set of subgradients. If this, however, proves to be impossible, a way to improve the approximation of the collective Bouligand subdifferential nevertheless would be to replace the rather static choice of points around the current iterate by a more sophisticated heuristic that would, for example, choose the next point depending on how close the subgradients of different points were together.

- In all our numerical tests, we have never observed a case in which any combination of the regularization parameters γ and ε would not have converged, i.e., by slightly changing the regularization parameters, the method always provided a result within the specified tolerances. This observation leaves room for the implementation of a path-following heuristic, i.e., an automatic control and adjustment of the regularization parameters during the runtime of the trust region method.

- During the numerical test, we observed that the semismooth Newton method repeatedly failed to converge, when choosing both regularization parameters below 10^{-4}, and thus leading to inaccurate transport plans. However, since small regularization parameters provided the most accurate results, see Figure 6.2, it would certainly be worth investing more work to improve this point and, for example, choose a different regularization parameter $\tilde{\varepsilon} = \Phi(\gamma)$ that is a function of γ and tuned in a way that the product $\gamma\tilde{\varepsilon}$ does not go towards 0 too quickly.

Appendix

Appendix A

On the Convolution of Marginals With Mollifiers

We use the first chapter of the appendix, to investigate the properties of the convolution of a measure with a mollifier. Let $d \in \mathbb{N}$ and some compact subset $X \subset \mathbb{R}^d$ be given. We start by recalling the definition of the convolution of a regular Borel measure with a mollifier.

Definition A.1. Let $\mu \in \mathfrak{M}(X)$ with $\mu \geq 0$ be a nonnegative regular Borel measure and let φ_δ, given some $\delta > 0$, be a mollifier in the sense of Definition 3.12, i.e., $\varphi_\delta \in C_c^\infty(\mathbb{R}^d)$ such that $\operatorname{supp}(\varphi_\delta) = \overline{B(0;\delta)}$, $\varphi_\delta \geq 0$, and $\int_{\mathbb{R}^d} \varphi_\delta \, dx = 1$. Then, the *convolution (of μ with φ_δ)* is defined by

$$(\varphi_\delta * \mu)(x) \coloneqq \int_X \varphi_\delta(x-y) \, d\mu(y) \quad \text{for all } x \in \mathbb{R}^d.$$

Remark A.2. One could generalize the above definition by allowing arbitrary measures on $\mathbb{R}^d$ and arbitrary measurable functions, see e.g. [69, Definition 14.4].

However, as we are about to see, the convolution of a regular Borel measure with a mollifier has some advantageous properties. Hence, we restrict ourselves to the setting given in Definition A.1. ○

Theorem A.3. *Let μ be a nonnegative regular Borel measure and φ_δ be a mollifier. Then, the convolution $\varphi_\delta * \mu \colon \mathbb{R}^d \to \mathbb{R}_+$ is a smooth, measurable, nonnegative, bounded, and compactly supported function whose support satisfies*

$$\operatorname{supp}(\varphi_\delta * \mu) \subset \overline{B(0;\delta)} + X.$$

Proof. To begin with, the nonnegativity and the boundedness of the mollifier as well as the nonnegativity of the measure directly yield that

$$0 \leq (\varphi_\delta * \mu)(x) \leq \|\varphi_\delta\|_\infty \|\mu\|_{\mathfrak{M}(X)} < \infty$$

for all $x \in \mathbb{R}^d$, so that $\varphi_\delta * \mu$ indeed is a finite, nonnegative, and bounded function on $\mathbb{R}^d$.

To see that the convolution is continuous, let $x_0 \in \mathbb{R}^d$ and $\varepsilon > 0$ be arbitrary. Since μ is (inner) regular, there is a compact subset $K \subset X$ with

$$\mu(X \setminus K) < \frac{\varepsilon}{4\|\varphi_\delta\|_\infty}.$$

We then calculate

$$\begin{aligned}|(\varphi_\delta * \mu)(x) - (\varphi_\delta * \mu)(x_0)| &\leq \int_X |\varphi_\delta(x-y) - \varphi_\delta(x_0-y)|\, d\mu(y) \\ &\leq \int_{X\setminus K} |\varphi_\delta(x-y)| + |\varphi_\delta(x_0-y)|\, d\mu(y) \\ &\quad + \int_K |\varphi_\delta(x-y) - \varphi_\delta(x_0-y)|\, d\mu(y) \\ &< \frac{\varepsilon}{2} + \int_K |\varphi_\delta(x-y) - \varphi_\delta(x_0-y)|\, d\mu(y)\,.\end{aligned}$$

Because the mapping $(x,y) \mapsto \varphi_\delta(x-y)$ is continuous on the compact set $\overline{B(x_0;1)} \times K$, the Heine-Cantor theorem ensures its uniform continuity. That is, there exists some $\rho \in (0,1)$ such that

$$|\varphi_\delta(x-y) - \varphi_\delta(\tilde{x}-\tilde{y})| < \frac{\varepsilon}{2\mu(K)}$$

as long as $(x,y),(\tilde{x},\tilde{y}) \in \overline{B(x_0;1)} \times K$ and $\|(x,y)-(\tilde{x},\tilde{y})\| < \rho$. For any $x \in B(x_0;\rho)$, we find that $\|(x,y)-(x_0,y)\| = \|x-x_0\| < \rho$ and thus

$$|\varphi_\delta(x-y) - \varphi_\delta(x_0-y)| < \frac{\varepsilon}{2\mu(K)} \quad \text{for all } y \in K.$$

Together with the above, we conclude that

$$|(\varphi_\delta * \mu)(x) - (\varphi_\delta * \mu)(x_0)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for all $x \in B(x_0;\rho)$. Since $\varepsilon > 0$ and $x_0 \in \mathbb{R}^d$ were arbitrary, the convolution $\varphi_\delta * \mu$ is a continuous (and hence measurable) function.

Next, we are going to convince ourselves that $\varphi_\delta * \mu$ is a smooth function on all of $\mathbb{R}^d$. To this end, let $x_0 \in \mathbb{R}^d$ be an arbitrary point and $j \in \{1,...,d\}$ an arbitrary index. We denote the j-th standard unit vector of $\mathbb{R}^d$ by e_j and compute the j-th partial derivative of the convolution to be equal to

$$\begin{aligned}\frac{\partial}{\partial x_j}(\varphi_\delta * \mu)(x_0) &= \lim_{h\to 0} \frac{(\varphi_\delta * \mu)(x_0 + he_j) - (\varphi_\delta * \mu)(x_0)}{h} \\ &= \lim_{h\to 0} \int_X \frac{\varphi_\delta(x_0 + he_j - y) - \varphi_\delta(x_0 - y)}{h}\, d\mu(y)\,.\end{aligned} \tag{A.1}$$

For arbitrary $y \in X$, the differentiability of φ_δ implies that

$$\lim_{h\to 0} \frac{\varphi_\delta(x_0 + he_j - y) - \varphi_\delta(x_0 - y)}{h} = \frac{\partial}{\partial x_j}\varphi_\delta(x_0 - y).$$

Thus, the integrand in (A.1) converges pointwisely to $\frac{\partial}{\partial x_j}\varphi_\delta(x_0 - \cdot\,)$ and is, independently of x_0, y and h, bounded by

$$\left|\frac{\varphi_\delta(x_0 + he_j - y) - \varphi_\delta(x_0 - y)}{h}\right| \leq L_{\varphi_\delta} \frac{\|(x_0 + he_j - y) - (x_0 - y)\|}{h} = L_{\varphi_\delta},$$

where $L_{\varphi_\delta} > 0$ denotes the Lipschitz constant of φ_δ. Consequently, we apply Lebesgue's dominated convergence theorem to receive

$$\frac{\partial}{\partial x_j}(\varphi_\delta * \mu)(x_0) = \int_X \frac{\partial}{\partial x_j}\varphi_\delta(x_0 - y)\ d\mu(y) = \left(\frac{\partial \varphi_\delta}{\partial x_j} * \mu\right)(x_0)$$

for all $x_0 \in \mathbb{R}^d$. Thanks to our former considerations, we know that $\frac{\partial \varphi_\delta}{\partial x_j} * \mu$ is a continuous function on $\mathbb{R}^d$ and so is $\frac{\partial}{\partial x_j}(\varphi_\delta * \mu)$. Since j was arbitrary, the convolution is (totally) differentiable with derivative

$$D(\varphi_\delta * \mu) = \begin{pmatrix} \frac{\partial \varphi_\delta}{\partial x_1} * \mu & \dots & \frac{\partial \varphi_\delta}{\partial x_d} * \mu \end{pmatrix} \tag{A.2}$$

and thus $\varphi_\delta * \mu \in C^1(\mathbb{R}^d)$. Using the relation in (A.2), an induction argument can be applied to arrive at $\varphi_\delta * \mu \in C^\infty(\mathbb{R}^d)$.

It remains to show the statement concerning the support of $\varphi_\delta * \mu$. To this end, choose an arbitrary point $z \in \mathbb{R}^d$ with $z \notin \operatorname{supp}(\varphi_\delta) + \operatorname{supp}(\mu)$. Because $\operatorname{supp}(\varphi_\delta)$ and $\operatorname{supp}(\mu)$ are both closed and bounded sets and because the Minkowski addition preserves closedness[1], we can find an open neighborhood $U_z \subset \mathbb{R}^d$ of z that is disjoint to $\operatorname{supp}(\varphi_\delta) + \operatorname{supp}(\mu)$. Hence, for any $\tilde{z} \in U_z$,

$$\tilde{z} \neq x + y \quad \text{for all } x \in \operatorname{supp}(\varphi_\delta),\ y \in \operatorname{supp}(\mu),$$

or equivalently,

$$\tilde{z} - y \notin \operatorname{supp}(\varphi_\delta) \quad \text{for all } y \in \operatorname{supp}(\mu).$$

Thereby,

$$(\varphi_\delta * \mu)(\tilde{z}) = \int_X \varphi_\delta(\tilde{z} - y)\ d\mu(y) = \int_{\operatorname{supp}(\mu)} \varphi_\delta(\tilde{z} - y)\ d\mu(y) = 0$$

for all $\tilde{z} \in U_z$. This implies $z \notin \operatorname{supp}(\varphi_\delta * \mu)$ and, therefore,

$$\operatorname{supp}(\varphi_\delta * \mu) \subset \operatorname{supp}(\varphi_\delta) + \operatorname{supp}(\mu) \subset \overline{B(x;\delta)} + X$$

as claimed.

In particular, since $\operatorname{supp}(\varphi_\delta) \subset \overline{B(0;\delta)}$ and $\operatorname{supp}(\mu) \subset X$ are compact sets and the Minkowski addition preserves compactness, we receive that $\varphi_\delta * \mu \in C_c^\infty(\mathbb{R}^d)$. □

A quick calculation shows that we can bound the L^p norm of the convolution:

Lemma A.4. *For $p \in [1, \infty)$ as well as μ and φ_δ as in Definition A.1,*

$$\|\varphi_\delta * \mu\|_{L^p(\mathbb{R}^d)} \leq \|\varphi_\delta\|_{L^p(\overline{B(0;\delta)})} \|\mu\|_{\mathfrak{M}(X)}$$

and equation holds for $p = 1$.

Proof. For $p > 1$, we apply Minkowski's integral inequality to estimate

$$\|\varphi_\delta * \mu\|_{L^p(\mathbb{R}^d)} = \left(\int_{\mathbb{R}^d} \left| \int_X \varphi_\delta(x - y)\ d\mu(y) \right|^p dx\right)^{\frac{1}{p}}$$

[1] In fact, even if one of these sets were unbounded, the closedness is still preserved.

$$\leq \int_X \left(\int_{\mathbb{R}^d} |\varphi_\delta(x-y)|^p \, dx \right)^{\frac{1}{p}} d\mu(y)$$
$$= \int_X \left(\int_{\overline{B(0;\delta)}} |\varphi_\delta(x)|^p \, dx \right)^{\frac{1}{p}} d\mu(y) = \|\varphi_\delta\|_{L^p(\overline{B(0;\delta)})} \|\mu\|_{\mathfrak{M}(X)}.$$

In the case $p = 1$, the equation is due to Fubini's theorem. □

We now consider a monotonously vanishing sequence of radii $\delta_n \searrow 0$ and the corresponding sequence of scaled mollifiers $(\varphi_n)_{n\in\mathbb{N}} \subset C_c^\infty(\mathbb{R}^d)$ with $\varphi_n \geq 0$, $\operatorname{supp}(\varphi_n) \subset \overline{B(0;\delta_n)}$, and $\int_{\mathbb{R}^d} \varphi_n \, dx = 1$ for all $n \in \mathbb{N}$. Abbreviate $X_n := X + \overline{B(0;\delta_n)}$.

In the following, we convince ourselves that the sequence of convolutions $(\varphi_n * \mu)_{n\in\mathbb{N}}$ converges weakly* towards μ, if we interpret the former as a sequence of density functions of measures.

Lemma A.5. *Let $0 \leq \mu \in \mathfrak{M}(X)$ be a nonnegative regular Borel measure. Then,*

$$\int_{X_n} v(\varphi_n * \mu) \, d\lambda \xrightarrow[n\to\infty]{} \langle \mu, v|_X \rangle_{C(X)^*, C(X)} \quad \text{for all } v \in C(\mathbb{R}^d).$$

Moreover, if $(\mu_n)_{n\in\mathbb{N}} \subset \mathfrak{M}(X)$, with $\mu_n \geq 0$ for all $n \in \mathbb{N}$, is a sequence of nonnegative regular Borel measures that converges weakly towards some $\mu \in \mathfrak{M}(X)$, then*

$$\int_{X_n} v(\varphi_n * \mu_n) \, d\lambda \xrightarrow[n\to\infty]{} \langle \mu, v|_X \rangle_{C(X)^*, C(X)} \quad \text{for all } v \in C(\mathbb{R}^d).$$

Proof. To prove the first part, we once again apply Fubini's theorem to obtain that

$$\begin{aligned}
\int_{X_n} v(\varphi_n * \mu) \, d\lambda &= \int_{X_n} v(x) \int_X \varphi_n(x-y) \, d\mu(y) \, dx \\
&= \int_X \int_{X_n} v(x) \varphi_n(x-y) \, dx \, d\mu(y) \\
&= \int_X \int_{\mathbb{R}^d} v(x) \tilde{\varphi}_n(y-x) \, dx \, d\mu(y) \\
&= \langle \mu, (v * \tilde{\varphi}_n)|_X \rangle_{C(X)^*, C(X)}.
\end{aligned}$$

In the above, $\tilde{\varphi}_n(x) := \varphi_n(-x)$ denotes the reflection of φ_n, which inherits the same properties. It is broadly known that the convolution (of functions) satisfies $v * \tilde{\varphi}_n \to v$ in $C(\mathbb{R}^d)$ as $n \to \infty$, see e.g. [48, Theorem 1.3.2]. In particular, $(v * \tilde{\varphi}_n)|_X \to v|_X$ in $C(X)$. Since, according to the Riesz-Markov theorem, $\mu \in \mathfrak{M}(X)$ can be identified[2] with a continuous functional on $C(X)$, we obtain that

$$\int_{X_n} v(\varphi_n * \mu) \, d\lambda \xrightarrow[n\to\infty]{} \langle \mu, v|_X \rangle_{C(X)^*, C(X)}.$$

[2] To ease notation, we refrain from explicitly stating the isometric isomorphism that maps $\mathfrak{M}(X)$ onto $C(X)^*$. A reference on this topic can be found in, e.g. [31].

The proof to the second part is analogous. Again by Fubini's theorem,

$$\int_{X_n} v(\varphi_n * \mu_n)\ d\lambda = \langle \mu_n, (v * \tilde{\varphi}_n)|_X \rangle_{C(X)^*, C(X)}$$

and because of

$$\begin{aligned} &\left| \langle \mu_n, (v * \tilde{\varphi}_n)|_X \rangle_{C(X)^*, C(X)} - \langle \mu, v|_X \rangle_{C(X)^*, C(X)} \right| \\ &\qquad \leq \|\mu_n\|_{\mathfrak{M}(X)} \|(v * \tilde{\varphi}_n)|_X - v|_X\|_{C(X)} + \left| \langle \mu - \mu_n, v \rangle_{C(X)^*, C(X)} \right| \end{aligned}$$

and the boundedness of the sequence $(\mu_n)_{n\in\mathbb{N}}$, this directly shows that

$$\int_{X_n} v(\varphi_n * \mu_n)\ d\lambda \ \xrightarrow[n\to\infty]{}\ \langle \mu, v|_X \rangle_{C(X)^*, C(X)}.$$

□

Appendix B

On the Theory of Measure & Integration

Definition B.1. Given measurable spaces $(X_1, \mathfrak{A}_1)$ and $(X_2, \mathfrak{A}_2)$, we define the *product σ-algebra* of $\mathfrak{A}_1$ and $\mathfrak{A}_2$ on $X_1 \times X_2$ by

$$\mathfrak{A}_1 \otimes \mathfrak{A}_2 \coloneqq \sigma\left(\{P_i^{-1}(A_i) \colon A_i \in \mathfrak{A}_i,\ i = 1, 2\}\right).$$

Here, σ denotes the σ-operator (which generates the smallest σ-algebra that contains the set $\mathcal{M}$) and $P_i \colon X_1 \times X_2 \to X_i$ as usual denotes the projection map $(x_1, x_2) \mapsto x_i$.

Lemma B.2. *Consider arbitrary measurable spaces $(X_1, \mathfrak{A}_1)$ and $(X_2, \mathfrak{A}_2)$ and $\mathfrak{A}_1$-$\mathfrak{B}(\mathbb{R})$- and $\mathfrak{A}_2$-$\mathfrak{B}(\mathbb{R})$-measurable functions $f_1 \colon X_1 \to \mathbb{R}$ and $f_2 \colon X_2 \to \mathbb{R}$, respectively. Then, for any continuous function $g \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the function $g \circ (f_1, f_2) \colon X_1 \times X_2 \to \mathbb{R}$ is an $(\mathfrak{A}_1 \otimes \mathfrak{A}_2)$-$\mathfrak{B}(\mathbb{R})$-measurable function.*

Proof. The function g is continuous on $\mathbb{R} \times \mathbb{R}$. Hence, it is $\big(\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})\big)$-$\mathcal{B}(\mathbb{R})$-measurable. It suffices to show the $(\mathfrak{A}_1 \otimes \mathfrak{A}_2)$-$\big(\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})\big)$-measurability of the mapping $(f_1 \times f_2)(x, y) \coloneqq (f_1(x), f_2(x))$. For this purpose, we choose some element B of the generator of $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$, i.e., $B \in \{\pi_i^{-1}(A) \colon A \in \mathcal{B}(\mathbb{R}),\ i = 1, 2\}$, where π_i is the projection from $\mathbb{R} \times \mathbb{R}$ onto the i-th component. Without loss of generality, we assume that $B \in \{\pi_1^{-1}(A) \colon A \in \mathcal{B}(\mathbb{R})\}$. Then, $B = A \times \mathbb{R}$ with some $A \in \mathcal{B}(\mathbb{R})$ and we calculate

$$\begin{aligned}(f_1 \times f_2)^{-1}(B) &= \{(x_1, x_2) \colon (f_1(x_1), f_2(x_2)) \in A \times \mathbb{R}\} \\ &= f_1^{-1}(A) \times X_2 = P_{X_1}^{-1}(f_1^{-1}(A)).\end{aligned}$$

Since f_1 is $\mathfrak{A}_1$-$\mathcal{B}(\mathbb{R})$-measurable, this is some element of the generator of $\mathfrak{A}_1 \otimes \mathfrak{A}_2$ and thus an element of the latter, which implies the desired measurability. □

Corollary B.3. *Given non-empty and compact Euclidean sets X_1 and X_2 as well as functions $f_1 \in L^2(X_1)$ and $f_2 \in L^2(X_2)$, the functions $(f_1 \oplus f_2)(x_1, x_2) \coloneqq f_1(x_1) + f_2(x_2)$ and $(f_1 \otimes f_2)(x_1, x_2) \coloneqq f_1(x_1) f_2(x_2)$ (both defined λ-a.e. on $X_1 \times X_2$) are elements of $L^2(X_1 \times X_2)$.*

Proof. By definition of $L^2(X_i)$, the function f_i is $\mathfrak{B}(X_i)$-$\mathfrak{B}(\mathbb{R})$-measurable. Both mappings $(r_1, r_2) \mapsto r_1 + r_2$ and $(s_1, s_2) \mapsto s_1 s_2$ are continuous on $\mathbb{R} \times \mathbb{R}$.

Hence, by Lemma B.2, $f_1 \oplus f_2$ and $f_1 \otimes f_2$ are measurable. We set $X \coloneqq X_1 \times X_2$ and $\lambda \coloneqq \lambda_1 \otimes \lambda_2$. Due to Fubini's theorem,

$$\begin{aligned}\|f_1 \otimes f_2\|_{L^2(X)}^2 &= \int_X |f_1 \otimes f_2|^2 \, d\lambda \\ &= \int_{X_1} |f_1|^2 \int_{X_2} |f_2|^2 \, d\lambda_2 \, d\lambda_1 = \|f_1\|_{L^2(X_1)}^2 \|f_2\|_{L^2(X_2)}^2 < \infty,\end{aligned}$$

and, additionally using the Cauchy-Schwarz inequality,

$$\begin{aligned}&\|f_1 \oplus f_2\|_{L^2(X)}^2 \\ &\quad= \int_X |f_1 \oplus f_2|^2 \, d\lambda \\ &\quad= \int_X |f_1|^2 \oplus |f_2|^2 \, d\lambda + 2 \int_X f_1 \otimes f_2 \, d\lambda \\ &\quad\le |X_2| \int_{X_1} |f_1|^2 \, d\lambda_1 + |X_1| \int_{X_2} |f_2|^2 \, d\lambda_2 + 2 \|\mathbb{1}\|_{L^2(X)} \|f_1 \otimes f_2\|_{L^2(X)} \\ &\quad= |X_2| \|f_1\|_{L^2(X_1)}^2 + |X_1| \|f_2\|_{L^2(X_2)}^2 + 2 |X_1| |X_2| \|f_1 \otimes f_2\|_{L^2(X)} < \infty,\end{aligned}$$

since X_1 and X_2 are bounded. □

Remark B.4. For the rest of this chapter, $(X, \mathfrak{A})$ will be an arbitrary measurable space. ○

Definition B.5. A *signed measure* on $(X, \mathfrak{A})$ is a set function $\mu \colon \mathfrak{A} \to \mathbb{R}$ that satisfies

1. $\mu(\emptyset) = 0$,
2. $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$ for all sequences of pairwise disjoint sets $\{A_n\}_{n \in \mathbb{N}} \subset \mathfrak{A}$.

Remark B.6. According to the above definition, a signed measure μ is a function which maps elements of $\mathfrak{A}$ to the real numbers and never takes the values $\pm\infty$. However, there are authors who allow the extended real or complex numbers as choices for a signed measure's codomain. As this is not required for our purposes, we will not follow this approach. ○

Definition B.7. We define the *variation measure* of a signed measure μ by

$$\begin{aligned}|\mu|(A) &\coloneqq \sup\left\{\sum_{i \in \mathbb{N}} |\mu(A_i)| \colon A_i \in \mathfrak{A} \text{ disjoint}, \ A = \dot{\bigcup_{i \in \mathbb{N}}} A_i\right\} \\ &= \sup\left\{\sum_{i=1}^{n} |\mu(A_i)| \colon A \supset A_1, \ldots, A_n \in \mathfrak{A} \text{ disjoint}, \ n \in \mathbb{N}\right\}\end{aligned}$$

for all $A \in \mathfrak{A}$.

Lemma B.8. *The variation measure $|\mu|$ of a finite signed measure μ is indeed a measure on $(X, \mathfrak{A})$, i.e.,*

1. $|\mu| \colon \mathfrak{A} \to [0, \infty]$,

2. $|\mu|(\emptyset) = 0$*, and for countably many disjoint* $A_i \in \mathfrak{A}$ *we have*

$$|\mu|\left(\dot{\bigcup_{i\in\mathbb{N}}} A_i\right) = \sum_{i\in\mathbb{N}} |\mu|(A_i).$$

Furthermore, it holds that

3. $|\mu|$ *is a finite measure,*

4. if $\mu \geq 0$*, then* $|\mu| = \mu$*.*

Proof. Ad *1.*: This property is clear from the definition of the variation measure.

Ad *2.*: $|\mu|(\emptyset) = 0$ follows from $\mu(\emptyset) = 0$. To show the σ-additivity, let $(A_i)_{i\in\mathbb{N}} \subset \mathfrak{A}$ be an arbitrary sequence of disjoint sets.

Ad "$\leq$": Let $B_1, \dots, B_m \in \mathfrak{A}$ be disjoint with $B_1, \dots, B_m \subset \bigcup_{i\in\mathbb{N}} A_i$. Then, because of $B_j = B_j \cap \bigcup_{i\in\mathbb{N}} A_i = \bigcup_{i\in\mathbb{N}} B_j \cap A_i$ and the σ additivity of μ,

$$\begin{aligned}\sum_{j=1}^{m} |\mu(B_j)| = \sum_{j=1}^{m} \left|\mu\Big(\bigcup_{i\in\mathbb{N}} B_j \cap A_i\Big)\right| &= \sum_{j=1}^{m} \left|\sum_{i\in\mathbb{N}} \mu(B_j \cap A_i)\right| \\ &\leq \sum_{j=1}^{m} \sum_{i\in\mathbb{N}} |\mu(B_j \cap A_i)| \\ &= \sum_{i\in\mathbb{N}} \sum_{j=1}^{m} |\mu(B_j \cap A_i)| \leq \sum_{i\in\mathbb{N}} |\mu|(A_i),\end{aligned}$$

since $(B_1 \cap A_i), \dots, (B_m \cap A_i) \subset A_i$ are disjoint. Taking the supremum over all disjoint partitions $B_1, \dots, B_m \subset \bigcup_{i\in\mathbb{N}} A_i$ yields the claim.

Ad "$\geq$": We first show that, for arbitrary $M \in \mathbb{N}$,

$$s_M \coloneqq \sum_{i=1}^{M} |\mu|(A_i) \leq |\mu|\Big(\bigcup_{i\in\mathbb{N}} A_i\Big). \tag{B.1}$$

Given $i \in \{1, \dots, M\}$, choose arbitrary disjoint sets $B_1^i, \dots, B_{m_i}^i \subset A_i$. Then, the sets

$$B_1^1, \dots, B_{m_1}^1, \dots\dots, B_1^M, \dots, B_{m_M}^M \subset \bigcup_{i=1}^{M} A_i \subset \bigcup_{i\in\mathbb{N}} A_i$$

are disjoint and therefore

$$\sum_{i=1}^{M} \sum_{j=1}^{m_i} |\mu(B_j^i)| \leq |\mu|\Big(\bigcup_{i\in\mathbb{N}} A_i\Big).$$

If we take, for each $i = 1, \dots, M$, the supremum over all disjoint partitions, we arrive at (B.1). The claim then follows by passing to the limit $M \to \infty$ (s_M is monotonically increasing and bounded, thus convergent).

Ad *3.*: If the sigma algebra $\mathfrak{A}$ contains only finitely many elements, then the finiteness of $|\mu|$ follows trivially from the finiteness of μ. Otherwise, assume

that $|\mu|(X) = \infty$. The definition of the variation measure and the finiteness of μ imply the existence of a disjoint partition $A_1 \dot\cup \ldots \dot\cup A_n = X$ with

$$\sum_{i=1}^{n} |\mu(A_i)| > 2(|\mu(X)| + 1). \tag{B.2}$$

At least two of the sets $A_1, \ldots, A_n$ are non-empty, otherwise (B.2) would be violated. We choose one of the non-empty sets A_i. If $|\mu(A_i)| > |\mu(X)| + 1$, we set $A \coloneqq A_i$ and $A \coloneqq X \setminus A_i$ otherwise. By construction,

$$|\mu(A)| > |\mu(X)| + 1 \geq 1.$$

If we set $B \coloneqq X \setminus A$, we find that

$$|\mu(B)| = |\mu(X) - \mu(A)| \geq |\mu(A)| - |\mu(X)| > 1.$$

The additivity of $|\mu|$ then implies that

$$\infty = |\mu|(X) = |\mu|(A) + |\mu|(B),$$

i.e., either $|\mu|(A) = \infty$ or $|\mu|(B) = \infty$. If $|\mu|(A) = \infty$, then we set $E_1 \coloneqq B$ (otherwise $E_1 \coloneqq B$) and repeat the above argument for the set A (otherwise B) to receive a set E_2. This approach yields a sequence of disjoint sets $(E_i)_{i\in\mathbb{N}}$ with $|\mu(E_i)| > 1$ for all $i \in \mathbb{N}$. In particular, the series $\sum_{i=1}^{\infty} \mu(E_i)$ cannot converge, which is a contradiction to

$$\sum_{i=1}^{\infty} \mu(E_i) = \mu\left(\bigcup_{i=1}^{\infty} E_i\right) \in \mathbb{R}.$$

Consequently, it must hold that $|\mu|(X) < \infty$.

Ad *4.*: Let $A \in \mathfrak{A}$ be an arbitrary measurable set. By assumption, μ is a positive measure. Together with μ's σ-additivity, we immediately obtain that

$$\sum_{i=1}^{n} |\mu(A_i)| = \sum_{i=1}^{n} \mu(A_i) = \mu\left(\bigcup_{i=1}^{n} A_i\right) = \mu(A),$$

for any disjoint partition $A_1, A_2, \ldots$ of A. Thus,

$$|\mu|(A) = \sup\left\{\mu(A)\colon A_1, \ldots, A_n,\ A = \bigcup_{i=1}^{n} A_i,\ n \in \mathbb{N}\right\} = \mu(A).$$

□

The following lemma summarizes a number of general results in measure theory that revolve around the so-called Jordan decomposition of signed measures.

Lemma B.9 ([31, Kapitel VII])**.** *For a signed measure μ there are finite measures μ^+ and μ^- on $(X, \mathfrak{A})$, which together we call the* Jordan decomposition *of μ, such that*

1. $\mu^+, \mu^- \geq 0$,

2. $\mu = \mu^+ - \mu^-$,

3. $|\mu| = \mu^+ + \mu^-$,

4. $\mu^+ \perp \mu^-$, *i.e., there is a disjoint decomposition of* $X = A \dot\cup B$ *with* $A, B \in \mathfrak{A}$ *and* $\mu^+(A) = 0 = \mu^-(B)$,

5. *the Jordan decomposition is unique (except for* μ *null sets).*

We sometimes call μ^+ *the* nonnegative part *and* μ^- *the* nonpositive part *of* μ.

Definition B.10. Let μ be a signed measure with Jordan decomposition $\mu = \mu^+ - \mu^-$. For an $\mathfrak{A}$-$\mathfrak{B}(\overline{\mathbb{R}})$-measurable function $f\colon X \to \overline{\mathbb{R}}$, we define the *(signed) integral* of f with respect to the signed measure μ by

$$\int_X f \, d\mu = \int_X f \, d\mu^+ - \int_X f \, d\bar{\mu},$$

whenever at least one of the Lebesgue integrals on the right hand side is finite.

We say an $\mathfrak{A}$-$\mathfrak{B}(\overline{\mathbb{R}})$-measurable function f is *μ-integrable* if $\left|\int_X f \, d\mu\right| < \infty$, or equivalently $\left|\int_X f \, d\mu^{\pm}\right| < \infty$.

Lemma B.11. *Given a signed measure* $\mu\colon \mathfrak{A} \to \mathbb{R}$ *and a* μ*-integrable function* $f\colon X \to \overline{\mathbb{R}}$. *It holds that*

1. $\int_X \chi_A \, d\mu = \mu(A)$, *for all* $A \in \mathfrak{A}$,

2. f *is* $\mu^{\pm}$*-integrable.*

If $g\colon X \to \overline{\mathbb{R}}$ *is another* μ*-integrable function and* $\alpha, \beta \in \mathbb{R}$, *it also holds that*

3. $\int_X \alpha f + \beta g \, d\mu = \alpha \int_X f \, d\mu + \beta \int_X g \, d\mu$.

Proof. These properties follow immediately from the definition of the (signed) integral and the respective properties of the Lebesgue integral. □

Lemma B.12. *The (signed) integration of a function is linear w.r.t. the measure, i.e., for any two signed measures* $\mu_1, \mu_2\colon \mathfrak{A} \to \mathbb{R}$ *as well as any* μ_1*- and* μ_2*-integrable function* $f\colon X \to \overline{\mathbb{R}}$, *it holds that*

$$\int_X f \, d(\alpha_1\mu_1 + \alpha_2\mu_2) = \alpha_1 \int_X f \, d\mu_1 + \alpha_2 \int_X f \, d\mu_2$$

for all $\alpha_1, \alpha_2 \in \mathbb{R}$. *Moreover, this directly implies that* f *is* $(\alpha_1\mu_1 + \alpha_2\mu_2)$*-integrable.*

Proof. First, it is easy to verify that every linear combination of signed measures is a signed measure again. Second, for all $A \in \mathfrak{A}$, we find that

$$\begin{aligned}\int_X \chi_A \, d(\alpha_1\mu_1 + \alpha_2\mu_2) &= (\alpha_1\mu_1 + \alpha_2\mu_2)(A) \\ &= \alpha_1\mu_1(A) + \alpha_2\mu_2(A) = \alpha_1 \int_X \chi_A \, d\mu_1 + \alpha_2 \int_X \chi_A \, d\mu_2,\end{aligned}$$

i.e., the claim is valid for characteristic functions. We now consider an arbitrary nonnegative simple function $u = \sum_{i=1}^{n} a_i \chi_{A_i}$ with $a_i \in \mathbb{R}$ and $A_i \in \mathfrak{A}$ for $i = 1, \dots, n$. Then, by linearity of the integrand, see Lemma B.11,

$$\begin{aligned}\int_X u \; d(\alpha_1\mu_1 + \alpha_2\mu_2) &= \sum_{i=1}^{n} a_i \int_X \chi_{A_i} \; d(\alpha_1\mu_1 + \alpha_2\mu_2) \\ &= \alpha_1 \sum_{i=1}^{n} a_i \int_X \chi_{A_i} \; d\mu_1 + \alpha_2 \sum_{i=1}^{n} a_i \int_X \chi_{A_i} \; d\mu_2 \\ &= \alpha_1 \int_X u \; d\mu_1 + \alpha_2 \int_X u \; d\mu_2,\end{aligned}$$

i.e., the claim is valid for simple functions. We note that the above equation also implies that

$$\begin{aligned}&\int_X u \; d(\alpha_1\mu_1 + \alpha_2\mu_2)^+ - \int_X u \; d(\alpha_1\mu_1 + \alpha_2\mu_2)^- \\ &\quad = \alpha_1 \left(\int_X u \; d\mu_1^+ - \int_X u \; d\mu_1^- \right) + \alpha_1 \left(\int_X u \; d\mu_2^+ - \int_X u \; d\mu_2^- \right).\end{aligned} \tag{B.3}$$

Now, let $f \colon X \to \overline{\mathbb{R}}$ be a μ_1- and μ_2-integrable function. Set $f^+(x) := \max(f(x), 0)$ and $f^-(x) := -\min(f(x), 0)$ for all $x \in X$. Then f^+ and f^- are nonnegative and μ_1- and μ_2-integrable functions on $(X, \mathfrak{A})$ with $f = f^+ - f^-$. There exists a sequence of nonnegative simple functions $(u_n)_{n\in\mathbb{N}}$ with $u_n \to f^+$ pointwisely and $0 \le u_n \le u_{n+1}$ for all $n \in \mathbb{N}$. Applying Beppo Levi's lemma and (B.3), we obtain that

$$\begin{aligned}&\int_X f^+ \; d(\alpha_1\mu_1 + \alpha_2\mu_2) \\ &\quad = \int_X f^+ \; d(\alpha_1\mu_1 + \alpha_2\mu_2)^+ - \int_X f^+ \; d(\alpha_1\mu_1 + \alpha_2\mu_2)^- \\ &\quad = \lim_{n\to\infty} \int_X u_n \; d(\alpha_1\mu_1 + \alpha_2\mu_2)^+ - \lim_{n\to\infty} \int_X u_n \; d(\alpha_1\mu_1 + \alpha_2\mu_2)^- \\ &\quad = \lim_{n\to\infty} \left(\alpha_1 \int_X u_n \; d\mu_1^+ + \alpha_2 \int_X u_n \; d\mu_2^+ \right) \\ &\qquad - \lim_{n\to\infty} \left(\alpha_1 \int_X u_n \; d\mu_1^- + \alpha_2 \int_X u_n \; d\mu_2^- \right) \\ &\quad = \alpha_1 \int_X f^+ \; d\mu_1^+ + \alpha_2 \int_X f^+ \; d\mu_2^+ - \alpha_1 \int_X f^+ \; d\mu_1^+ - \alpha_2 \int_X f^+ \; d\mu_2^+ \\ &\quad = \alpha_1 \int_X f^+ \; d\mu_1 + \alpha_2 \int_X f^+ \; d\mu_2 \,.\end{aligned}$$

Due to $(\alpha_1\mu_1 + \alpha_2\mu_2)^+ \perp (\alpha_1\mu_1 + \alpha_2\mu_2)^-$, there exists a partition $X = P \cup N$ with $(\alpha_1\mu_1 + \alpha_2\mu_2)^+(N) = 0$ and $(\alpha_1\mu_1 + \alpha_2\mu_2)^-(P) = 0$, see Lemma B.9. Then $\chi_P f^+$ is μ_1- and μ_2-integrable and it holds that $\chi_P u_n \nearrow \chi_P f^+$ pointwisely. Therefore,

$$\alpha_1 \int_X \chi_P f^+ \; d\mu_1 + \alpha_2 \int_X \chi_P f^+ \; d\mu_2$$

$$
\begin{aligned}
&= \alpha_1 \left(\int_X \chi_P f^+ \, d\mu_1^+ - \int_X \chi_P f^+ \, d\mu_1^- \right) \\
&\quad + \alpha_2 \left(\int_X \chi_P f^+ \, d\mu_2^+ - \int_X \chi_P f^+ \, d\mu_2^- \right) \\
&= \lim_{n\to\infty} \alpha_1 \left(\int_X \chi_P u_n \, d\mu_1^+ - \int_X \chi_P u_n \, d\mu_1^- \right) \\
&\quad + \lim_{n\to\infty} \alpha_2 \left(\int_X \chi_P u_n \, d\mu_2^+ - \int_X \chi_P u_n \, d\mu_2^- \right) \\
&= \lim_{n\to\infty} \left(\int_X \chi_P u_n \, d(\alpha_1\mu_1 + \alpha_2\mu_2)^+ - \int_X \chi_P u_n \, d(\alpha_1\mu_1 + \alpha_2\mu_2)^- \right) \\
&= \int_X \chi_P f^+ \, d(\alpha_1\mu_1 + \alpha_2\mu_2)^+ - \int_X \chi_P f^+ \, d(\alpha_1\mu_1 + \alpha_2\mu_2)^- \\
&= \int_X \chi_P f^+ \, d(\alpha_1\mu_1 + \alpha_2\mu_2),
\end{aligned}
$$

where the convergence follows from Beppo Levi's lemma again. Analogously, we obtain the same relation for $\chi_N f^+$ and hence

$$
\int_X f^+ \, d(\alpha_1\mu_1 + \alpha_2\mu_2) = \alpha_1 \int_X f^+ \, d\mu_1 + \alpha_2 \int_X f^+ \, d\mu_2 \,.
$$

We use the same argumentation for the negative part f^- to ultimately arrive at

$$
\begin{aligned}
&\int_X f \, d(\alpha_1\mu_1 + \alpha_2\mu_2) \\
&\quad = \int_X f^+ \, d(\alpha_1\mu_1 + \alpha_2\mu_2) - \int_X f^- \, d(\alpha_1\mu_1 + \alpha_2\mu_2) \\
&\quad = \alpha_1 \left(\int_X f^+ \, d\mu_1 - \int_X f^- \, d\mu_1 \right) + \alpha_2 \left(\int_X f^+ \, d\mu_2 - \int_X f^- \, d\mu_2 \right) \\
&\quad = \alpha_1 \int_X f \, d\mu_1 + \alpha_2 \int_X f \, d\mu_2 \,.
\end{aligned}
$$

□

Lemma B.13. *Let $\mu\colon \mathfrak{A} \to \mathbb{R}$ be a finite signed measure and $f\colon X \to \mathbb{R}$ be an μ-integrable function. Denote the total variation of μ by $|\mu|$. Then it holds that*

$$
\left| \int_X f \, d\mu \right| \leq \int_X |f| \, d|\mu| \,.
$$

Proof. Given the Jordan decomposition of μ, namely $\mu^+, \mu^-\colon \mathfrak{A} \to \overline{\mathbb{R}}$, we know that $\mu^+, \mu^- \geq 0$ and $|\mu| = \mu^+ + \mu^-$. Then,

$$
\begin{aligned}
\left| \int_X f \, d\mu \right| &= \left| \int_X f^+ \, d\mu^+ - \int_X f^- \, d\mu^+ - \int_X f^+ \, d\mu^- + \int_X f^- \, d\mu^- \right| \\
&\leq \int_X f^+ \, d\mu^+ + \int_X f^- \, d\mu^+ + \int_X f^+ \, d\mu^- + \int_X f^- \, d\mu^- = \int_X |f| \, d|\mu|,
\end{aligned}
$$

because all of the integrals $\int_X f^\pm \, d\mu^+$, $\int_X f^\pm \, d\mu^-$ are nonnegative and Lemma B.12. □

We recall the definition of the space $\mathfrak{M}(X)$ from Chapter 2:

Definition B.14. We denote the set of signed measures $\mu\colon \mathfrak{B}(X) \to \mathbb{R}$ whose total variation measure $|\mu|$ is an (inner and outer) regular measure by $\mathfrak{M}(X)$. We call those measures *(signed) regular Borel measures*. $\mathfrak{M}(X)$ is a Banach space w.r.t. the *total variation norm*

$$\|\mu\|_{\mathfrak{M}(X)} \coloneqq |\mu|(X),$$

see e.g. [2, Theorem 6.21].

Lemma B.15. *Let $X = \{x_1, \dots, x_n\}$, $n \in \mathbb{N}$, be an arbitrary finite set and let $\mathfrak{A}$ be the* discrete σ-algebra *on X (i.e., $\mathfrak{A}$ is defined to be the power set of X). Then, $\big(\mathfrak{M}(X), \|\cdot\|_{\mathfrak{M}(X)}\big)$ is isometric isomorphic to $\big(\mathbb{R}^n, \|\cdot\|_1\big)$.*

Proof. Let $\mu \in \mathfrak{M}(X)$ and $\mathfrak{A} \ni A = \bigcup_{i\in I} x_i$ with $I \subset \{1, \dots, n\}$ be arbitrary. Then μ's additivity implies that

$$\mu\Big(\bigcup_{i\in I} x_i\Big) = \sum_{i\in I} \mu(x_i), \tag{B.4}$$

i.e., $\mu(A)$ is uniquely determined by $\{\mu(x_i)\}_{i\in I}$. Hence, the mapping $\phi\colon \mathbb{R}^n \to \mathfrak{M}(X)$ defined via

$$\phi(a_1, \dots, a_n) \coloneqq \mu \quad \text{with } \mu(\emptyset) = 0 \text{ and } \mu(x_i) = a_i \text{ for all } i = 1, \dots, n,$$

is one-to-one and a homomorphism. Moreover, (B.4) shows that ϕ is an isometry, since

$$\|\phi(a)\|_{\mathfrak{M}(X)} = |\mu|(X) = \sum_{i=1}^{n} |\mu(x_i)| = \sum_{i=1}^{n} |a_i| = \|a\|_1,$$

for all $a = (a_1, \dots, a_n)^\top \in \mathbb{R}^m$. □

Theorem B.16. *Let $X \subset \mathbb{R}^d$ be compact. Then, it holds that $L^1(X) \hookrightarrow \mathfrak{M}(X)$.*

Proof. Consider the operator

$$\iota\colon L^1(X) \to \mathfrak{M}(X), \quad \iota(f)(B) \coloneqq \int_B f \, d\lambda, \quad f \in L^1(X),\ B \in \mathfrak{B}(X),$$

with λ being the Lebesgue measure on the measurable space $\big(X, \mathfrak{B}(X)\big)$. Note that λ is finite on X, because X is bounded. We first convince ourselves that ι is well-defined.

Let $f \in L^1(X)$ be fixed but arbitrary. By construction, $\iota(f)(\emptyset) = 0$. For every $B \in \mathfrak{B}(X)$, we find that

$$|\iota(f)(B)| = \Big|\int_X \chi_B f \, d\lambda\Big| \leq \int_X |\chi_B||f| \, d\lambda \leq \int_X |f| \, d\lambda = \|f\|_{L^1(X)} < \infty,$$

where we used Lemma B.13, the monotonicity of integration, and the nonnegativity of the Lebesgue measure. Hence, $\iota(f)\colon \mathfrak{B}(X) \to \mathbb{R}$ is a finite set function.

To see that $\iota(f)$ is σ-additive, we first assume w.l.o.g. that f is nonnegative almost everywhere. Otherwise, the following argument can be made separately for its positive and negative parts. We then consider a sequence $(B_i)_{i\in\mathbb{N}}$ of

disjoint measurable sets $B_i \in \mathfrak{B}(X)$. Because $\mathfrak{B}(X)$ is a σ-algebra, $\bigcup_{i\in\mathbb{N}} B_i \in \mathfrak{B}(X)$.

$$\iota(f)\Big(\bigcup_{i\in\mathbb{N}} B_i\Big) = \int_{\bigcup_{i\in\mathbb{N}} B_i} f \, d\lambda = \int_X \chi_{\bigcup_{i\in\mathbb{N}} B_i} f \, d\lambda \overset{(1)}{=} \int_X \sum_{i\in\mathbb{N}} \big(\chi_{B_i} f\big) \, d\lambda$$
$$\overset{(2)}{=} \sum_{i\in\mathbb{N}} \int_X \chi_{B_i} f \, d\lambda = \sum_{i\in\mathbb{N}} \int_{B_i} f \, d\lambda = \sum_{i\in\mathbb{N}} \iota(f)(B_i),$$

where we used the disjointness of the sets $B_1, B_2, \ldots$ for (1) and the monotone convergence theorem together with the nonnegativity and measurability of the functions $\chi_{B_1} f, \chi_{B_2} f, \ldots$ for (2). Altogether, this shows that $\iota(f)$ is a finite (signed) Borel measure on $\big(X, \mathfrak{B}(X)\big)$.

Because X is a compact subset of $\mathbb{R}^d$, it is Polish. Ulam's theorem, see e.g. [31, Satz VIII.1.16], then ensures that $\iota(f) \in \mathfrak{M}(X)$ so that ι indeed is well-defined.

The linearity of ι follows trivially from the linearity of the Lebesgue integral. Moreover, we observe that

$$\iota(f)(B) = \int_B f_+ \, d\lambda - \int_B f_- \, d\lambda \quad \text{for all } B \in \mathfrak{B}(X).$$

Both integrals on the right-hand side are nonnegative. Therefore, the uniqueness of the Jordan decomposition, see Lemma B.9 *5.*, implies that

$$\big(\iota(f)\big)^+(B) = \int_B f_+ \, d\lambda \quad \text{and} \quad \big(\iota(f)\big)^-(B) = \int_B f_- \, d\lambda$$

for all $B \in \mathfrak{B}(X)$. Using this and Lemma B.9 *3.*,

$$\|\iota(f)\|_{\mathfrak{M}(X)} = |\iota(f)|(X) = \int_X f_+ \, d\lambda + \int_X f_+ \, d\lambda = \int_X |f| \, d\lambda = \|\iota(f)\|_{L^1(X)}$$

shows that the linear operator ι is bounded and therefore continuous.

To convince ourselves that ι is injective, let $f_1, f_2 \in L^1(X)$ with $f_1 \neq f_2$ be given. Then, there exists a subset $E \subset X$ with $\lambda(E) > 0$ such that $(f_1 - f_2)(x) > 0$ for all $x \in E$ or $(f_1 - f_2)(x) < 0$ for all $x \in E$. The subset E is measurable, because it is the preimage of $(0, \infty) \in \mathfrak{B}(\mathbb{R})$ or $(-\infty, 0) \in \mathfrak{B}(\mathbb{R})$ w.r.t. the $\mathfrak{B}(X)$-$\mathfrak{B}(\mathbb{R})$-measurable function $f_1 - f_2$. Therefore, $E \in \mathfrak{B}(X)$ and

$$\iota(f_1)(E) - \iota(f_2)(E) = \iota(f_1 - f_2)(E) = \int_E f_1 - f_2 \, d\lambda \gneqq 0$$

so that $\iota(f_1)(E) \gneqq \iota(f_2)(E)$, in particular $\iota(f_1) \neq \iota(f_2)$.

To finally establish that ι is an embedding, it remains to show that ι is an open map between $L^1(X)$ and its image $\iota\big(L^1(X)\big) \subset \mathfrak{M}(X)$. Let $O \subset L^1(X)$ be open and $\mu \in \iota(O)$ be arbitrary. Then, there must exist some $f \in O$ with $\mu = \iota(f)$ and, because of O's openness, a radius $\delta > 0$ such that $B(f; \delta) \subset O$. For every $\nu \in B(\mu; \delta) \subset \iota\big(L^1(X)\big)$ there exists some $g \in L^1(X)$ such that $\nu = \iota(g)$ and $\|\mu - \nu\|_{\mathfrak{M}(X)} < \delta$. We immediately receive that

$$\|f - g\|_{L^1(X)} = \int_X |f - g| \, d\lambda = |\iota(f - g)|(X) = \|\mu - \nu\|_{\mathfrak{M}(X)} < \delta,$$

i.e., $g \in B(f;\delta) \subset O$ and thus $\nu = \iota(g) \in \iota(O)$. Because ν was arbitrary, this shows that $B(\mu;\delta) \subset \iota(O)$ and, because μ was arbitrary, that $\iota(O)$ is an open set in $L^1(X)$. Therefore, ι maps open sets onto open sets, i.e., ι is an open map. This concludes the proof. □

Lemma B.17. *For $X \subset \mathbb{R}^d$ compact, some regular Borel measure $\mu \in \mathfrak{M}(X)$, and some measurable set $B \in \mathfrak{B}(X)$, there exists a sequence of nonnegative continuous (and thus measurable) functions $(v_\varepsilon)_{\varepsilon \searrow 0}$ on X such that*

$$\int_X v_\varepsilon \, d\mu \;\to\; \mu(B) \quad \text{as} \quad \varepsilon \searrow 0.$$

Proof. Denote by μ^+ and μ^- the Jordan decomposition of $\mu \in \mathfrak{M}(X)$, see Lemma B.9. By Ulam's theorem, μ^+ and μ^- are regular Borel measures (see, e.g. [31, Folgerung VIII.2.22]). Hence, for $\varepsilon > 0$ arbitrary, there exist both a compact set $K^+ \subset X$ and an open set $U^+ \subset X$ such that $K^+ \subset B \subset U^+$ and

$$\mu^+(B) - \varepsilon \le \mu^+(K^+) \le \mu^+(B) \le \mu^+(U^+) \le \mu^+(B) + \varepsilon.$$

Analogously, there exist both a compact set $K^- \subset X$ and an open set $U^- \subset X$ such that the above estimates hold for μ^-. We define $K \coloneqq K^+ \cup K^- \subset B$ and $U \coloneqq U^+ \cap U^- \supset B$ and notice, by Urysohn's lemma, that there exists a continuous (and bounded) function $v_\varepsilon \colon X \to [0,1]$ with $v_\varepsilon(x) = 1$ for all $x \in K$ and $v_\varepsilon(x) = 0$ for all $x \in X \setminus U$. Hence,

$$\begin{aligned}
\mu^+(B) - \varepsilon \le \mu^+(K) &= \int_X \chi_K \, d\mu^+ \\
&\le \int_X v_\varepsilon \, d\mu^+ \\
&\le \int_X \chi_U \, d\mu^+ = \mu^+(U) \le \mu^+(B) + \varepsilon,
\end{aligned}$$

and we find a similar estimate for μ^-. By definition of the (signed) integral,

$$\left| \int_X v_\varepsilon \, d\mu - \mu(B) \right| \le \left| \int_X v_\varepsilon \, d\mu^+ - \mu^+(B) \right| + \left| \int_X v_\varepsilon \, d\mu^- - \mu^-(B) \right| \le 2\varepsilon \;\to\; 0,$$

as $\varepsilon \searrow 0$, which gives the desired result. □

Lemma B.18. *Let $X \subset \mathbb{R}^d$ be compact and $\mu \in \mathfrak{M}(X)$ be a nonnegative (signed) regular Borel measure. For any $B \in \mathfrak{B}(X)$ with $B \cap \operatorname{supp}(\mu) = \emptyset$, it holds that $\mu(B) = 0$.*

Proof. Let $B \in \mathfrak{B}(X)$ be such that $B \cap \operatorname{supp}(\mu) = \emptyset$. Because μ is nonnegative and (inner) regular, for every $\varepsilon > 0$ there exists a compact set $K \subset B$ such that $\mu(K) \ge \mu(B) - \varepsilon$. Since K is a subset of B, we have that $K \cap \operatorname{supp}(\mu) = \emptyset$ and thus for any $x \in K$, by definition of $\operatorname{supp}(\mu)$, there exists an open neighborhood $N_x \in \mathfrak{B}(X)$ of x such that $\mu(N_x) = 0$. Because K is compact, the open cover $\bigcup_{x \in K} N_x \supset K$ admits a finite (possibly non-disjoint) subcover $N_{x_1} \cup \cdots \cup N_{x_N} \supset K$, where $N \in \mathbb{N}$ and $x_1, \ldots, x_N \in K$. Consequently,

$$0 = \mu(N_{x_1}) + \cdots + \mu(N_{x_n}) \ge \mu(N_{x_1} \cup \cdots \cup N_{x_N}) \ge \mu(K) \ge \mu(B) - \varepsilon$$

i.e., $\mu(B) \in [0, \varepsilon]$. Because $\varepsilon > 0$ was arbitrary, this yields that $\mu(B) = 0$. □

Appendix C

On the Theory of Optimal Transport

Lemma C.1. *Given the measurable spaces* $(X_1, \mathfrak{A}_1)$ *and* $(X_2, \mathfrak{A}_2)$ *as well as a nonnegative coupling* $\pi \in \Pi(\mu_1, \mu_2)$ *between the nonnegative marginals* $\mu_1 \in \mathfrak{M}(X_1)$ *and* $\mu_2 \in \mathfrak{M}(X_2)$*. Then, it holds that*

$$\operatorname{supp}(\pi) \subset \operatorname{supp}(\mu_1) \times \operatorname{supp}(\mu_2).$$

Proof. We argue by contradiction and assume the contrary, i.e., we assume that there exists some $x \in \operatorname{supp}(\pi) \setminus (\operatorname{supp}(\mu_1) \times \operatorname{supp}(\mu_2))$. Since both $\operatorname{supp}(\mu_1)$ and $\operatorname{supp}(\mu_2)$ are closed, so is their product $\operatorname{supp}(\mu_1) \times \operatorname{supp}(\mu_2)$ and hence there exists a radius $\rho > 0$ such that $B(x;\rho) \cap (\operatorname{supp}(\mu_1) \times \operatorname{supp}(\mu_2)) = \emptyset$.

Then, there must exist a radius $\rho > 0$ such that

$$B(x;\rho) \cap \big(\operatorname{supp}(\mu_1) \times X_2\big) = \emptyset \quad \text{and/or} \quad B(x;\rho) \cap \big(X_1 \times \operatorname{supp}(\mu_2)\big) = \emptyset.$$

If this were not the case, then we would find the points $y_n \in B\big(x; \frac{1}{n}\big) \cap \big(\operatorname{supp}(\mu_1) \times X_2\big)$ as well as $z_n \in B\big(x; \frac{1}{n}\big) \cap \big(X_1 \times \operatorname{supp}(\mu_2)\big)$ for every $n \in \mathbb{N}$. As n approaches infinity, $y_n \to x$ as well as $z_n \to x$ and the closedness of both $\operatorname{supp}(\mu_1) \times X_2$ and $X_1 \times \operatorname{supp}(\mu_2)$ would imply that

$$x \in (\operatorname{supp}(\mu_1) \times X_2) \cap (X_1 \times \operatorname{supp}(\mu_2)) = \operatorname{supp}(\mu_1) \times \operatorname{supp}(\mu_2),$$

contrary to our initial assumption.

Now, if $B(x;\rho) \cap (\operatorname{supp}(\mu_1) \times X_2) = \emptyset$, then because π is a nonnegative coupling between μ_1 and μ_2,

$$\begin{aligned}\mu_1(X_1) = \pi(X_1 \times X_2) &\geq \pi\big((\operatorname{supp}(\mu_1) \times X_2) \cup B(x;\rho)\big) \\ &= \pi\big((\operatorname{supp}(\mu_1) \times X_2)\big) + \pi\big(B(x;\rho)\big) \\ &= \mu_1\big(\operatorname{supp}(\mu_1)\big) + \pi\big(B(x;\rho)\big) = \mu_1(X_1) + \pi\big(B(x;\rho)\big)\end{aligned}$$

and consequently $\pi\big(B(x;\rho)\big) = 0$, which contradicts $x \in \operatorname{supp}(\pi)$. If $B(x;\rho) \cap \big(X_1 \times \operatorname{supp}(\mu_2)\big) = \emptyset$, we argue analogously to derive the same contradiction. □

Lemma C.2. *Consider some measurable space* $\big(X, \mathfrak{B}(X)\big)$*. If* $\pi \in \Pi(\mu_1, \mu_2)$ *is an optimal transport plan between the marginals* $\mu_1 \in \mathfrak{M}(X)$ *and* $\mu_2 \in \mathfrak{M}(X)$

w.r.t. the $\mathfrak{B}(X)$-$\mathfrak{B}(\mathbb{R})$-measurable and symmetric cost function $c\colon X \times X \to \mathbb{R}$, then there exists an optimal transport plan $\pi' \in \Pi(\mu_2, \mu_1)$ between μ_2 and μ_1 w.r.t. c which satisfies

$$\pi'(B_2 \times B_1) = \pi(B_1 \times B_2) \quad \textit{for all } B_1, B_2 \in \mathfrak{B}(X). \tag{C.1}$$

Proof. Because the system $\mathfrak{B}(X) \times \mathfrak{B}(X)$ is a generator of the sigma algebra $\mathfrak{B}(X \times X)$ that is closed under finite intersections and because π is a finite measure, the relation in (C.1) is actually the definition of the measure π', see e.g. [31, Theorem II.5.6]. Furthermore, we directly obtain that $\pi' \in \Pi(\mu_2, \mu_1)$. Consequently, there is a one-to-one correspondence between $\Pi(\mu_1, \mu_2)$ and $\Pi(\mu_2, \mu_1)$.

For each $n \in \mathbb{N}$, we choose a disjoint partition $B_1, \dots, B_{k_n} \in \mathfrak{B}(X) \setminus \{\emptyset\}$ of X[1] and symmetric simple functions

$$c_n = \sum_{i,j=1}^{k_n} c_n^{ij}\, \chi_{B_i \times B_j} \quad \text{with } c_n^{ij} = c_n^{ji} \in \mathbb{R} \text{ for all } i, j = 1, \dots, k_n, \tag{C.2}$$

in a way that the sequence $(c_n)_{n\in\mathbb{N}}$ approximates the symmetric function c uniformly. With this choice, the definition of the Lebesgue integral together with (C.1) and (C.2) yields that

$$\begin{aligned}
&\int_X c(x_1, x_2)\; d\pi(x_1, x_2) \\
&\quad = \lim_{n\to\infty} \int_X c_n(x_1, x_2)\; d\pi(x_1, x_2) \\
&\quad = \lim_{n\to\infty} \sum_{i,j=1}^{n} c_n^{ij}\, \pi(B_i \times B_j) = \lim_{n\to\infty} \sum_{i,j=1}^{n} c_n^{ji}\, \pi'(B_j \times B_i) \\
&\quad = \lim_{n\to\infty} \int_{X_2 \times X_1} c_n(x_2, x_1)\; d\pi'(x_2, x_1) = \int_X c(x_2, x_1)\; d\pi'(x_2, x_1),
\end{aligned}$$

i.e., π and π' have the same target values in their respective Kantorovich problems. This is sufficient to conclude that uniqueness of the optimal transport plan between μ_2 and μ_1 implies the uniqueness of the optimal transport plan between μ_1 and μ_2 with respect to the cost c. □

[1] Note that $\{B_1, \dots, B_{k_n}\} \times \{B_1, \dots, B_{k_n}\}$ is a disjoint partition of $X \times X$.

Appendix D

On Functional Analysis

Lemma D.1. *Suppose that* $(X, \|\cdot\|_X)$ *and* $(Y, \|\cdot\|_Y)$ *are normed vector spaces over* $\mathbb{R}$ *(or* $\mathbb{C}$*). Denote their (topological) dual spaces by* $(X^*, \|\cdot\|_{X^*})$ *and* $(Y^*, \|\cdot\|_{Y^*})$ *with the usual definitions*

$$\|f\|_{X^*} \coloneqq \sup_{\substack{x\in X,\\ \|x\|_X\leq 1}} |f(x)| = \sup_{\substack{x\in X,\\ x\neq 0}} \frac{|f(x)|}{\|x\|_X}$$

and

$$\|g\|_{Y^*} \coloneqq \sup_{\substack{y\in Y,\\ \|y\|_Y\leq 1}} |g(y)| = \sup_{\substack{y\in Y,\\ y\neq 0}} \frac{|g(y)|}{\|y\|_Y}.$$

On $X \times Y$ *and* $X^* \times Y^*$*, we define the norms*

$$\|(x,y)\|_{X\times Y} \coloneqq \|x\|_X + \|y\|_Y \quad \text{and} \quad \|(f,g)\|_{X^*\times Y^*} \coloneqq \max\{\|f\|_{X^*}, \|g\|_{Y^*}\},$$

respectively. With the above definitions, the spaces $(X \times Y, \|\cdot\|_{X\times Y})$ *and* $(X^* \times Y^*, \|\cdot\|_{X^*\times Y^*})$ *become Banach spaces (the proof is left to the reader).*

Then, the mapping

$$J\colon X^* \times Y^* \to (X \times Y)^*, \ \big(J(f,g)\big)(x,y) \coloneqq f(x) + g(y)$$

is an isometric isomorphism. Hence, $(X \times Y)^* \cong X^* \times Y^*$*.*

Remark D.2. Note that the above result only provides an isometric isomorphism between the spaces $\big((X \times Y)^*, \|\cdot\|_{(X\times Y)^*}\big)$ and $\big(X^* \times Y^*, \|\cdot\|_{X^*\times Y^*}\big)$, where

$$\|h\|_{(X\times Y)^*} = \sup_{\substack{(x,y)\in X\times Y,\\ \|(x,y)\|_{X\times Y}\leq 1}} |h(x,y)|$$

and

$$\|(f,g)\|_{X^*\times Y^*} = \max\{\|f\|_{X^*}, \|g\|_{Y^*}\}.$$

If we equip the latter space with any of the norms

$$\|(f,g)\|_p \coloneqq \big(\|f\|_{X^*}^p + \|g\|_{Y^*}^p\big)^{\frac{1}{p}}, \quad p \in [1,\infty),$$

there cannot exist an **isometric** isomorphism between the above spaces. This is due to the fact that the norms on $\mathbb{R}^2$ are equivalent but not equal. Nevertheless, the spaces are still isomorphic. ○

Proof of Lemma D.1. We first convince ourselves that J is well defined. For arbitrary $f \in X^*$ and $g \in Y^*$, the mapping $J(f,g)$ is obviously linear w.r.t. x and y. Furthermore,

$$\begin{aligned}|\big(J(f,g)\big)(x,y)| &\leq |f(x)| + |g(y)| \\ &\leq \|f\|_{X^*}\|x\|_X + \|g\|_{Y^*}\|y\|_Y \\ &\leq \max\{\|f\|_{X^*}, \|g\|_{Y^*}\}\|(x,y)\|_{X\times Y}.\end{aligned}$$

Thus $J(f,g)$ is a bounded linear operator and therefore continuous so that J is well defined.

Clearly, J is linear w.r.t. f and g and

$$\begin{aligned}\|J(f,g)\|_{(X\times Y)^*} &= \sup_{\|(x,y)\|_{X\times Y}\leq 1} |f(x)+g(y)| \\ &\leq \sup_{\|(x,y)\|_{X\times Y}\leq 1} |f(x)| + |g(y)| \\ &\leq \sup_{\|(x,y)\|_{X\times Y}\leq 1} \|f\|_{X^*}\|x\|_X + \|g\|_{Y^*}\|y\|_Y \\ &\leq \max\{\|f\|_{X^*}, \|g\|_{Y^*}\} \cdot \sup_{\|(x,y)\|_{X\times Y}\leq 1} \|x\|_X + \|y\|_Y \\ &= \|(f,g)\|_{X^*\times Y^*},\end{aligned}$$

so that J is bounded and thus continuous on $X^* \times Y^*$. To show the converse estimate, we assume without loss of generality that $\|f\|_{X^*} = \max\{\|f\|_{X^*}, \|g\|_{Y^*}\}$. Then, because of $g(0) = 0$,

$$\begin{aligned}\|J(f,g)\|_{(X\times Y)^*} &= \sup_{\|(x,y)\|_{X\times Y}\leq 1} |f(x)+g(y)| \\ &\geq \sup_{\|(x,0)\|_{X\times Y}\leq 1} |f(x)+g(0)| \\ &= \sup_{\|x\|_X\leq 1} |f(x)| = \|f\|_{X^*} = \|(f,g)\|_{X^*\times Y^*},\end{aligned}$$

which, together with the previous estimate, implies that $\|J(f,g)\|_{(X\times Y)^*} = \|(f,g)\|_{X^*\times Y^*}$, i.e., J is an isometry. Furthermore, this means that J must also be injective, because $J(f_1,g_1) = J(f_2,g_2)$ implies

$$\begin{aligned}0 = \|J(f_1,g_1) - J(f_2,g_2)\|_{(X\times Y)^*} &= \|J(f_1-f_2, g_1-g_2)\|_{(X\times Y)^*} \\ &= \|(f_1-f_2, g_1-g_2)\|_{X^*\times Y^*} \\ &= \|f_1-f_2\|_{X^*} + \|g_1-g_2\|_{Y^*},\end{aligned}$$

that is, $f_1 = f_2$ and $g_1 = g_2$ for all $(f_1,g_1),(f_2,g_2) \in X^* \times Y^*$. It remains to show the surjectivity of J. However, we see this immediately because

$$J^{-1}\colon (X\times Y)^* \to X^* \times Y^*,\ \big(J^{-1}(h)\big)(x,y) \coloneqq (h(x,0), h(0,y))$$

is the inverse function to J. □

Lemma D.3 (A Scaled Young's inequality)**.** *For any two real numbers $a, b \in \mathbb{R}$ and any positive scalar $\rho > 0$ it holds that*

$$ab \leq \frac{\rho}{2}a^2 + \frac{1}{2\rho}b^2 \leq \rho a^2 + \frac{b^2}{\rho}.$$

Proof. By binomial expansion,

$$0 \leq \left(\sqrt{\rho}a - \frac{b}{\sqrt{\rho}}\right)^2 = \rho a^2 - 2ab + \frac{b^2}{\rho}.$$

Hence,

$$ab \leq \frac{\rho}{2}a^2 + \frac{1}{2\rho}b^2 \leq \rho a^2 + \frac{b^2}{\rho}.$$

□

Lemma D.4. *Let $(X, \|\cdot\|_X)$ be a Banach space and denote by X^* its topological dual space. Then the norm on X^* is weak* lower semi-continuous.*

Proof. The norm on X^* is defined by

$$\|x^*\|_{X^*} \coloneqq \sup_{0 \neq x \in X} \frac{\left|\langle x^*, x\rangle_{X^*,X}\right|}{\|x\|_X}.$$

Accordingly,

$$\|x^*\|_{X^*}\|x\|_X \geq \left|\langle x^*, x\rangle_{X^*,X}\right| \quad \text{for all } x \in X,\ x^* \in X^*. \tag{D.1}$$

Now, consider a sequence $(x_n^*)_{n\in\mathbb{N}}$ in X^* that is weakly* convergent to some element $\bar{x}^* \in X^*$, i.e., $\langle x_n^*, x\rangle_{X^*,X} \to \langle \bar{x}^*, x\rangle_{X^*,X}$ as $n \to \infty$. By (D.1),

$$\left|\langle \bar{x}^*, x\rangle_{X^*,X}\right| = \liminf_{n\to\infty} \left|\langle x_n^*, x\rangle_{X^*,X}\right| \leq \|x\|_X \liminf_{n\to\infty} \|x_n^*\|_{X^*} \quad \text{for all } x \in X,$$

which implies

$$\frac{\left|\langle \bar{x}^*, x\rangle_{X^*,X}\right|}{\|x\|_X} \leq \liminf_{n\to\infty} \|x_n^*\|_{X^*} \quad \text{for all } x \in X$$

and therefore

$$\|\bar{x}^*\|_{X^*} \leq \liminf_{n\to\infty} \|x_n^*\|_{X^*},$$

which corresponds to the lower semi-continuity of the dual norm. □

Lemma D.5 (Lemma ohne Namen)**.** *Let (X, d) be a metric space. If $(x_n)_{n\in\mathbb{N}}$ and $x \in X$ are a sequence and a point, respectively, such that any subsequence $(x_{n_k})_{k\in\mathbb{N}}$ possesses another subsequence $(x_{n_{k_l}})_{l\in\mathbb{N}}$ with $x_{n_{k_l}} \xrightarrow[l\to\infty]{} x$ in X, then $x_n \xrightarrow[n\to\infty]{} x$ in X.*

Proof. We argue by contradiction and assume that $x_n \not\to x$ as $n \to \infty$. This, however, provides us with some $\epsilon > 0$ and some subsequence $(n_k)_{k\in\mathbb{N}}$ such that $d(x_{n_k}, x) > \epsilon$ for all $k \in \mathbb{N}$. This subsequence cannot possess a subsequence convergent to x which is a contradiction to the assumptions. Therefore, it must hold that $x_n \to x$ in X as $n \to \infty$. □

Definition D.6. We say that a mapping $f\colon X \to Y$ between the Banach spaces X and Y is *Hadamard differentiable* at a point $x \in X$, if

$$f'(x; h) = \lim_{\substack{\tilde{h}\to h,\\ t\to 0}} \frac{f(x + t\tilde{h}) - f(x)}{t},$$

for each direction $h \in X$, see e.g. [70, Definition 2.2]. Note that, in the above, $f'(x;h)$ denotes the usual *directional derivative*

$$f'(x;h) \coloneqq \lim_{t \to 0} \frac{f(x+th) - f(x)}{t} \in Y.$$

Lemma D.7. *Let $f\colon X \to Y$ be a mapping between the Banach spaces X and Y that is directionally differentiable at a point $x \in X$ in each direction $h \in X$. If, in addition, f is (locally) Lipschitz continuous, then f is Hadamard differentiable at x.*

Proof. Consider arbitrary sequences $(h_n)_{n\in\mathbb{N}} \subset X$ and $(t_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ with $h_n \to h$ and $t_n \to 0$ as $n \to \infty$. We denote f's Lipschitz constant at x by $L_{f,x} > 0$. Then, for $n \in \mathbb{N}$ large enough, it holds that

$$\begin{aligned}
&\left\| \frac{f(x+t_n h_n) - f(x)}{t_n} - f'(x;h) \right\|_Y \\
&\quad \leq \left\| \frac{f(x+t_n h_n) - f(x+t_n h)}{t_n} \right\|_Y + \left\| \frac{f(x+t_n h) - f(x)}{t_n} - f'(x;h) \right\|_Y \\
&\quad \leq L_{f,x} \|h_n - h\|_X + \left\| \frac{f(x+t_n h) - f(x)}{t_n} - f'(x;h) \right\|_Y \xrightarrow[n\to\infty]{} 0.
\end{aligned}$$

Thus, f is Hadamard differentiable at x. □

Bibliography

[1] Robert A. Adams and John J.F. Fournier, eds. *Sobolev Spaces*. Second Edition. Vol. 140. Pure and Applied Mathematics. Elsevier, 2003. ISBN: 978-0-12-044143-3. DOI: https://doi.org/10.1016/S0079-8169(03)80012-0. URL: https://www.sciencedirect.com/science/article/pii/S0079816903800120.

[2] Hans Wilhelm Alt. *Linear Functional Analysis. An Application-Oriented Introduction*. Universitext. Springer, London, 2016. ISBN: 978-1-4471-7280-2. DOI: https://doi.org/10.1007/978-1-4471-7280-2.

[3] Luigi Ambrosio and Nicola Gigli. "A user's guide to optimal transport". In: *Modelling and Optimisation of Flows on Networks* (2013), pp. 1–155.

[4] Adil Bagirov, Napsu Karmitsa, and Marko M. Mäkelä. *Introduction to Nonsmooth Optimization: theory, practice and software. Theory, Practice and Software*. Vol. 12. Springer Cham, 2014. ISBN: 978-3-319-08114-4. DOI: https://doi.org/10.1007/978-3-319-08114-4.

[5] Adil M. Bagirov, L. Jin, Napsu Karmitsa, Alia Al Nuaimat, and Napsu Sultanova. "Subgradient method for nonconvex nonsmooth optimization". In: *Journal of optimization theory and applications* 157 (2013), pp. 416–435.

[6] Martin Beckmann. "A Continuous Model of Transportation". In: *Econometrica* 20.4 (1952), pp. 643–660. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1907646 (visited on 10/09/2023).

[7] Jean-David Benamou and Yann Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem". In: *Numerische Mathematik* 84.3 (2000), pp. 375–393. URL: https://doi.org/10.1007/s002110050002.

[8] Berkin Bilgic, Itthi Chatnuntawech, Audrey P. Fan, Kawin Setsompop, Stephen F. Cauley, Lawrence L. Wald, and Elfar Adalsteinsson. "Fast image reconstruction with L2-regularization". In: *Journal of magnetic resonance imaging* 40.1 (2014), pp. 181–191.

[9] Vladimir I. Bogachev. *Measure Theory. Volume I*. Springer, Berlin, Heidelberg, 2007. ISBN: 978-3-540-34513-8.

[10] Vladimir I. Bogachev and Svetlana Popova. *Optimal transportation of measures with a parameter*. 2021. DOI: 10.48550/ARXIV.2111.13014. URL: https://arxiv.org/abs/2111.13014.

[11] Nicolas Bonneel and Julie Digne. "A survey of optimal transport for computer graphics and computer vision". In: *Computer Graphics Forum*. Vol. 42. 2. Wiley Online Library. 2023, pp. 439–460.

[12] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. DOI: 10.1017/CBO9780511804441.

[13] Kristian Bredies and Hanna K. Pikkarainen. "Inverse problems in spaces of measures". In: *ESAIM: Control, Optimisation and Calculus of Variations* 19.1 (2013), pp. 190–218.

[14] James Burke, Adrian Lewis, and Michael Overton. "A Robust Gradient Sampling Algorithm for Nonsmooth, Nonconvex Optimization". In: *SIAM Journal on Optimization* 15 (2005), pp. 751–779. DOI: 10.1137/030601296.

[15] Giuseppe Buttazzo and Filippo Santambrogio. "A model for the optimal planning of an urban area". In: *SIAM Journal on Mathematical Analysis* 37.2 (2005), pp. 514–530.

[16] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. "Convergence of entropic schemes for optimal transport and gradient flows". In: *SIAM Journal on Mathematical Analysis* 49.2 (2017), pp. 1385–1418.

[17] Eduardo Casas, Christian Clason, and Karl Kunisch. "Approximation of elliptic control problems in measure spaces with sparse solutions". In: *SIAM Journal on Control and Optimization* 50.4 (2012), pp. 1735–1752.

[18] Eduardo Casas and Karl Kunisch. "Optimal control of semilinear elliptic equations in measure spaces". In: *SIAM Journal on Control and Optimization* 52.1 (2014), pp. 339–364.

[19] Eduardo Casas and Karl Kunisch. "Optimal control of the two-dimensional stationary Navier–Stokes equations with measure valued controls". In: *SIAM Journal on Control and Optimization* 57.2 (2019), pp. 1328–1354.

[20] Yunmei Chen and Xiaojing Ye. "Projection onto a simplex". In: *arXiv preprint arXiv:1101.6081* (2011).

[21] Constantin Christof, Juan C. De los Reyes, and Christian Meyer. "A nonsmooth trust-region method for locally Lipschitz functions with application to optimization problems constrained by variational inequalities". In: *SIAM Journal on Optimization* 30.3 (2020), pp. 2163–2196.

[22] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM: Society for Industrial and Applied Mathematics, 1990. DOI: https://doi.org/10.1137/1.9781611971309.

[23] Christian Clason and Karl Kunisch. "A measure space approach to optimal source placement". In: *Computational optimization and applications* 53 (2012), pp. 155–171.

[24] Christian Clason, Dirk A. Lorenz, Hinrich Mahler, and Benedikt Wirth. "Entropic regularization of continuous optimal transport problems". In: *Journal of Mathematical Analysis and Applications* 494.1 (2021).

[25] Andrew R. Conn, Nicholas I.M. Gould, and Philippe L. Toint. *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000. DOI: 10.1137/1.9780898719857.

[26] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. "Optimal Transport for Domain Adaptation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9 (2017), pp. 1853–1865. DOI: 10.1109/TPAMI.2016.2615921.

[27] Ying Cui and Jong-Shi Pang. *Modern Nonconvex Nondifferentiable Optimization*. Society for Industrial and Applied Mathematics, 2021. DOI: 10.1137/1.9781611976748.

[28] Marco Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in neural information processing systems* 26 (2013).

[29] Marco Cuturi and Arnaud Doucet. "Fast computation of Wasserstein barycenters". In: *International conference on machine learning*. PMLR. 2014, pp. 685–693.

[30] Vincent Duval and Gabriel Peyré. "Exact support recovery for sparse spikes deconvolution". In: *Foundations of Computational Mathematics* 15.5 (2015), pp. 1315–1355.

[31] Jürgen Elstrodt. *Maß- und Integrationstheorie*. Springer Spektrum, Berlin, Heidelberg, 2018. ISBN: 978-3-662-57939-8. DOI: https://doi.org/10.1007/978-3-662-57939-8.

[32] Björn Engquist, Brittany D. Froese, and Yunan Yang. "Optimal transport for seismic full waveform inversion". In: *Communications in Mathematical Sciences* 14.8 (2016), pp. 2309–2330.

[33] Björn Engquist, Kui Ren, and Yunan Yang. "The quadratic Wasserstein metric for inverse data matching". In: *Inverse Problems* 36.5 (2020).

[34] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research and Financial Engineering. Springer New York, NY, 2007. ISBN: 978-0-387-21814-4. DOI: https://doi.org/10.1007/b97543.

[35] Charles Feinstein and Shmuel S. Oren. "A Newton-Type Algorithm for the Solution of the Implicit Programming Problem". In: *Mathematics of Operations Research - MOR* 9 (1984), pp. 75–86. DOI: 10.1287/moor.9.1.75.

[36] Charles D. Feinstein and David G. Luenberger. "Analysis of the Asymptotic Behavior of Optimal Control Trajectories: The Implicit Programming Problem". In: *SIAM Journal on Control and Optimization* 19.5 (1981), pp. 561–585. DOI: 10.1137/0319035. eprint: https://doi.org/10.1137/0319035. URL: https://doi.org/10.1137/0319035.

[37] Lester R. Ford and Delbert R. Fulkerson. "A Simple Algorithm for Finding Maximal Network Flows and an Application to the Hitchcock Problem". In: *Canadian Journal of Mathematics* 9 (1957), pp. 210–218. DOI: 10.4153/CJM-1957-024-0.

[38] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A. Poggio. "Learning with a Wasserstein loss". In: *Advances in neural information processing systems* 28 (2015).

[39] Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, 2018.

[40] Thomas O. Gallouët and Quentin Mérigot. "A Lagrangian Scheme á la Brenier for the Incompressible Euler Equations". In: *Foundations of Computational Mathematics* 18.4 (2018), pp. 835–865.

[41] Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. "Robust hypothesis testing using Wasserstein uncertainty sets". In: *Advances in Neural Information Processing Systems* 31 (2018).

[42] Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains*. Society for Industrial and Applied Mathematics, 2011. DOI: 10.1137/1.9781611972030.

[43] Lukas Hertlein and Michael Ulbrich. "An inexact bundle algorithm for nonconvex nonsmooth minimization in Hilbert space". In: *SIAM Journal on Control and Optimization* 57.5 (2019), pp. 3137–3165.

[44] Sebastian Hillbrecht, Paul Manns, and Christian Meyer. "Bilevel Optimization of the Kantorovich Problem and its Quadratic Regularization Part II: Convergence Analysis". In: *Applied Mathematics and Optimization* (2024). In press.

[45] Sebastian Hillbrecht and Christian Meyer. "Bilevel Optimization of the Kantorovich Problem and its Quadratic Regularization Part I: Existence Results". In: *Applied Mathematics and Optimization* (2024). In press.

[46] Michael Hintermüller and Thomas M. Surowiec. "A bundle-free implicit programming approach for a class of elliptic MPECs in function space". In: *Mathematical Programming* 160 (2016). DOI: 10.1007/s10107-016-0983-9.

[47] Frank L. Hitchcock. "The distribution of a product from several sources to numerous localities". In: *Journal of Mathematics and Physics. Massachusetts Institute of Technology* (20 1941), pp. 224–230. ISSN: 0097-1421. DOI: 10.1002/sapm1941201224. URL: https://doi.org/10.1002/sapm1941201224.

[48] Lars Hörmander. *The Analysis of Linear Partial Differential Operators I*. Springer-Verlag Berlin Heidelberg, 2003. ISBN: 978-3-642-61497-2. DOI: https://doi.org/10.1007/978-3-642-61497-2_5.

[49] Leonid V. Kantorovich. "On the translocation of masses". In: *Doklady Akademii Nauk SSSR* (37 1942). English translation in J. Math. Sci. 133, 4 (2006),1381–1382, pp. 199–201.

[50] Erwin Kreyszig. *Introductory functional analysis with applications*. John Wiley & Sons, 1991. ISBN: 978-0-471-50731-4.

[51] Dirk Lorenz and Hinrich Mahler. "Orlicz space regularization of continuous optimal transport problems". In: *Applied Mathematics & Optimization* 85.2 (2022), p. 14.

[52] Dirk A. Lorenz, Paul Manns, and Christian Meyer. "Quadratically Regularized Optimal Transport". In: *Applied Mathematics & Optimization* (83 2019), 1919–1949 (2021). DOI: https://doi.org/10.1007/s00245-019-09614-w.

[53] David G. Luenberger and Yinyu Ye. *Linear and nonlinear programming*. Vol. 2. Springer, 1984.

[54] Zhi-Quan Luo, Jong-Shi Pang, and Daniel Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, 1996. DOI: 10.1017/CBO9780511983658.

[55] Hinrich Mahler. "Bilevel Optimal Transport Problems: Existence, Regularization and Convergence". PhD thesis. TU Braunschweig, 2022.

[56] Ludovic Métivier, Romain Brossier, Quentin Mérigot, Edouard Oudet, and Jean Virieux. "An optimal transport approach for seismic tomography: application to 3D full waveform inversion". In: *Inverse Problems* 32.11 (2016). DOI: 10.1088/0266-5611/32/11/115008.

[57] Gaspard Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Mem. Math. Phys. Acad. Royale Sci.* (1781), pp. 666–704.

[58] Constantin P. Niculescu and Lars-Erik Persson. *Convex Functions and Their Applications: A Contemporary Approach*. CMS Books in Mathematics. Springer International Publishing, 2018. ISBN: 9783319783376. URL: https://books.google.de/books?id=SnZfDwAAQBAJ.

[59] Jiři Outrata and Michal Cervinka. "On the implicit programming approach in a class of mathematical programs with equilibrium constraints". In: *Control and Cybernetics* 38 (2009).

[60] Jiři Outrata, Michal Kočvara, and Jochem Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*. Vol. 28. Springer Science & Business Media, 1998.

[61] Gabriel Peyré and Marco Cuturi. "Computational optimal transport". In: *Center for Research in Economics and Statistics Working Papers* 86 (2017).

[62] Konstantin Pieper and Boris Vexler. "A Priori Error Analysis for Discretization of Sparse Elliptic Optimal Control Problems in Measure Space". In: *SIAM Journal on Control and Optimization* 51.4 (2013), pp. 2788–2808. DOI: 10.1137/120889137.

[63] Frank Pörner and Daniel Wachsmuth. "Tikhonov regularization of optimal control problems governed by semi-linear partial differential equations". In: *Mathematical Control and Related Fields* 8.1 (2018), pp. 315–335.

[64] Liqun Qi and Jie Sun. "A trust region algorithm for minimization of locally Lipschitzian functions". In: *Mathematical Programming* 66.1 (1994), pp. 25–43. DOI: 10.1007/BF01581136.

[65] Antoine Rolet, Marco Cuturi, and Gabriel Peyré. "Fast dictionary learning with a smoothed Wasserstein loss". In: *Artificial Intelligence and Statistics*. PMLR. 2016, pp. 630–638.

[66] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. "The earth mover's distance as a metric for image retrieval". In: *International journal of computer vision* 40.2 (2000), p. 99.

[67] Michael Růžička. *Nichtlineare Funktionalanalysis. Eine Einführung*. Masterclass. Springer Berlin, 2004. ISBN: 978-3-540-35022-4.

[68] Filippo Santambrogio. "Optimal transport for applied mathematicians". In: *Birkäuser, NY* 55 (2015).

[69] René L. Schilling. *Measures, Integrals and Martingales*. Cambridge University Press, 2005. DOI: 10.1017/CBO9780511810886.

[70] Alexander Shapiro. "On concepts of directional differentiability". In: *Journal of optimization theory and applications* 66 (1990), pp. 477–487.

[71] Richard Sinkhorn. "A relationship between arbitrary positive matrices and doubly stochastic matrices". In: *The annals of mathematical statistics* 35.2 (1964), pp. 876–879.

[72] Justin Solomo, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains". In: *ACM Transactions on Graphics (ToG)* 34.4 (2015), pp. 1–11.

[73] Andrew M. Stuart and Marie-Therese Wolfram. "Inverse optimal transport". In: *SIAM Journal on Applied Mathematics* 80.1 (2020), pp. 599–619.

[74] Michael Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. Society for Industrial and Applied Mathematics (SIAM), 2011. DOI: 10.1137/1.9781611970692.

[75] Cédric Villani. *Optimal transport, old and new*. Springer, 2008.

[76] Cédric Villani. *Topics in optimal transportation*. Vol. 58. American Mathematical Soc., 2021. ISBN: 978-1-4704-6726-5.

[77] Daniel Wachsmuth and Gerd Wachsmuth. "Regularization error estimates and discrepancy principle for optimal control problems with inequality constraints". In: *Control and Cybernetics* 40.4 (2011), pp. 1125–1158.

[78] Wei Wang, Dejan Slepčev, Saurav Basu, John A. Ozolek, and Gustavo K. Rohde. "A linear optimal transportation framework for quantifying and visualizing variations in sets of images". In: *International journal of computer vision* 101 (2013), pp. 254–269.

[79] Ralph A. Willoughby. "Solutions of Ill-Posed Problems (A. N. Tikhonov and V. Y. Arsenin)". In: *SIAM Review* 21.2 (1979), pp. 266–267. DOI: 10.1137/1021044.

Index

Active set, 91
Adjoint state, 122
Assignment function, 71

Ball
    Closed ..., 9
    Open ..., 9
Biactive set, 91
Bilevel Hitchcock problem, 61
    Reduced ..., 105
    Regularized ..., 68
    Twice regularized ..., 104
Bilevel Kantorovich problem, 15
    Regularized ..., 22, 34
    Relaxing ..., 49
        Regularized ..., 47
Bouligand subdifferential, 95
    Collective ..., 119
Bounded on bounded sets, 15, 61

Cauchy decrease condition
    Constrained ..., 115
        Modified ..., 115
    Generalized
        Modified ..., 110
    Generalized ..., 110
Characteristic matrix, 93
Convolution, xiii, 20
Cost function, 15
    Set of ..., 23
Cost matrix, 60
Coupling, 14, 60
    Set of ..., 14

Discrete σ-algebra, xxvi
Distance parameter, 47
Domain, 8, 59
    Compact ..., 13
    Extension ..., 13, 47
    Smoothed ..., 21
Dual variable, 19, 64
    Zero-mean ..., 27

Extension ...
    of marginal, 47
    of target functional, 15
    Operator, 21

Farka's lemma, 74
First-order stationary, 112

Generalized Jacobian, 95

Hitchcock problem, 60
    Regularized ..., 63
        Dual problem, 67

Implicit programming, 85
Inactive set, 91

Jordan decomposition, xxii

Kantorovich problem, 15
    Regularized ..., 18
        Dual problem, 19

Lebesgue null set, 8

Marginal-to-transport-plan
        mapping
    Regularized ..., 99
Marginals, 14, 60
    Compatible ..., 14
        Set of ..., 23, 68
Mask, 100
Model function, 109
Mollifier, 20

Nonnegative part ...
    of function, 19

of matrix, 65
of measure, xxiii
Nonpositive part ...
of function, 19
of matrix, 65
of measure, xxiii
Nonsmooth problem, 109
Constrained ..., 112
Norm
1-..., 7
∞-..., 7
Euclidean ..., 7
Frobenius ..., 7
Spectral ..., 7
Total variation ..., xxvi, 8
Uniform ..., 8

Outer structure, 96
Outer sum ...
of functions, 19
of vectors, 64

Probability measure, 8
Finite moment, 54
Product σ-algebra, xix
Projection map, 14
Prototypical bilevel problem, 2
Pushforward measure, 14

Quality indicator, 110
Modified ..., 111

Recovery sequence, 46, 70
Reduced cost vector, 73
Reduced system matrix, 73
Regular Borel measure, xxvi
Space of ..., 8
Regularization parameter, 17, 64

Scalar product, 9
Frobenius ..., 7
Signed integration, xxiii
Smoothing parameter, 20
Solution operator, 23, 68, 87
Stationarity measure, 113
Modified ..., 113
Support ...
of function, 10
of measure, 10

Tensor product, 26
Tracking-type objective, 53, 122
Transport plan, 14, 60
Set of ..., 14
Transportation identification problem, 123
Trust region subproblem, 110
Constrained ..., 114
Modified ..., 115
Modified ..., 110

Variation measure, xx
Variational inequality, 112
Generalized ..., 112

Wasserstein ...
Distance, 54
Inverse problem, 54
Weak* lower semicontinuity, 15