Optimal Transport and Large-Scale Mapping Estimation

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Peter Bernd Oehme

Supervisor: Professor Dr. Jan-Frederik Pietschmann

15 July 2021

Contents

Introduction

Optimal transport is the theory of moving a unit of mass whilst minimizing the incurring costs associated with the underlying transportation. Over the years different formulations have been posed, all related to each other. With the advance of computing technology, implementations that are based on optimal transport have entered application in various ways. One challenge this sets is the incorporation of large-scale problems, as previous algorithms do not scale well with increasing dimensions. In this thesis one particular approach for solving large-scale optimal transport will be outlined.

The first optimal transport problem was posed in 1781 by Gaspard Monge [Mon81]. Considering a cost function $c : X \times Y \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, its modern formulation for two probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ is

$$\inf \left\{ \int_X c(x, T(x)) \, \mathrm{d}\mu(x) : T \in \mathfrak{M}(X, Y), T\#\mu = \nu \right\},$$

where $\mathfrak{M}(X, Y)$ is the set of all measurable functions from X to Y, and $T\#\mu$ describes the push-forward of μ under some T. It suffers from the problem of not allowing mass splitting, limiting its usage. In 1942, Leonid Kantorovich [Kan42] put forward a relaxed form of the original problem, allowing for the application of primal-dual optimization theory. Instead of directly optimizing over *transport maps* between two spaces, the objective now takes a joint measure, which attains the starting measures as its marginals, and assigning a cost to the *transport plan* by integrating the underlying cost function with respect to it:

$$\inf \left\{ \int_{X \times Y} c(x, y) \, \mathrm{d}\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\}.$$

Here, $\Pi(\mu, \nu)$ is the set of all joint measures on $X \times Y$ that have μ and ν as marginals. To aid in the approximation of optimal transport, entropic

regularization may be performed, resulting in

$$\inf\left\{\int_{X\times Y} c\,\mathrm{d}\gamma + \int_{X\times Y} \gamma(x,y)\big(\log(\gamma(x,y))-1\big)\,\mathrm{d}(x,y) : \gamma\in\Pi(\mu,\nu)\right\}.$$

Recently, optimal transport has been used to advance areas of applied mathematics like machine learning, image processing and auction processes. In these applications it is often impractical to calculate an exact solution to the optimal transport problem, instead an approximated solution is used. The main tool for this is the so-called *Sinkhorn algorithm*, which iteratively improves dual variables of the optimal transport problem and solves the entropically regularized transport problem. An approximation with near-linear time cost of this algorithm has been achieved by [AWR18]. In contrast to this traditional approach, [GCPB16] proposed methods expanding the application of optimal transport approximation to samples of large scales by using stochastic optimization. Building on this work, [SDF$^+$18] have applied a stochastic gradient descent method to describe an approach that works well with large-scale problems. They reduced the complexity to a constant cost by training neural networks on samples of fixed size to estimate an optimal transport plan, and approximate an optimal transport map from the plan via the *barycentric projection*. In particular, this approach lends itself to the handling of continuous problems, in contrast to previous restrictions on discrete formulations of regularized transport, see e.g. [FPPA13]. The estimation of optimal transport plans and maps using this method is the main objective described here.

This thesis consists of two main parts. The first part will function as a general introduction to optimal transport and provide related results. In Section 1.1 the Monge formulation of optimal transport is stated, while in Section 1.2 the Kantorovich relaxation is introduced and results about the feasibility of the relaxed problem are given. In Section 1.3 duality theory is used to derive a dual problem, give conditions on the existence of a dual solution, and outline presumptions necessary for strong duality to hold.

The second part will build upon the given theory and introduce a method of estimating solutions of the large-scale optimal transport problem as given by [SDF$^+$18]. To achieve this, Section 2.1 introduces regularization of the original transport problem, Section 2.2 gives a tool to obtain transport maps from transport plans, and Section 2.3 highlights the regularized dual problem. The Sinkhorn algorithm is then used as a starting point to expand onto the estimation of optimal transport plans and maps via stochastic gradient descent in Section 2.4. References to a number of applications in theoretical and applied mathematics will be provided at the end.

Notation

Throughout this thesis we will be using the following notation:

- σ-algebras will not be explicitly stated for measures and measurable spaces. Measurable functions can thus be written as $f : X \to Y$ instead of $f : (X, \mathcal{X}) \to (Y, \mathcal{Y})$.
- The notion of weak convergence of measures ($\mu_n \rightharpoonup \mu$) will be used instead of weak-* convergence.
- $\mathfrak{M}(X, Y)$: The set of all measurable functions $f : X \to Y$.
- $\mathcal{M}(X)$: The set of all finite measures on the underlying set X.
- $\mathcal{P}(X)$: The set of all probability measures on X.
- $\Pi(\mu, \nu)$: The set of all transport plans from μ to ν.
- $T\#\mu$: The push-forward of the measure μ. Sometimes also called the image measure.
- $C(X)$: The set of all continuous functions $f : X \to \mathbb{R}$.
- $C_b(X)$: The set of all bounded $f \in C(X)$.
- $\overline{\mathbb{R}}$: The set $\mathbb{R} \cup \{-\infty, \infty\}$.

1 Optimal Transport

This introductory part will follow [San15, Chapter 1] closely in content as well as notation. It will contain the Monge and Kantorovich formulations of the optimal transport problem, as well as a dual problem for the Kantorovich problem. Afterwards, we will introduce regularized optimal transport and work towards mapping estimation in the second part.

1.1 The Monge Problem

There are two important paradigms in the theory of optimal transport. The original formulation is the so-called Monge Problem. It was later relaxed by Kantorovich to a more general problem which is easier to handle. To motivate the modern formulation of the Monge problem, which was first described by [Mon81] in 1781, we consider a mass distribution modelled by a measure on X to be moved to another mass distribution, again modelled by a measure, on Y. One may for example imagine a pile of sand or a number of particles being moved from one location to another. This transport will incur costs and the goal of optimal transport is to find a "way of transporting" the mass, which accumulates the smallest cost possible while making sure that no mass is lost.

For this purpose, we consider a non-negative *cost function* $c(x, y)$, which will give us the cost of moving a particle from position x to position y. Ideally this cost function will be continuous or semi-continuous, though later on we will see that we can obtain optimality and uniqueness in a simpler fashion by making use of more constraints on c.

To attain the Monge formulation, the transport is modelled via the *push-forward measure*, or *image measure*, of our original measure on X.

Definition 1.1 (Push-Forward; adapted from [Bog07, Section 9.1])**.** Given a measure $\mu \in \mathcal{M}(X)$ and a measurable map $T : X \to Y$, we define the **push-forward measure** of μ under T as a measure on Y by

$$(T\#\mu)(y) := \mu\big(T^{-1}(y)\big).$$

A very useful consequence of this definition is that we can use it to state the conservation of mass throughout the transport. Take a starting measure μ on X and a target measure ν on Y. In order to not lose any mass, it is necessary for

$$T\#\mu = \nu$$

to hold. This means that all elements of the underlying σ-algebra $\mathcal{Y}$ on Y coincide under the measures ν and $T\#\mu$, i.e.

$$\forall A \in \mathcal{Y} : (T\#\mu)(A) = \nu(A).$$

Another useful step to take is to normalize the overall mass. This can be achieved by considering only probability measures on X and Y.

We can now state the Monge problem in its modern form:

Definition 1.2 (Monge Problem; adapted from [San15, Problem 1.1])**.** Let $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ be our starting and target measures, and consider $c : X \times Y \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ to be a cost function. Then we define the **Monge Problem** as

$$\inf \left\{ M(T) := \int_X c\big(x, T(x)\big) \, \mathrm{d}\mu(x) : T \in \mathfrak{M}(X, Y), T\#\mu = \nu \right\}.$$

From here on, we will be using the abbreviation (MP) as the name of the Monge Problem, and $\inf(\mathrm{MP})$ to denote the attained infimum. A function $T \in \mathfrak{M}(X, Y,)$ is called an **optimal transport map**, if it satisfies

$$\int_X c\big(x, T(x)\big) \, \mathrm{d}\mu(x) = \inf(\mathrm{MP}).$$

We can already provide a couple of results about the general existence of transport maps — though these must not be optimal for (MP). The most prevalent restriction on our starting measures μ regards the existence of so-called *atoms*.

Definition 1.3 (Atoms; adapted from [Bog07, Volume 1, Definition 1.12.7])**.** Let $\mu \in \mathcal{M}(X)$ be a measure and $\mathcal{X}$ be a σ-algebra on X. A set $A \in \mathcal{X}$ is called an **atom** of μ if $\forall B \in \mathcal{X}, B \subseteq A$: either $\mu(B) = 0$ or $\mu(B) = \mu(A)$ holds. If μ has no atoms, we say it is **atomless**.

When there are no atoms for μ and the underlying measurable space is compact, we can first achieve existence of transport maps on the real line.

Lemma 1.4 (Adapted from [San15, Lemma 1.27]). *Let $\mu, \nu \in \mathcal{P}(\Omega)$ be two measures with μ being atomless, and $\Omega \subseteq \mathbb{R}$ being compact. Then, there exists a transport map $T : \Omega \to \Omega$ such that $T\#\mu = \nu$.*

Proof. Cf. the proof of [San15, Lemma 1.27]. □

By applying the injective map $\sigma_d : \mathbb{R}^d \to \mathbb{R}$ from [San15, Lemma 1.29], we can extend Lemma 1.4 onto compact subsets of $\mathbb{R}^d$.

Theorem 1.5 (Adapted from [San15, Corollary 1.29]). *Let $\mu, \nu \in \mathcal{P}(\Omega)$ be two measures with μ being atomless, and $\Omega \subseteq \mathbb{R}^d$ being compact. Then, there exists at least a transport map $T : \Omega \to \Omega$ such that $T\#\mu = \nu$.*

Proof. Cf. the proof of [San15, Corollary 1.29]. □

In general, (MP) is not easy to solve. In order to point out the difficulties we may face, we require the notion of *weak convergence* of a measure.

Definition 1.6 (Weak Convergence of Measures; adapted from [Bog07, Volume 2, Definition 8.1.1]). Let X be a measurable space and $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of measures which satisfies $\forall n \in \mathbb{N} : \mu_n \in \mathcal{M}(X)$. This sequence is said to **converge weakly to a measure** $\mu \in \mathcal{M}(X)$, if for all $f \in C_b(X)$

$$\lim_{n\in\mathbb{N}} \int_X f(x)\ \mathrm{d}\mu_n(x) = \int_X f(x)\ \mathrm{d}\mu(x)$$

holds. In this case we will write $\mu_n \rightharpoonup \mu$ for $n \to \infty$.

Remark. Technically, the space of radon measures is the dual of the space of continuous functions, see e.g. [Kap57, Section 4]. Following this duality, one can define weak-* convergence for locally finite non-negative measures in accordance with [Cam08, Definition 2.4]. In Definition 1.6, the observed measures are globally finite, meaning that it should instead refer to the *weak-* convergence* of these measures. It is however not uncommon to see the usage of weak convergence in relevant literature, e.g. [Bog07, Volume 2, Definition 8.1.1]. Due to this, we will continue to use this term, while at the same time acknowledging that it technically is not correct.

We can now consider the following two points to illustrate the difficulties in dealing with (MP):

Firstly, if one of the two measures were to be discrete, for example of the form $\mu = \delta_{x_0}, x_0 \in X$, and the other is not, (MP) would turn infeasible. In order for T to be a transport map for (MP), no *mass splitting* must occur. With a semi-discrete transport problem, this is not possible.

Secondly, as [San15] pointed out, the constraint on T in (MP) is not closed under weak convergence (cf. Definition 1.6). Consider the sequence $T_n(x) := \sin(nx)$ on $[0, 2\pi]$ for $n \in \mathbb{N}$. We get $(T_n \# \mu)(y) = \mu\left(\frac{\arcsin(y)}{n}\right)$ and notice that for all $\varepsilon > 0$ there exists an $n \in \mathbb{N}$ such that the inner term on the right-hand side is less than ε for all $y \in Y$. If $T_n \# \mu$ is weakly convergent, it can thus only converge to $\nu \equiv 0$, i.e. for $n \in \mathbb{N}$ sufficiently large, T_n no longer induces a probability measure via the push-forward.

Both of these problems can be avoided by considering a different way of formulating the transport constraint. We will later see that this relaxed description indeed includes (MP).

1.2 The Kantorovich Relaxation

In 1942, [Kan42] gave a more general formulation for the optimal transport problem. Instead of formulating the transport via the push-forward under a transport map, Kantorovich used so-called *transport plans*, where the original measures appeared as marginals of a joint measure. This, as [San15] puts it, allowed for a description of how many particles are moved from x to y instead of specifying the destination of x under T.

Definition 1.7 (Transport Plans; adapted from [San15, Problem 1.2])**.** For two measures $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ we define the set of all **transport plans** from μ to ν as

$$\Pi(\mu, \nu) := \Big\{\gamma \in \mathcal{P}(X \times Y) : (\pi_X)\#\gamma = \mu, (\pi_Y)\#\gamma = \nu\Big\},$$

where π_X and π_Y are the canonical projections onto X and Y, i.e.

$$\pi_X(x, y) := x, \text{ and } \pi_Y(x, y) := y.$$

Using this terminology we can now describe the Kantorovich Problem as a more general problem when compared to (MP).

Definition 1.8 (Kantorovich Problem; adapted from [San15, Problem 1.2])**.** Let $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ be our starting measures, and consider the function $c : X \times Y \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ to be a cost function. Then we call the following optimization problem the **Kantorovich Problem**:

$$\inf \left\{ K(\gamma) := \int_{X \times Y} c(x, y) \, \mathrm{d}\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\}.$$

In analogy to Definition 1.2, we will be using the abbreviation (KP) to describe the problem itself and denominate the attained infimum as inf (KP). We shall further define **optimal transport plans** as those specific transport plans $\gamma \in \Pi(\mu, \nu)$ that satisfy

$$\int_{X \times Y} c(x, y) \, \mathrm{d}\gamma(x, y) = \inf (\mathrm{KP}).$$

One direct benefit of this formulation is that we are no longer bound by the prohibition of mass splitting like in (MP). We can however recover the (MP) formulation via (KP) if $\gamma = (id, T)\#\mu$ is an optimal transport plan for $T \in \mathfrak{M}(X, Y)$, in which case T would once again be an optimal transport map. If given two measures $\mu \in \mathcal{M}(X), \nu \in \mathcal{M}(Y)$ and a transport map $T \in \mathfrak{M}(X, Y)$ with $T\#\mu = \nu$, we can always induce a transport plan from T by defining

$$\gamma_T := (id, T)\#\mu. \tag{1.1}$$

We are now interested in the relation between (KP) and (MP). First we define

$$J(\gamma) := \begin{cases} K(\gamma) = M(T), & \gamma = \gamma_T \\ \infty, & \text{otherwise} \end{cases}.$$

The functional J immediately allows us to consider (MP) on the same set of admissible objects as (KP), that is all transport plans induced by a transport map. As Kantorovich replaced J with K the question to be asked changes to: does $\inf K = \inf J$ hold?

To answer this question, it can be shown that the set of plans induced by a map is dense in $\Pi(\mu, \nu)$, whenever certain conditions are met. This will then allow the conclusion that, indeed, both infima coincide. For these results, we will always require compactness for a subset $\Omega \subseteq \mathbb{R}^d$, however [San15] points out that this is just for simplicity and more general statements can be made.

Theorem 1.9 (Plans Induced by Maps are Dense; taken from [San15, Theorem 1.32])**.** *Let $\Omega \subseteq \mathbb{R}^d$ be a compact subset, and $\mu, \nu \in \mathcal{P}(\Omega)$ be two probability measures. If μ is atomless, then the set of all transport plans γ_T induced by a transport map T as defined in Equation 1.1 is dense in the set of all transport plans $\Pi(\mu, \nu)$.*

Proof. Cf. the proof of [San15, Theorem 1.32]. □

Before we can prove strong duality for continuous cost functions and atomless μ, we need to define the term *lower semi-continuous*. This definition will then be used to consider the *relaxation* of the functional J in the following proof.

Definition 1.10 (Lower Semi-Continuous Functions; adapted from [Kes09, Definition 5.1.2])**.** Let X be an arbitrary topological space. Then a function $f : X \to \mathbb{R}$ is called **lower semi-continuous** (l.s.c.), if the set

$$\left\{x \in X : f(x) \leq \alpha\right\}$$

is closed in X for all $\alpha \in \mathbb{R}$. In particular, if $f : X \to \mathbb{R}$ is continuous, it is also l.s.c.

Theorem 1.11 (Infima of J and K coincide; adapted from [San15, Theorem 1.33])**.** *Let $\Omega \subseteq \mathbb{R}^d$ be a compact subset, $\mu, \nu \in \mathcal{P}(\Omega)$ be two measures on Ω with μ being atomless, and $c : \Omega \times \Omega \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ be a continuous cost function. Then* $\inf J = \inf K$ *holds, with the infimum being taken over* $\Pi(\mu, \nu)$.

Proof. This proof was adapted from [San15, the proof of Theorem 1.33 and the memo on relaxations, Box 1.10].

We consider the relaxation $\overline{J}$ of J defined by

$$\forall \gamma \in \Pi(\mu, \nu) : \overline{J}(\gamma) := \inf \left\{ \liminf_{n \to \infty} J(\gamma_n) : \gamma_n \rightharpoonup \gamma, n \to \infty \right\}.$$

Since $J \geq \overline{J}$ per definition implies $\inf J \geq \inf \overline{J}$, and as $J \geq \inf J$ results in $\overline{J} \geq \inf J$, because $\inf J$ is constant and thus l.s.c., we get that $\inf J = \inf \overline{J}$. As K is also l.s.c. with $K \leq J$ we obtain $\inf K \leq \inf J$.

To show $\inf K = \inf J$ we need to find a sequence of transport maps $(T_n)_{n \in \mathbb{N}}, T_n \in \mathfrak{M}(\mathbb{R}^d, \mathbb{R}^d)$ with $(T_n)\#\mu = \nu$ for every $\gamma \in \Pi(\mu, \nu)$, such that $\gamma_{T_n} \rightharpoonup \gamma$ and $J(\gamma_{T_n}) \to K(\gamma)$ for $n \to \infty$. Because K is continuous and for $\gamma = \gamma_{T_n}$ it holds that $K(\gamma) = J(\gamma)$, it is already sufficient for a sequence $(T_n)_{n \in \mathbb{N}}$ to exist with $\gamma_{T_n} \rightharpoonup \gamma$ for $n \to \infty$.

Such a sequence exists, since the set of transport plans induced by a transport map is dense in $\Pi(\mu, \nu)$ by Theorem 1.9. With this minimizing sequence we finally have $\inf K = \inf \overline{J} = \inf J$. □

So far we have avoided the question, if (KP) even admits a solution. We will now give a basic result for continuous cost functions. One of the essential ideas here is the continuity of the functional K. For this, an explicit notion of convergence is needed. We will be referring to the continuity of K in terms of the weak convergence of measures (cf. Definition 1.6).

Lemma 1.12 (Taken from [San15, Theorem 1.4])**.** *Let X and Y be compact metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ be probability measures, and consider $c : X \times Y \to \mathbb{R}$ to be a continuous function. Then* (KP) *admits a solution.*

Proof. This proof was adapted from the proof of [San15, Theorem 1.4]. It is sufficient to show that $\Pi(\mu,\nu)$ is compact and $\gamma \mapsto K(\gamma)$ is continuous, as any continuous function attains its minimum over a compact set.

To show the continuity, let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence of probability measures over $X \times Y$ converging weakly to $\gamma \in \mathcal{P}(X \times Y)$. As c is continuous and especially measurable, we have

$$\lim_{n\in\mathbb{N}} K(\gamma_n) = \lim_{n\in\mathbb{N}} \int_{X\times Y} c \,\mathrm{d}\gamma_n = \int_{X\times Y} c \,\mathrm{d}\gamma = K(\gamma).$$

To show the compactness, let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence with $\gamma_n \in \Pi(\mu,\nu)$ for every $n \in \mathbb{N}$. Let $f \in C_b(X \times Y)$. We define $\hat{f}_n := \int f \,\mathrm{d}\gamma_n$. With $|\int f \,\mathrm{d}\gamma_n| \leq \int |f| \,\mathrm{d}\gamma_n \leq \sup_{X\times Y} |f| < \infty$, because f is bounded and continuous on a compact set, we see that the sequence $(\hat{f}_n)_{n\in\mathbb{N}}$ is bounded. Hence, for every $f \in C_b(X \times Y)$ there exists a convergent subsequence $(\hat{f}_{n_k})_{k\in\mathbb{N}}$. □

Remark. Compactness and sequential compactness are the same on metric spaces. To be able to view the set of transport plans as a compact set in a metric space, one may consider the linear space of all bounded and signed measures over $X \times Y$ with the norm $\|\pi\| := |\pi|(X)$. The triangle inequality follows from the Hahn decomposition $X \times Y = A \dot{\cup} B$, such that $(\sigma+\tau)(A) \leq 0$, and $(\sigma+\tau)(B) \geq 0$ (cf. [Bog07, Volume 1, Theorem 3.1.1 and Corollary 3.1.2]), using

$$\|\sigma+\tau\| = |\sigma+\tau|(X) = (\sigma+\tau)\big((X\times Y)\cap A\big) - (\sigma+\tau)\big((X\times Y)\cap B\big)$$

$$= (\sigma+\tau)(A) - (\sigma+\tau)(B) \leq |\sigma|(A) + |\sigma|(B) + |\tau|(A) + |\tau|(B) = \|\sigma\| + \|\tau\|.$$

As our transport plans are probability measures on $X \times Y$, they are bounded and countably additive. We can thus assume $\Pi(\mu,\nu)$ to be a subset of the metric space of signed measures.

In most cases, our cost functions will be sufficiently continuous in order to apply Lemma 1.12 and guarantee the existence of a solution to (KP). However, it is still possible to generalize to l.s.c. cost functions on compact subspaces, and even l.s.c. cost functions on complete and separable metric spaces (i.e. *Polish spaces*).

Lemma 1.13 (Taken from [San15, Theorem 1.5])**.** *Let X and Y be any compact metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ be probability measures, and finally $c : X \times Y \to \mathbb{R} \cup \{\infty\}$ be an l.s.c. function bounded from below. Then* (KP) *admits a solution.*

Proof. Cf. the proof of [San15, Theorem 1.5]. □

Theorem 1.14 (Adapted from [San15, Theorem 1.7])**.** *Let X and Y be complete and separable metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ be probability measures, and $c : X \times Y \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ be l.s.c. Then* (KP) *admits a solution.*

Proof. This proof was adapted from the proof of [San15, Theorem 1.7]. Our objective is to show that any sequence in $\Pi(\mu, \nu)$ is tight (cf. A.1) and apply the Theorem of Prokhorov (A.1), as we can no longer rely on the finite maximum of continuous functions over our (in Lemma 1.13 still compact) space $X \times Y$.

To this end, let $\varepsilon > 0$. As a sequence made up of a single transport plan γ is necessarily tight, we can find two compact sets $K_X \subseteq X, K_Y \subseteq Y$, such that $\mu(X \setminus K_X) < \frac{\varepsilon}{2}$ and $\nu(Y \setminus K_Y) < \frac{\varepsilon}{2}$ hold true. Hence, the set $K_X \times K_Y \subseteq X \times Y$ is compact too, resulting in

$$\forall \gamma_n \in \Pi(\mu, \nu) : \gamma_n\big((X \times Y) \setminus (K_X \times K_Y)\big)$$

$$\leq \gamma_n\big((X \setminus K_X) \times Y\big) + \gamma_n\big(X \times (Y \setminus K_Y)\big) = \mu(X \setminus K_X) + \nu(Y \setminus K_Y) < \varepsilon.$$

With Prokhorov (A.1), we get $\gamma_n \rightharpoonup \gamma$ for $n \to \infty$. As our transport plan γ was chosen arbitrarily, this property holds for all transport plans and thus all sequences of transport plans. Finally, the claim can be seen using the compactness of $\Pi(\mu, \nu)$ and the continuity of K with regard to the weak convergence of measures (cf. Definition 1.6). □

It is beneficial to investigate the dual problem of (KP), because this dual problem will enable us to consider a regularized dual optimal problem in the second part of the thesis. This regularized dual will ultimately aid in the estimation of optimal transport plans and consequently in the estimation of maps as well.

1.3 Kantorovich Duality

As [San15], Section 1.2, points out, the problem (KP) is a linear optimization problem with convex constraints. Hence, it is plausible to investigate the *dual problem*. Beforehand, we introduce the notion of the *support* of a function.

Definition 1.15 (Support; adapted from [San15, Definition 1.14])**.** Let X be a separable metric space, and $\mu \in \mathcal{M}(X)$ be a measure on X. We define the **support** of μ as $\operatorname{supp}(\mu) := \bigcap \big\{A \subseteq X : A \text{ closed and } \mu(X \setminus A) = 0\big\}$.

To get a dual problem, it is useful to restate the constraint "$\gamma \in \Pi(\mu, \nu)$" in terms of continuous bounded functions. We notice that for a general measure $\gamma \in \mathcal{M}(X \times Y)$ the term

$$\sup_{\varphi \in C_b(X), \psi \in C_b(Y)} \int_X \varphi(x) \, \mathrm{d}\mu(x) + \int_Y \psi(y) \, \mathrm{d}\nu(y) - \int_{X \times Y} \varphi(x) + \psi(y) \, \mathrm{d}\gamma(x, y)$$

equals 0 if $\gamma \in \Pi(\mu, \nu)$, and ∞ if otherwise. This effectively constitutes a restatement of our constraint if we add it to the original formulation of (KP): when the constraint is fulfilled, nothing has changed, and when γ is not a transport plan, (KP) stays infeasible. For ease of notation, we will be using $(\varphi \oplus \psi)(x, y) := \varphi(x) + \psi(y)$. Exchanging the inf and the sup in the modified (KP), we obtain

$$\sup_{\varphi \in C_b(X), \psi \in C_b(Y)} \int_X \varphi \, \mathrm{d}\mu + \int_Y \psi \, \mathrm{d}\nu + \inf_{\gamma \in \mathcal{M}(X \times Y)} \int_{X \times Y} c - \varphi \oplus \psi \, \mathrm{d}\gamma.$$

We moreover want to restate the above infimum in γ as a constraint on φ and ψ. We get

$$\inf_{\gamma \in \mathcal{M}(X \times Y)} \int_{X \times Y} c - \varphi \oplus \psi \, \mathrm{d}\gamma = \begin{cases} 0, & \varphi \oplus \psi \leq c \text{ on } X \times Y \\ -\infty, & \text{else} \end{cases}.$$

This equation holds, [San15] argues, because if $\varphi \oplus \psi > c$ somewhere in $X \times Y$, we can choose a measure γ supported on that region with a mass tending to ∞. Such a γ would leave the expression unbounded from below, i.e. the problem infeasible.

This argument allows us to state the dual problem to the Kantorovich relaxation of (MP). With this problem, we will then investigate the duality between (KP) and its dual.

Definition 1.16 (Kantorovich Dual Problem; adapted from [San15, Problem 1.9])**.** Let $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ be our starting measures, and consider $c : X \times Y \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ to be a cost function. Then we call the following optimization problem the **Kantorovich Dual Problem**:

$$\sup \left\{ \int_X \varphi \, \mathrm{d}\mu + \int_Y \psi \, \mathrm{d}\nu : \varphi \in C_b(X), \psi \in C_b(Y), \varphi \oplus \psi \leq c \right\}.$$

In analogy to Definitions 1.2 and 1.8, we will be using the abbreviation (DP) to describe the problem itself and sup (DP) to denominate its supremum.

With the formulation of (DP) and the previous characterization of the constraint "$\gamma \in \Pi(\mu, \nu)$", it is now possible to reach a statement about strong duality. For all admissible $\gamma \in \Pi(\mu, \nu)$ (i.e. the constraint from (KP)) and $\varphi \in C_b(X), \psi \in C_b(Y)$ with $\varphi \oplus \psi \leq c$ (i.e. the constraint from (DP)) we have

$$\sup_{\varphi \in C_b(X), \psi \in C_b(Y)} \int_X \varphi(x) \, \mathrm{d}\mu(x) + \int_Y \psi(y) \, \mathrm{d}\nu(y)$$

$$= \sup_{\varphi \in C_b(X), \psi \in C_b(Y)} \int_{X \times Y} (\varphi \oplus \psi)(x, y) \, \mathrm{d}\gamma(x, y) \leq \int_{X \times Y} c(x, y) \, \mathrm{d}\gamma(x, y).$$

As the left-hand side of this equality is constant, we can further consider the infimum over all $\gamma \in \Pi(\mu, \nu)$ and obtain $\sup(\mathrm{DP}) \leq \inf(\mathrm{KP})$.

So far, we do not know if the supremum on the left-hand side does exist. To handle this problem, we will be further transforming (DP) using so-called *c-transforms*, or *c-conjugate functions*. With these transformations it will then be possible to formulate (DP) as an optimization problem over just one dual variable.

Definition 1.17 (c-Transform, $\bar{c}$-Transform, c-Concavity, $\bar{c}$-Concavity; taken from [San15, Definition 1.10])**.** Given $\chi : X \to \overline{\mathbb{R}}$ and $\zeta : Y \to \overline{\mathbb{R}}$ two functions, as well as $c : X \times Y \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ a cost function, we define the ***c*-transform** of χ by

$$\chi^c : Y \to \overline{\mathbb{R}}, \; y \mapsto \inf_{x \in X} c(x, y) - \chi(x)$$

and the **$\bar{c}$-transform** of ζ by

$$\zeta^{\bar{c}} : X \to \overline{\mathbb{R}}, \; x \mapsto \inf_{y \in Y} c(x, y) - \zeta(y).$$

We further define the notion for a function $\psi : Y \to \overline{\mathbb{R}}$ to be **$\bar{c}$-concave** if there exists a function $\chi : Y \to \overline{\mathbb{R}}$, such that $\psi = \chi^c$. The set of all $\bar{c}$-concave functions over Y will be denoted by $\bar{c}\text{-}conc(Y)$. Analogously, a function $\varphi : X \to \overline{\mathbb{R}}$ is said to be ***c*-concave** if there exists a function $\zeta : X \to \overline{\mathbb{R}}$, such that $\varphi = \zeta^{\bar{c}}$, and the set of all such c-concave functions over X will accordingly be described by $c\text{-}conc(X)$.

Theorem 1.18 (Adapted from [San15, Proposition 1.11])**.** *Let X and Y be compact metric spaces and $c : X \times Y \to \mathbb{R}_{\geq 0}$ be a continuous cost function. Then there exists a solution $(\varphi, \psi) \in C_b(X) \times C_b(Y)$ of* (DP) *with the form $\varphi \in c\text{-}conc(X), \psi \in \bar{c}\text{-}conc(Y)$ with $\psi = \varphi^c$. In particular, we can restate*

$$\sup(\mathrm{DP}) = \max_{\varphi \in c\text{-}conc(X)} \int_X \varphi \, \mathrm{d}\mu + \int_Y \varphi^c \, \mathrm{d}\nu.$$

Proof. This proof was adapted from the proof of [San15, Proposition 1.11].

We consider a maximizing sequence $(\varphi_n, \psi_n)_{n\in\mathbb{N}}$, with $\varphi_n \oplus \psi_n \leq c$ and $(\varphi_n, \psi_n) \in C_b(X) \times C_b(Y)$ for all $n \in \mathbb{N}$, that is a pair of admissible functions for (DP). By this assumption we have $\forall x \in X, y \in Y$:

$$\psi_n(y) \leq c(x,y) - \varphi_n(x),$$

and note that $\int_Y \psi_n \, \mathrm{d}\nu \leq \int_Y \inf_{x\in X} c(x,y) - \varphi_n(x) \, \mathrm{d}\nu(y) = \int_Y \varphi_n^c \, \mathrm{d}\nu$. Thus it will suffice to consider a maximizing sequence $(\varphi_n, \varphi_n^c)_{n\in\mathbb{N}}$. The same argument applied to φ_n instead of ψ_n results in a sequence $(\psi^{\bar{c}}, \psi_n)_{n\in\mathbb{N}}$. This already means that for an arbitrary pair (φ, ψ) to be a solution, it is necessary that $\varphi^c = \psi$ and $\psi^{\bar{c}} = \varphi$ hold, i.e. that they must be $\bar{c}$-concave and c-concave respectively, since the integrals otherwise could be further enlarged by considering the c- and $\bar{c}$-transforms of said solution.

From the proof of [San15, Proposition 1.34], we have $\varphi^{c\bar{c}} = \varphi, \psi^{\bar{c}c} = \psi$, as

$$\forall x \in X : \eta^{c\bar{c}}(x) = \inf_{y\in Y} c(x,y) - \eta^c(y) = \inf_{y\in Y} c(x,y) - \inf_{z\in X}(c(z,y) - \eta(z))$$

$$\geq \inf_{y\in Y} c(x,y) - c(x,y) + \eta(x) = \eta(x),$$

and for $\zeta^{\bar{c}c} \geq \zeta$ analogously. For η c-concave we hence have

$$\eta^c = \chi^{\bar{c}c} \geq \chi \Rightarrow \eta^{c\bar{c}}(x) = \inf_{y\in Y} c(x,y) - \eta^c(y) \leq \chi^{\bar{c}}(x) = \eta(x) \Rightarrow \eta^{c\bar{c}} = \eta,$$

where $\chi : Y \to \overline{\mathbb{R}}$ with $\eta = \chi^{\bar{c}}$. As such, any further iteration of these transformations cannot increase the dual value of sequence indefinitely.

We now consider a sequence of c- or $\bar{c}$-transforms of our original sequence and once again label it as $(\varphi_n, \psi_n)_{n\in\mathbb{N}}$. To be able to show convergence of this sequence, we want to apply the Theorem of Arzela-Ascoli (A.2). To do so, we still need to show the equiboundedness and equicontinuity of our sequence.

Let us assume that the sequence $(\varphi_n, \psi_n)_{n\in\mathbb{N}}$ is not equibounded. In this case we have $\forall C > 0 \; \exists (x,y) \in X \times Y, N \in \mathbb{N}$:

$$|(\varphi_N, \psi_N)(x,y)| > C.$$

Since φ_n and ψ_n are both bounded, we can assume that

$$\forall n \in \mathbb{N} \; \exists C_n > 0 \; \forall (x,y) \in X \times Y : |(\varphi_n, \psi_n)(x,y)| < C_n,$$

and because $(\varphi_n, \psi_n)_{n\in\mathbb{N}}$ is increasing, we can assume that the sequence $(C_n)_{n\in\mathbb{N}}$ is increasing too. This means that for the specific $N \in \mathbb{N}$ from

the assumption we get $C_N > \max\{C_n : n < N\}$, thus $(c_n)_{n\in\mathbb{N}}$ is strictly increasing. If $(C_n)_{n\in\mathbb{N}}$ were to converge to a constant $\tilde{C} > 0$ for $n \to \infty$, the assumption is contradicted for all $C > \tilde{C}$, as there could never exist a suitable $n \in \mathbb{N}$. If $(C_n)_{n\in\mathbb{N}}$ does not converge, there needs to exist a pair $(\hat{x}, \hat{y}) \in X \times Y$ with $|(\varphi_n, \psi_n)(\hat{x}, \hat{y})| > C_{n-1}$. This too would lead to a contradiction, because $C_n \to \infty$ for $n \to \infty$, and hence for a certain $\tilde{N} \in \mathbb{N}\ \forall n \geq \tilde{N} : \varphi_n \oplus \psi_n > c$, as c is continuous on a compact space and must thus attain a finite maximum. In this case, our sequence would no longer be admissible to (DP).

Let us now assume, that the sequence $(\varphi_n, \psi_n)_{n\in\mathbb{N}}$ is not equicontinuous, i.e. there exists an $\varepsilon > 0$, such that $\forall\delta > 0\ \exists n \in \mathbb{N}, (x, y), (\hat{x}, \hat{y}) \in X \times Y$, $\big|\big(d_X(x, \hat{x}), d_Y(y, \hat{y})\big)\big| < \delta : |(\varphi_n, \psi_n)(x, y) - (\varphi_n, \psi_n)(\hat{x}, \hat{y})| \geq \varepsilon$. Because of this assumption however, we have already contradicted the uniform continuity of all (φ_n, ψ_n) on the compact space $X \times Y$.

We can now apply Arzela-Ascoli (A.2), in order to obtain a subsequence $(\varphi_{n_k}, \psi_{n_k})_{k\in\mathbb{N}}$, which uniformly converges to a continuous and furthermore bounded function $(\varphi, \psi) : X \times Y \to \mathbb{R}$. This subsequence is once again equibounded and convergent, such that we can further apply the Theorem of Lebesgue (A.3) to obtain

$$\lim_{k\in\mathbb{N}} \int_X \varphi_{n_k}\, \mathrm{d}\mu + \lim_{k\in\mathbb{N}} \int_Y \psi_{n_k}\, \mathrm{d}\nu = \int_X \varphi\, \mathrm{d}\mu + \int_Y \psi\, \mathrm{d}\nu.$$

To show the admissibility of (φ, ψ), we note that

$$\varphi_{n_k} \oplus \psi_{n_k} \leq c \Rightarrow \varphi \oplus \psi \leq c.$$

Thus, (φ, ψ) is a solution to (DP) satisfying the desired properties. □

So far we have seen that both (KP) and (DP) admit solutions, but up to now the relation between these solutions has only been $\sup(\text{DP}) \leq \inf(\text{KP})$. Going forward, we want to show $\inf(\text{KP}) \leq \sup(\text{DP})$. One way to do so, is the *c-cyclical monotonicity* of the underlying set $\Omega \subseteq X \times Y$.

Definition 1.19 (c-Cyclical Monotonicity; adapted from [San15, Definition 1.36])**.** For any given function $c : X \times Y \to \mathbb{R} \cup \{\infty\}$, we say that a set $\Omega \subseteq X \times Y$ is c-cyclically monotone (c-CM), if for every $k \in \mathbb{N}$, every permutation $\sigma \in S_k$ and every finite set of points $(x_1, y_1), \ldots, (x_k, y_k) \in \Omega$ we have

$$\sum_{i=1}^{k} c(x_i, y_i) \leq \sum_{i=1}^{k} c(x_i, y_{\sigma(i)}).$$

For such c-cyclically monotone $\Omega \subseteq X \times Y$, we can achieve the following result for c-concave functions and Ω.

Lemma 1.20 (Adapted from [San15, Theorem 1.37]). *Let $\Omega \subseteq X \times Y$ be a c-CM set with $\Omega \neq \emptyset$, and $c : X \times Y \to \mathbb{R}$ be an arbitrary function. Then there exists a c-concave function $\varphi : X \to \mathbb{R} \cup \{-\infty\}$, $\varphi \not\equiv -\infty$, such that*

$$\Omega \subseteq \big\{(x, y) \in X \times Y : \varphi(x) + \varphi^c(y) = c(x, y)\big\}.$$

Proof. Cf. the proof of [San15, Theorem 1.37]. □

In particular, it can be shown that the support of all optimal transport plans for (KP) are c-cyclically monotone.

Lemma 1.21 (Adapted from [San15, Theorem 1.38]). *Let γ be an optimal transport plan for* (KP) *with $c : X \times Y \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ a continuous cost function. Then* $\operatorname{supp}(\gamma)$ *is a c-CM set.*

Proof. Cf. the proof of [San15, Theorem 1.38]. □

In combination with the concepts of c- and $\bar{c}$-concavity, Lemma 1.21 then results in strong duality between (KP) and (DP) for complete and separable metric spaces, and uniformly continuous and bounded cost functions.

Theorem 1.22 (Adapted from [San15, Theorem 1.39]). *Let X and Y be complete and separable metric spaces, and $c : X \times Y \to \mathbb{R}$ be uniformly continuous and bounded. Then the problem* (DP) *admits a solution (φ, φ^c) and strong duality holds, i.e.* $\sup\,(\mathrm{DP}) = \inf\,(\mathrm{KP})$.

Proof. This proof was adapted from the proof of [San15, Theorem 1.39].

We will first consider (KP) in order to find an admissible pair for (DP). As c is continuous and especially l.s.c., we can apply Theorem 1.14 to obtain a solution γ of (KP). By Lemma 1.21, the set $\Omega := \operatorname{supp}(\gamma)$ is c-CM. Applying Lemma 1.20, we can find a c-concave and thus continuous function φ, such that $\Omega \subseteq \big\{(x, y) \in X \times Y : \varphi(x) + \varphi^c(y) = c(x, y)\big\}$. Both φ and φ^c must further be bounded: if we assume φ and φ^c to be unbounded, i.e. for all $C \in \mathbb{R} : \sup |\varphi \oplus \varphi^c| > C$, our c would also be unbounded.

Because (φ, φ^c) is an admissible pair for (DP), we have

$$\int_X \varphi \, \mathrm{d}\mu + \int_Y \varphi^c \, \mathrm{d}\nu \overset{(*)}{=} \int_\Omega \varphi \oplus \varphi^c \, \mathrm{d}\gamma = \int_\Omega c \, \mathrm{d}\gamma = \int_{X \times Y} c \, \mathrm{d}\gamma.$$

Here, the equality $(*)$ holds because γ is concentrated on Ω and $\varphi \oplus \varphi^c = c$ on Ω.

As γ is an optimal transport plan, we get $\sup\,(\mathrm{DP}) \geq \inf\,(\mathrm{KP})$. When this inequality is combined with our previously obtained $\sup\,(\mathrm{DP}) \leq \inf\,(\mathrm{KP})$ we get the claimed strong duality $\sup\,(\mathrm{DP}) = \inf\,(\mathrm{KP})$. □

At the end of Section 1.2, we have generalized the existence results for (KP) to l.s.c. cost functions. In a similar fashion, we would like to relax the constraint on c in Theorem 1.22. We first make the following observation for all l.s.c. cost functions bounded from below. Boundedness from below is usually a practical assumption for cost functions, because negative costs can often be avoided by reformulating the problem to only assign non-negative costs. After the lemma we can achieve strong duality for much more general settings.

Lemma 1.23 (Adapted from [San15, Lemma 1.41])**.** *Let c and c_n be l.s.c. and bounded from below for all $n \in \mathbb{N}$. If $c_n \nearrow c$ for $n \to \infty$, i.e. $(c_n)_{n\in\mathbb{N}}$ converges increasingly towards c, then*

$$\lim_{n\to\infty} \inf \left\{ \int c_n \, \mathrm{d}\gamma : \gamma \in \Pi(\mu,\nu) \right\} = \inf \left\{ \int c \, \mathrm{d}\gamma : \gamma \in \Pi(\mu,\nu) \right\}.$$

Proof. This proof was adapted from the proof of [San15, Lemma 1.41].

As $c_n \leq c, c_n \leq c_{n+1}$ for all $n \in \mathbb{N}$ we have for all $\gamma \in \Pi(\mu,\nu)$:

$$\forall n \in \mathbb{N} : \int\limits_{X\times Y} c_n \, \mathrm{d}\gamma \leq \int\limits_{X\times Y} c \, \mathrm{d}\gamma, \int\limits_{X\times Y} c_n \, \mathrm{d}\gamma \leq \int\limits_{X\times Y} c_{n+1} \, \mathrm{d}\gamma$$

$$\Rightarrow \forall n \in \mathbb{N} : \inf_{\gamma\in\Pi(\mu,\nu)} \int\limits_{X\times Y} c_n \, \mathrm{d}\gamma \leq \inf_{\gamma\in\Pi(\mu,\nu)} \int\limits_{X\times Y} c_{n+1} \, \mathrm{d}\gamma \leq \inf_{\gamma\in\Pi(\mu,\nu)} \int\limits_{X\times Y} c \, \mathrm{d}\gamma.$$

With the convergence of c_n we thus have $\lim\limits_{n\to\infty} \inf\limits_{\gamma\in\Pi(\mu,\nu)} \int c_n \, \mathrm{d}\gamma \leq \inf\limits_{\gamma\in\Pi(\mu,\nu)} \int c \, \mathrm{d}\gamma$.

In the proof of Theorem 1.14, we have already seen that every sequence from $\Pi(\mu,\nu)$ is tight, and from the theorem itself we know that for every c_n there exists an optimal transport map γ_n. Applying Prokhorov (A.1) once more, we obtain a subsequence $(\gamma_{n_k})_{k\in\mathbb{N}}$ which weakly converges to a transport plan $\tilde{\gamma}$. Picking an arbitrary $j \in \mathbb{N}$ we have $\forall n \geq j : c_n \geq c_j$ by induction from the monotonicity of our original sequence, and hence

$$\lim_{n\in\mathbb{N}} \inf_{\gamma\in\Pi(\mu,\nu)} \int\limits_{X\times Y} c_n \, \mathrm{d}\gamma = \lim_{n\in\mathbb{N}} \int\limits_{X\times Y} c_n \, \mathrm{d}\gamma_n \geq \liminf_{n\in\mathbb{N}} \int\limits_{X\times Y} c_j \, \mathrm{d}\gamma_n.$$

As $|c_j| \leq c$ we can now take the limit over j and apply Lebesgue (A.3) to get

$$\lim_{n\in\mathbb{N}} \inf_{\gamma\in\Pi(\mu,\nu)} \int\limits_{X\times Y} c_n \, \mathrm{d}\gamma \geq \liminf_{n\in\mathbb{N}} \lim_{j\in\mathbb{N}} \int\limits_{X\times Y} c_j \, \mathrm{d}\gamma_n = \liminf_{n\in\mathbb{N}} \int\limits_{X\times Y} c \, \mathrm{d}\gamma_n.$$

We finally see that $\liminf\limits_{n\in\mathbb{N}} \int c \, \mathrm{d}\gamma_n \geq \inf\limits_{\gamma\in\Pi(\mu,\nu)} \int c \, \mathrm{d}\gamma$, and thus our claim. □

With Lemma 1.23, we can obtain a theorem about strong duality with preliminaries resembling those of Theorem 1.14. This result will conclude the first part, having laid out a foundation of optimal transport theory. The focus will then turn towards regularization in optimal transport, applying analogous duality results, and estimating optimal transport plans and maps in large-scale problem settings.

Theorem 1.24 (Adapted from [San15, Theorem 1.42])**.** *Let X and Y be complete and separable metric spaces, and $c : X \times Y \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ be an l.s.c. cost function. Then,* $\sup(\mathrm{DP}) = \inf(\mathrm{KP})$ *holds.*

Proof. This proof was adapted from the proof of [San15, Theorem 1.42].

For this proof, we would like to apply Lemma 1.23, however this requires the existence of an l.s.c. sequence $(c_n)_{n\in\mathbb{N}}, \forall n \in \mathbb{N} : c_n : X \times Y \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ with $c_n \nearrow c$ for $n \to \infty$. It turns out that this is no problem: as [San15, Box 1.5] points out, c is l.s.c. if and only if there exists a sequence of K-Lipschitz functions with $c_n \nearrow c$ for $n \to \infty$. Here K-Lipschitz refers to $\forall n \in \mathbb{N}, z, \hat{z} \in X \times Y : |c_n(z) - c_n(\hat{z})| \leq K \cdot d_{X\times Y}(z, \hat{z})$.

With Lipschitz continuity implying uniform continuity, and by uniform continuity implying lower semi-continuity through regular continuity, this sequence is l.s.c. Using the uniform continuity and the bound obtained from c, we can apply Theorem 1.22 and $\varphi \oplus \psi \leq c_n \leq c$ to get

$$\inf_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c_n \, \mathrm{d}\gamma = \sup_{\varphi\in C_b(X),\psi\in C_b(Y),\varphi\oplus\psi\leq c_n} \int_X \varphi \, \mathrm{d}\mu + \int_Y \psi \, \mathrm{d}\nu$$

$$\leq \sup_{\varphi\in C_b(X),\psi\in C_b(Y),\varphi\oplus\psi\leq c} \int_X \varphi \, \mathrm{d}\mu + \int_Y \psi \, \mathrm{d}\nu.$$

Using Lemma 1.23 for $n \to \infty$, we have $\inf(\mathrm{KP}) \leq \sup(\mathrm{DP})$. Hence, in combination with $\sup(\mathrm{DP}) \leq \inf(\mathrm{KP})$, strong duality holds. □

This concludes the first part of the thesis. We will now work towards large-scale mapping estimation, making use of regularized optimal transport along the way.

2 Regularized Transport and Mapping Estimation

In the following, $\Omega_1 \subseteq \mathbb{R}^{n_1}$, $\Omega_2 \subseteq \mathbb{R}^{n_2}$ and $\Omega_1 \times \Omega_2 =: \Omega \subseteq \mathbb{R}^{n_1+n_2}$ will always be compact for some $n_1, n_2 \in \mathbb{N}$.

We will first introduce regularized optimal transport. The process of regularization imposes an additional penalty on the objective function of an optimization problem. This can serve multiple purposes: regularizing an ill-posed problem can allow for an approximate solution, result in uniqueness of the solution or, especially in the case covered in this thesis, allow for duality results which net themselves for further applications.

The barycentric projection will then be used to obtain a transport map from the optimal regularized transport plan. Due to the explicit computation of these transport solutions often being impractical for applications, it is necessary to estimate optimal transport plans and maps. This is traditionally done by formulating a regularized dual problem and applying the Sinkhorn algorithm, however this particular algorithm is not suited to handle large-scale problems. Hence, a different approach involving neural networks and stochastic mapping estimation is required.

The first sections on regularization and the corresponding regularized dual problem follow [CLMW21] closely, while the last section is adapted from [SDF+18, Section 3 and Section 4].

2.1 Regularized Optimal Transport

There are multiple ways in which regularization can take place, however we will focus on *entropic regularization*. This form of regularization deals with measurable functions with finite entropy, i.e. for $f \in \mathfrak{M}(\Omega, \mathbb{R})$

$$E(f) := \int_{\Omega} \big|f(z)\big| \log\big(|f(z)|\big) \, \mathrm{d}z < \infty,$$

where $0 \cdot \log(0) := 0$. From here we can define

$$L\log L(\Omega) := \left\{ f \in \mathfrak{M}(\Omega, \mathbb{R}) : \int_\Omega \big|f(z)\big| \log^+\big(|f(z)|\big)\ \mathrm{d}z < \infty \right\}$$

as a subset of all measurable functions with finite entropy, where we consider $\log^+(x) := \max\{0, \log(x)\}$.

The first conclusion about the entropic regularization that can be reached via [CLMW21, Proposition 2.1], is that $f \in \mathfrak{M}(\Omega, \mathbb{R}_{\geq 0})$ is equivalent to $E(f) < \infty$ and $f \in L\log L(\Omega)$.

The second conclusion regards the structure of $L\log L(\Omega)$ in accordance with [CLMW21, Definition 2.3]. When considering the *Luxemburg norm*

$$\|f\|_\Phi := \inf \left\{ C > 0 : \int_\Omega \Phi\left(\frac{|f(z)|}{C}\right)\ \mathrm{d}z \leq 1 \right\}$$

for measurable functions f, we denote $L^\Phi(\Omega) := (\mathcal{F}, \|\cdot\|_\Phi)$ the *Orlicz space* for a certain function Φ, where $\mathcal{F} := \{f \in \mathfrak{M}(\Omega, \mathbb{R}) : \|f\|_\Phi < \infty\}$. Using $\Phi_{\log}(x) := x\log^+(x)$, we have $L^{\Phi_{\log}}(\Omega) = L\log L(\Omega)$. In fact, by [CLMW21, Theorem 2.5], $L\log L(\Omega)$ is a Banach space with respect to its Luxemburg norm.

For a third conclusion we are interested in the dual space of $L\log L(\Omega)$. We define

$$\Phi_{\exp}(s) := \begin{cases} s, & 0 \leq s \leq 1 \\ \exp(s-1), & s > 1 \end{cases}$$

and $L_{\exp}(\Omega) := L^{\Phi_{\exp}}(\Omega)$ with its corresponding Luxemburg norm $\|\cdot\|_{\Phi_{\exp}}$. As shown in [CLMW21, Proposition 2.7], if Ω has finite Lebesgue measure (which it has by $\Omega \subseteq \mathbb{R}^{n_1+n_2}$ being compact and thus bounded), then

$$(L\log L(\Omega))' = L_{\exp}(\Omega).$$

Here, $(\cdot)'$ denotes the dual space in accordance with [RY08, Definition 4.28].

We can now state the formulation of the regularized optimal transport problem, based on (KP). Furthermore, a predual problem will be given.

Definition 2.1 (Regularized Problem; adapted from [CLMW21, (P), (D), and Proposition 3.1])**.** Consider $\mu \in \mathcal{P}(\Omega_1)$ and $\nu \in \mathcal{P}(\Omega_2)$ to be our starting measures, and $c : \Omega \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ to be a cost function. The **Regularized Problem** for some $\varepsilon > 0$ is defined as

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_\Omega c(x,y)\ \mathrm{d}\gamma(x,y) + \varepsilon \int_\Omega \gamma(x,y)\Big(\log\big(\gamma(x,y)\big) - 1\Big)\ \mathrm{d}(x,y).$$

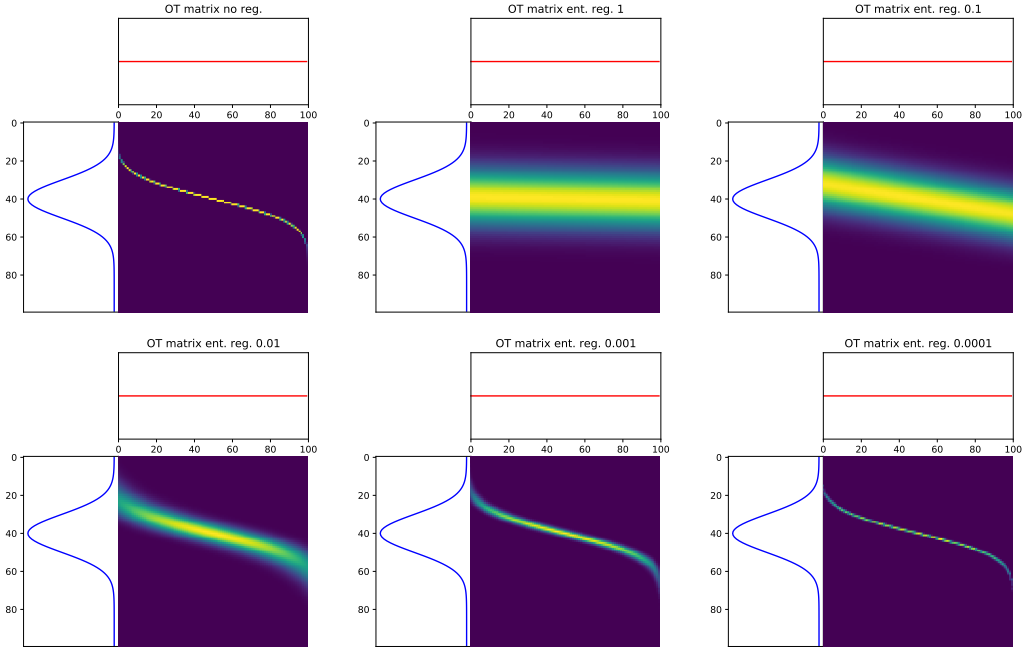

Figure 1: Solutions for unregularized and entropically regularized optimal transport problems, $\varepsilon \in \{1, 0.1, 0.01, 0.001, 0.0001\}$

Its **Predual Problem** for the same $\varepsilon > 0$ is defined as

$$\sup_{\varphi \in C_b(\Omega_1), \psi \in C_b(\Omega_2)} \int_{\Omega_1} \varphi(x) \, \mathrm{d}\mu(x) + \int_{\Omega_2} \psi(y) \, \mathrm{d}\nu(y) - \varepsilon \int_{\Omega} F_\varepsilon(x, y) \, \mathrm{d}(x, y),$$

with $F_\varepsilon(x, y) := \exp\left(\frac{1}{\varepsilon}\big(\varphi(x) + \psi(y) - c(x, y)\big)\right)$. Just as in the definitions of the previous problems, we will use the abbreviations (RP), inf (RP), (PD) and sup (PD) for the problems and their respective solutions. To distinguish from transport plans for (KP), we denote γ^ε as the optimal transport plan for (RP) and a specific ε. Further, by [CLMW21, Proposition 3.1] we have that inf (RP) = sup (PD), and, if sup (PD) is finite, (RP) admits a minimizer.

Remark. Figure 1 shows the effect of entropic regularization on solutions of the optimal transport problem. The software provided by [FCG+21] was used to generate the solutions between a Gaussian and a uniform distribution. The upper left image shows the unregularized result and the remaining images show the results for entropic regularization.

With Theorem 1.14, we already saw that an optimal transport plan exists for (KP). For (RP) however, we require more constraints on our marginal measures. Most importantly, we involve the $L \log L$ spaces from above.

Theorem 2.2 (Taken from [CLMW21, Theorem 3.3]). *The problem* (RP) *admits a minimizer $\gamma \in \Pi(\mu, \nu)$ if and only if we have $\mu \in L \log L(\Omega_1)$ and $\nu \in L \log L(\Omega_2)$. In this case, $\gamma \in L \log L(\Omega)$ is unique.*

Proof. Cf. the proof of [CLMW21, Theorem 3.3]. □

With a solution to (RP), we can achieve weak convergence in the sense of Definition 1.6 of the regularized optimal transport plans to a solution of (KP). For this we first consider X and Y as complete metric spaces, and $\Omega_1 \subseteq X, \Omega_2 \subseteq Y, \Omega := \Omega_1 \times \Omega_2$ as compact subsets. Given two probability measures $\mu \in \mathcal{P}(\Omega_1)$ and $\nu \in \mathcal{P}(\Omega_2)$, we further require the following two discrete probability measures, weakly converging to μ and ν for $n \to \infty$ respectively, which are defined as

$$\mu_n := \sum_{i=1}^{n} a_i \delta_{x_i}, \nu_n := \sum_{j=1}^{n} b_j \delta_{y_j},$$

with $x_i \in \Omega_1, a_i \geq 0, y_j \in \Omega_2, b_j \geq 0$ for $1 \leq i, j \leq n$.

With these preliminaries, the following relation between the solutions of (RP) and (KP) can be shown.

Theorem 2.3 (Adapted from [SDF$^+$18, Theorem 1]). *Consider X, Y, Ω_1, Ω_2, Ω, μ, μ_n, ν and ν_n as described above. Let $c : \Omega \to \mathbb{R}_{\geq 0}$ be a finite and continuous cost function, and $(\varepsilon_n)_{n\in\mathbb{N}}$ be a sequence converging to 0 sufficiently fast with $\varepsilon_n > 0$ for all $n \in \mathbb{N}$. Then for $(\gamma_n^{\varepsilon_n})_{n\in\mathbb{N}}$, the sequence of solutions of* (RP) *between μ_n and ν_n with $\varepsilon = \varepsilon_n$, there exists a subsequence $(\gamma_{n_k}^{\varepsilon_{n_k}})_{k\in\mathbb{N}}$ that weakly converges to a solution γ of* (KP) *between μ and ν for $n \to \infty$.*

Proof. This proof was adapted from the proof of [SDF$^+$18, Theorem 1].

We consider γ_n the solution of (KP) between μ_n and ν_n with maximum entropy. According to [Vil09, Theorem 5.20], there now exists a subsequence which weakly converges to a solution γ of the same problem between μ and ν. We continue to label this subsequence as γ_n, and further take the same indices to form a subsequence from the solutions of (RP) between μ_n and ν_n, labelling it $\gamma_n^{\varepsilon_n}$. The solutions of (RP) exist due to the discrete measures μ_n, ν_n reducing the integral over Ω_1 and Ω_2 respectively to a finite sum over their masses x_i, y_j, hence allowing for the application of Theorem 2.2. By expanding the following integral for $g \in C_b(\Omega)$

$$\left| \int_\Omega g \, \mathrm{d}\gamma_n^{\varepsilon_n} - \int_\Omega g \, \mathrm{d}\gamma \right| \leq \left| \int_\Omega g \, \mathrm{d}\gamma_n^{\varepsilon_n} - \int_\Omega g \, \mathrm{d}\gamma_n \right| + \left| \int_\Omega g \, \mathrm{d}\gamma_n - \int_\Omega g \, \mathrm{d}\gamma \right|,$$

we have to show that the left summand on the right-hand side converges to 0, because the other summand converges by the previously mentioned statement of [Vil09]. As $\gamma_n^{\varepsilon_n}$ and γ_n are solutions to optimal transport problems between discrete measures, we can replace the integral over Ω with sums over the masses of both measures and obtain

$$\left| \sum_{i,j=1}^{n} g(x_i, y_j)\gamma_n^{\varepsilon_n}(x_i, y_j) - \sum_{i,j=1}^{n} g(x_i, y_j)\gamma_n(x_i, y_j) \right| \leq M_g \|\gamma_n^{\varepsilon_n} - \gamma_n\|_1^{n\times n},$$

where $\|\gamma_n^{\varepsilon_n} - \gamma_n\|_1^{n\times n} := \sum_{i,j=1}^{n} \left|\gamma_n^{\varepsilon_n}(x_i, y_j) - \gamma_n(x_i, y_j)\right|$ and $M_g := \max g(x_i, y_j)$ over $1 \leq i, j \leq n$. Adapting a result from [CM94, specifically Equation 2 in the proof of Proposition 3.1], there exist $M_{c_n,\mu_n,\nu_n}, \lambda_{c_n,\mu_n,\nu_n} > 0$ such that

$$\|\gamma_n^{\varepsilon_n} - \gamma_n\|_1^{n\times n} \leq M_{c_n,\mu_n,\nu_n} \exp\left(\frac{-\lambda_{c_n,\mu_n,\nu_n}}{\varepsilon_n}\right),$$

with $c_n := \big(c(x_i, y_j)\big)_{i,j=1}^{n}$. To show the convergence of the right-hand side, we finally choose $\varepsilon_n = \lambda_{c_n,\mu_n,\nu_n} \cdot \ln\left(n\, M_{c_n,\mu_n,\nu_n}\right)^{-1} \to 0, n \to \infty$. □

Before handling the dual problem of (RP), we investigate the *barycentric projection* in the next section. The barycentric projection is a specific method of obtaining a transport map from a transport plan, allowing us to take a solution of (RP), and get a solution of (MP) in return.

The results of the following two sections will then be combined in the last section of this part to enable the estimation of optimal transport plans and maps for large-scale problems.

2.2 The Barycentric Projection

In the first part, specifically in Theorem 1.9, we already saw that the set of transport plans induced by transport maps is densely contained within the set of all transport maps $\Pi(\mu, \nu)$. We are now interested in finding a transport map from our regularized optimal transport plan, i.e. a solution to (RP) between μ and ν, since the handling of a mapping instead of a joint measure is often advantageous in direct applications. We have already seen in Theorem 1.11, that when μ is atomless, such a transport map is directly related to the accompanying transport plan. This suggests that we start with a weakly convergent subsequence of regularized transport plans and apply a transformation which will then result in a sequence weakly converging to an optimal transport map. The transformation chosen by [SDF+18], is the so-called barycentric projection.

Definition 2.4 (Barycentric Projection; adapted from [SDF+18, Definition 1])**.** We consider a transport plan $\gamma \in \Pi(\mu, \nu)$ and a convex cost function $d : \Omega_2 \times \Omega_2 \to \mathbb{R}_{\geq 0} \cup \{\infty\}$. Then the **barycentric projection** of γ at $x \in \Omega_1$ is defined as

$$\bar{\gamma}(x) := \operatorname*{arg\,min}_{z \in \Omega_2} \int_{\Omega_2} d(z, y) \, \mathrm{d}\gamma(x, y).$$

If $d(x, y) = \|x - y\|_2^2$, where $\| \cdot \|_2$ is the Euclidean norm, the barycentric projection has the closed form

$$\bar{\gamma}(x) = \int_{\Omega_2} y \, \mathrm{d}\gamma(x, y).$$

If $c(x, y) = \|x - y\|_2^2$, according to [SDF+18], it can be further shown that the barycentric projection of an optimal transport plan for (KP) is already an optimal transport map for (MP).

Similarly to Theorem 2.3, we will be considering $X = Y = \mathbb{R}^d$, with $\Omega_1, \Omega_2 \subseteq \mathbb{R}^d, \Omega := \Omega_1 \times \Omega_2$ compact, a continuous measure $\mu \in \mathcal{P}(\Omega_1)$ that satisfies [Vil09, Corollary 9.3], an arbitrary measure $\nu \in \mathcal{P}(\Omega_2)$, a continuous cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, as well as the discrete measures

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} a_i \delta_{x_i}, \nu_n := \frac{1}{n} \sum_{j=1}^{n} b_j \delta_{y_j},$$

with $x_i \in \Omega_1, a_i \geq 0, y_j \in \Omega_2, b_j \geq 0$ for $1 \leq i, j \leq n$, weakly converging to μ and ν respectively. This then allows us to show the following result, applying the barycentric projection on regularized optimal transport plans to obtain an optimal transport map for (MP).

Theorem 2.5 (Adapted from [SDF+18, Theorem 2])**.** *Consider X, Y, Ω_1, Ω_2, Ω, μ, ν, c, μ_n, ν_n as described above. Assume that the solutions γ_n to* (KP) *between μ_n and ν_n are unique for all $n \in \mathbb{N}$. Further let $(\varepsilon_n)_{n \in \mathbb{N}}$ be a sequence converging sufficiently fast to 0 with $\varepsilon_n > 0$ for all $n \in \mathbb{N}$, and d be a convex cost function on $\Omega_2 \times \Omega_2$. Then there exists a subsequence $\left(\bar{\gamma}_{n_k}^{\varepsilon_{n_k}}\right)_{k \in \mathbb{N}}$ such that $\left(id_{\Omega_1}, \bar{\gamma}_{n_k}^{\varepsilon_{n_k}}\right) \# \mu_n \rightharpoonup (id_{\Omega_1}, T) \# \mu$ for $n \to \infty$, where T is the map solving* (MP) *between μ and ν, id_{Ω_1} is the identity map on Ω_1, and $\bar{\gamma}_{n_k}^{\varepsilon_{n_k}}$ is the barycentric projection with respect to the cost function d of the solution $\gamma_{n_k}^{\varepsilon_{n_k}}$ of* (RP).

Proof. This proof was adapted from the proof of [SDF+18, Theorem 2].

We know that $\bar{\gamma}_n^{\varepsilon_n}\#\mu_n - T\#\mu = \bar{\gamma}_n^{\varepsilon_n}\#\mu_n - \bar{\gamma}_n\#\mu_n + \bar{\gamma}_n\#\mu_n - T\#\mu$. When considering the absolute value of the left-hand side and applying the triangle inequality to the right-hand side, we can prove the weak convergence of our sequence (cf. Definition 1.6) claimed in the theorem by proving it for both summands on the right-hand side. By [Vil09, Corollary 9.3], there exists a Monge map between μ and ν due to μ being uniformly continuous on Ω_1 and thus absolutely continuous with respect to the Lebesgue measure. When following the proof of [SDF+18, Theorem 2], we see that the summand involving $\bar{\gamma}_n\#\mu_n - T\#\mu$ converges.

We thus only have to show that for any Lipschitz continuous function g over Ω

$$\left| \int_\Omega g \, \mathrm{d}(id_{\Omega_1}, \bar{\gamma}_n^{\varepsilon_n})\#\mu_n - \int_\Omega g \, \mathrm{d}(id_{\Omega_1}, \bar{\gamma}_n)\#\mu_n \right| \to 0$$

for $n \to \infty$ and ε_n converging to 0 sufficiently fast, as $\gamma_n \rightharpoonup T$ by [Vil09, Theorem 5.20]. According to [SDF+18], the optimal transport plan γ_n is induced by an optimal map T_n, i.e. $\gamma_n = (id_{\Omega_1}, T_n)\#\mu_n$, which then further implies $(id_{\Omega_1}, \bar{\gamma}_n)\#\mu_n = (id_{\Omega_1}, T_n)\#\mu_n$. This allows us to apply the definition of the push forward and rewrite

$$\int_\Omega g \, \mathrm{d}(id_{\Omega_1}, \bar{\gamma}_n^{\varepsilon_n})\#\mu_n = \int_\Omega g\big(x, \bar{\gamma}_n^{\varepsilon_n}(x)\big) \, \mathrm{d}\mu_n(x) = \frac{1}{n}\sum_{i=1}^n g\big(x_i, \bar{\gamma}_n^{\varepsilon_n}(x_i)\big),$$

$$\int_\Omega g \, \mathrm{d}(id_{\Omega_1}, \bar{\gamma}_n)\#\mu_n = \int_\Omega g\big(x, T_n(x)\big) \, \mathrm{d}\mu_n(x) = \frac{1}{n}\sum_{i=1}^n g\big(x_i, T_n(x_i)\big).$$

We apply the closed form of the barycentric projection as given by [SDF+18] $\bar{\gamma}(x) = \gamma(x)Y_n$, with $Y_n := (y_1, \ldots, y_n)^T$ and $\gamma(x) := \big(\gamma(x, y_1), \ldots, \gamma(x, y_n)\big)$, to obtain

$$\left| \int_\Omega g\big(x, \bar{\gamma}_n^{\varepsilon_n}(x)\big) \, \mathrm{d}\mu_n(x) - \int_\Omega g\big(x, T_n(x)\big) \, \mathrm{d}\mu_n(x) \right| \leq \sum_{i=1}^n K_g \big|\bar{\gamma}_n^{\varepsilon_n}(x_i) - \bar{\gamma}_n(x_i)\big|$$

$$= n \; K_g \; \|\gamma_n^{\varepsilon_n} Y_n - \gamma_n Y_n\|_{\mathbb{R}^{n\times d},2} \leq n \; K_g \; \|Y_n\|_{\mathbb{R}^{n\times d},2}^{1/2} \; \|\gamma_n^{\varepsilon_n} - \gamma_n\|_{\mathbb{R}^{n\times n},2}^{1/2},$$

where we use $K_g < \infty$ the Lipschitz constant of g, $\|A\|_{\mathbb{R}^{d\times n},2} = \sigma_{\max}(A^TA)$, $\|A\|_{\mathbb{R}^{n\times n},2} = \sigma_{\max}(A^TA)$, $\gamma_n^{\varepsilon_n} = \big(\gamma_n^{\varepsilon_n}(x_i, y_j)\big)_{i,j=1}^n$, γ_n in a similar fashion, and Cauchy-Schwarz is used on the last inequality. Using the same result by [CM94] as in the proof of Theorem 2.3, we can now find $M_{c_n,\mu_n,\nu_n} > 0$ and $\lambda_{c_n,\mu_n,\nu_n} > 0$ such that

$$\|\gamma_n^{\varepsilon_n} - \gamma_n\|_{\mathbb{R}^{n\times n},2}^{1/2} \leq M_{c_n,\mu_n,\nu_n} \exp\left(\frac{-\lambda_{c_n,\mu_n,\nu_n}}{\varepsilon_n}\right).$$

Choosing $\varepsilon = \lambda_{c_n,\mu_n,\nu_n} \cdot \ln\left(n^2 \|Y\|^{1/2}_{\mathbb{R}^{n\times d},2} M_{c_n,\mu_n,\nu_n}\right)^{-1}$ suffices for the right-hand side of the previous equation to converge to 0 for any Lipschitz continuous g over Ω, showing $\bar{\gamma}_n^{\varepsilon_n} \rightharpoonup T$ for $n \to \infty$. □

As an immediate consequence, we can even avoid the notion of transport plans via the push-forward, and directly formulate the barycentric projection as a transport map.

Corollary 2.6 (Taken from [SDF+18, Corollary 1])**.** *With exactly the same assumptions as for Theorem 2.5, there exists a subsequence $(n_k)_{k\in\mathbb{N}}, n_k \in \mathbb{N}$ for all $k \in \mathbb{N}$, such that $\bar{\gamma}_{n_k}^{\varepsilon_{n_k}} \# \mu_{n_k} \rightharpoonup \nu$ for $n \to \infty$.*

Proof. This proof was adapted from the proof of [SDF+18, Corollary 1].

We consider $h \in C_b(\Omega_2)$ and further define $g : \Omega \to \mathbb{R}$ by $g(x, y) := h(y)$ for all $(x, y) \in \Omega$. As h is bounded and continuous, g is as well. With the same notation as in the proof of Theorem 2.5, we see that

$$\left| \int_{\Omega_2} h \, \mathrm{d}\bar{\gamma}_n^{\varepsilon_n} \# \mu_n - \int_{\Omega_2} h \, \mathrm{d}T \# \mu \right| = \left| \int_{\Omega} g \, \mathrm{d}(id_{\Omega_1}, \bar{\gamma}_n^{\varepsilon_n}) \# \mu_n - \int_{\Omega} g \, \mathrm{d}(id_{\Omega_1}, T) \# \mu \right|.$$

Applying the theorem mentioned above, we see the left-hand side converging to 0 for a subsequence if ε_n converges sufficiently fast. Finally, we can make use of $T\#\mu = \nu$ to prove the corollary. □

Before we are able to establish large-scale mapping estimation for the optimal transport problem, we need to investigate the dual of the regularized problem. This will occur in a similar way, as was done in Section 1.3 for the dual of (KP).

2.3 Regularized Duality

As we have previously seen, the regularization term is finite if and only if $\gamma \in L \log L(\Omega)$. By considering (RP) over $L \log L(\Omega)$ instead of $\Pi(\mu, \nu)$, we can also derive a dual problem over $L_{\exp}(\Omega_1) \times L_{\exp}(\Omega_2)$. Following the argument put forth in [CLMW21, Section 4], we define

$$\Phi(s) := \begin{cases} \infty, & s < 0 \\ s, & 0 \le s \le 1 \\ \exp(s-1), & s > 1 \end{cases}, \Psi(s) := \begin{cases} -\infty, & s \le 0 \\ \log(s), & 0 < s \le 1 \\ s - 1, & s > 1 \end{cases},$$

$$u_1 := \begin{cases} \exp\left(\frac{\varphi}{\varepsilon}\right), & \varphi \leq 0 \\ \frac{\varphi}{\varepsilon} + 1, & \varphi > 0 \end{cases}, \text{ and } u_2 := \begin{cases} \exp\left(\frac{\psi}{\varepsilon}\right), & \psi \leq 0 \\ \frac{\psi}{\varepsilon}, & \psi > 0 \end{cases}$$

and obtain $\varphi = \varepsilon \log\big(\Phi(u_1)\big) = \varepsilon\Psi(u_1)$ and $\psi = \varepsilon \log\big(\Phi(u_2)\big) = \varepsilon\Psi(u_2)$. When put back into (PD), this gives us

$$\int_{\Omega_1} \varphi(x)\ \mathrm{d}\mu(x) + \int_{\Omega_2} \psi(y)\ \mathrm{d}\nu(y) - \varepsilon \int_{\Omega} \exp\left(\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon}\right)\ \mathrm{d}(x,y)$$

$$= \varepsilon \int_{\Omega_1} \Psi\big(u_1(x)\big)\ \mathrm{d}\mu(x) + \varepsilon \int_{\Omega_2} \Psi\big(u_2(y)\big)\ \mathrm{d}\nu(y) - \varepsilon \int F_\varepsilon^\Phi(x,y)\ \mathrm{d}(x,y),$$

with $F_\varepsilon^\Phi(x,y) := \Phi\big(u_1(x)\big)\Phi\big(u_2(y)\big) \exp\left(\frac{-c(x,y)}{\varepsilon}\right)$. We note that $\Phi = \Phi_{\exp}$ on $\mathbb{R}_{\geq 0}$, suggesting an optimization over $(u_1, u_2) \in L_{\exp}(\Omega_1) \times L_{\exp}(\Omega_2)$ with $u_1, u_2 \geq 0$ (as the problem would be infeasible for $u_1, u_2 < 0$).

This argumentation gives rise to the following regularized dual problem.

Definition 2.7 (Regularized Dual; adapted from [CLMW21, ($\mathrm{D}_{\exp}$)])**.** Let μ, ν, ε as well as c be the same as in Definition 2.1. We consider the set $\mathcal{F} := \big\{(f,g) \in L_{\exp}(\Omega_1) \times L_{\exp}(\Omega_2) : f, g \geq 0\big\}$ and define the **Regularized Dual Problem** as

$$\sup_{\mathcal{F}} \left\{ \varepsilon \int_{\Omega_1} \Psi\big(u_1(x)\big)\ \mathrm{d}\mu(x) + \varepsilon \int_{\Omega_2} \Psi\big(u_2(y)\big)\ \mathrm{d}\nu(y) - \varepsilon \int_{\Omega} F_\varepsilon^\Phi(x,y)\ \mathrm{d}(x,y) \right\}.$$

This problem will accordingly be labelled as (RD) and its supremum as $\sup(\mathrm{RD})$. The restriction on $\mathcal{F}$ ensures that all integrals are well-defined.

Theorem 2.2 gave an existence result for (RP). An analogous statement for (RD) is given by the following theorem, however the class of functions is now $L_{\exp}$ instead of $L \log L$. Following this result, we will investigate the duality between (RP) and (RD), like we investigated the duality between (KP) and (DP) in Section 1.3. Finally, these duality results will provide a way to formulate dual optimality conditions, which will serve as a starting point for plan and mapping estimation for the regularized transport problem.

Theorem 2.8 (Adapted from [CLMW21, Theorem 4.6])**.** *The problem* (RD) *admits a solution* $(\bar{u}_1, \bar{u}_2) \in L_{\exp}(\Omega_1) \times L_{\exp}(\Omega_2)$.

Proof. This proof was adapted from the proof of [CLMW21, Theorem 4.6].

We define the functional

$$B(u_1, u_2) := \int_\Omega \Phi\big(u_1(x)\big)\Phi\big(u_2(y)\big) \exp\left(\frac{-c(x,y)}{\varepsilon}\right) \mathrm{d}(x,y)$$

$$- \int_{\Omega_1} \Psi(u_1) \, \mathrm{d}\mu - \int_{\Omega_2} \Psi(u_2) \, \mathrm{d}\nu$$

and show that it possesses a minimizer. The value taken by B is finite for e.g. $u_1 \equiv u_2 \equiv 1$, allowing us to consider a minimizing sequence $(u_1^n, u_2^n)_{n\in\mathbb{N}}$ with $(u_1^n, u_2^n) \in L_{\exp}(\Omega_1) \times L_{\exp}(\Omega_2)$ for all $n \in \mathbb{N}$. By [CLMW21, Lemma 4.2], we can assume $\int \Phi_{\exp}(u_1^n) \, \mathrm{d}x = 1$ without loss of generality. With [CLMW21, Lemma 4.4], we observe that the norm $\|u_2^n\|_{\Phi_{\exp}}$ is bounded. By applying the Theorem of Banach-Alaoglu (A.4), we can find a weakly-* convergent subsequence $(u_1^{n_k}, u_2^{n_k})$. Making use of [CLMW21, Lemma 4.5], the functional B is sequentially weakly-* lower semi-continuous on $L_{\exp}(\Omega_1) \times L_{\exp}(\Omega_2)$, showing that the minimum is attained over $L_{\exp}(\Omega_1) \times L_{\exp}(\Omega_2)$. □

Continuing the previous argument, we can now substitute $\bar{\phi} = \varepsilon\Psi(\bar{u}_1)$ as well as $\bar{\psi} = \varepsilon\Psi(\bar{u}_2)$ from our dual solutions. We can further assume $\bar{u}_1 > 0$ μ-a.e. and $\bar{u}_2 > 0$ ν-a.e., as otherwise $\int \Psi(\bar{u}_1) \, \mathrm{d}\mu + \int \Psi(\bar{u}_2) \, \mathrm{d}\nu = -\infty$, rendering the problem infeasible. The pair $(\bar{\phi}, \bar{\psi})$ however is not guaranteed to be admissible for (PD), i.e. $C_b(\Omega_1) \times C_b(\Omega_2)$, as even for $\bar{u}_1, \bar{u}_2 > 0$ we do not necessarily get bounds on $\bar{\phi}$ and $\bar{\psi}$, since $\Psi(x) \to -\infty$ for $x \to 0$.

If we instead restrict ourselves to the assumptions from Theorem 2.2, we obtain strong duality between (RP) and (RD).

Theorem 2.9 (Adapted from [CLMW21, Proposition 4.7])**.** *We consider $\mu \in L\log L(\Omega_1), \nu \in L\log L(\Omega_2)$ and $c \in C(\Omega)$. Then* (RP) *admits a solution $\bar{\gamma} \in L\log L(\Omega)$,* (RD) *admits a solution $(\bar{u}_1, \bar{u}_2) \in L_{\exp}(\Omega_1) \times L_{\exp}(\Omega_2)$, and strong duality holds, i.e.* $\sup(\mathrm{RD}) = \inf(\mathrm{RP})$.

Proof. This proof was taken from the proof [CLMW21, Proposition 4.7]. The existence of a solution follows from Theorem 2.2 and Theorem 2.8.

According to [CLMW21, Proposition 3.1], it suffices to show that strong duality holds between the predual problem and the regularized dual problem, i.e. $\sup(\mathrm{PD}) = \sup(\mathrm{RD})$. For any arbitrary $\alpha \in C(\Omega_1), \beta \in C(\Omega_2)$ we set $u_1 := \Psi^{-1}(\alpha/\varepsilon), u_2 := \Psi(\beta/\varepsilon)$ and see that

$$\int_{\Omega_1} \alpha \, \mathrm{d}\mu + \int_{\Omega_2} \beta \, \mathrm{d}\nu - \varepsilon \int_\Omega \exp\left(\frac{-c+\alpha+\beta}{\varepsilon}\right) \mathrm{d}(x,y) \leq \max_{u_1,u_2\geq 0} -\varepsilon B(u_1,u_2),$$

with $B(\cdot,\cdot)$ being defined as in the proof of Theorem 2.8. By taking the supremum over all α and β, we get $\sup(\mathrm{PD}) \leq \sup(\mathrm{RD})$.

To see $\sup(\mathrm{RD}) \leq \sup(\mathrm{PD})$, one may follow the argument put forth by [CLMW21], using the theorems of monotone convergence applied to B and dominated convergence applied to a uniformly bounded, weakly-* convergent subsequence. □

As the final result of this section, the following corollary provides two optimality conditions for solutions of (RD), as well as a characterization of primal solutions to (RP) in terms of its optimal dual variables.

Corollary 2.10 (Conditions on Dual Optimality; adapted from [CLMW21, Theorem 4.8])**.** *Let $\mu \in L\log L(\Omega_1), \nu \in L\log L(\Omega_2)$ and $c \in C(\Omega)$, as in Theorem 2.9. Then, for μ-a.e. $x \in \Omega_1$ and ν-a.e. $y \in \Omega_2$ all dual solutions $(\bar{u}_1, \bar{u}_2) \in L_{\exp}(\Omega_1) \times L_{\exp}(\Omega_2)$ of* (RD) *satisfy*

$$\mu(x) = \Phi\big(\bar{u}_1(x)\big) \int_{\Omega_2} \Phi\big(\bar{u}_2(y)\big) \exp\left(\frac{-c(x,y)}{\varepsilon}\right) \mathrm{d}y, \tag{2.1}$$

$$\nu(y) = \Phi\big(\bar{u}_2(y)\big) \int_{\Omega_1} \Phi\big(\bar{u}_1(x)\big) \exp\left(\frac{-c(x,y)}{\varepsilon}\right) \mathrm{d}x. \tag{2.2}$$

Furthermore, a solution $\bar{\gamma} \in L\log L(\Omega)$ of (RP) *is defined by*

$$\bar{\gamma}(x,y) = \Phi\big(\bar{u}_1(x)\big)\Phi\big(\bar{u}_2(y)\big) \exp\left(\frac{-c(x,y)}{\varepsilon}\right).$$

Proof. Cf. the proof of [CLMW21, Theorem 4.8]. □

The final section of the second part will contain three algorithms used for mapping estimation. The first makes use of the preceding optimality conditions in Corollary 2.10, while the second and the third algorithms will improve on the first to be applicable to large-scale problems. In particular, the dual solutions derived in this section will be estimated by the second algorithm, and a barycentric projection from Section 2.2 approximated for these estimates in the third algorithm.

2.4 Stochastic Plan and Mapping Estimation

We can use the optimality conditions from Corollary 2.10 to derive the *Sinkhorn algorithm*, as [CLMW21, Remark 4.9], points out. The idea of

this algorithm is twofold: apply Equations 2.1 and 2.2 to iteratively improve dual variables, and then use the last equation to obtain an estimated optimal transport plan.

For a starting value $u_2^0 \in L_{\exp}(\Omega_2)$ we define

$$T_1^{n+1}(x) := \Phi^{-1}\left(\frac{\mu(x)}{\int_{\Omega_2} \Phi\big(u_2^n(y)\big) \exp\left(\frac{-c(x,y)}{\varepsilon}\right) \mathrm{d}y}\right),$$

and

$$T_2^{n+1}(y) := \Phi^{-1}\left(\frac{\nu(y)}{\int_{\Omega_2} \Phi\big(u_1^{n+1}(x)\big) \exp\left(\frac{-c(x,y)}{\varepsilon}\right) \mathrm{d}x}\right)$$

for all $n \in \mathbb{N}$, $(x, y) \in \Omega$.

The Sinkhorn algorithm expressed in this setting is given by:

Algorithm 1: Sinkhorn Algorithm; adapted from [CLMW21, Remark 4.9]

Result: Optimal dual variables $\bar{u}_1, \bar{u}_2$
Input: $\mu \in \mathcal{P}(\Omega_1), \nu \in \mathcal{P}(\Omega_2), (x, y) \in \Omega, \varepsilon > 0$, cost function c, starting variable $u_2^0 \in L_{\exp}(\Omega_2)$

$n := 0$;
while not converged **do**
    $u_1^{n+1}(x) := T_1^{n+1}(x)$;
    $u_2^{n+1}(x) := T_2^{n+1}(y)$;
    $n := n + 1$;

Stopping after $N - 1, N > 1$, iterations, we get

$$\gamma_N(x, y) = \Phi\big(u_1^N(x)\big)\Phi\big(u_2^N(y)\big) \exp\left(\frac{-c(x, y)}{\varepsilon}\right)$$

as an approximate value of transport plan for (RP) at a predetermined value (x, y). Substituting once more with $\exp\left(\frac{\varphi_N(x)}{\varepsilon}\right) = \Phi\big(u_1^N(x)\big)$ as well as $\exp\left(\frac{\psi_N(y)}{\varepsilon}\right) = \Phi\big(u_2^N(y)\big)$, we obtain

$$\gamma_N(x, y) = \exp\left(\frac{\varphi_N(x) + \psi_N(y) - c(x, y)}{\varepsilon}\right). \tag{2.3}$$

This γ_N now functions as the approximation of an optimal transport plan for the regularized problem. To get an optimal transport map, the

barycentric projection of γ_N is calculated. In combination with Theorem 2.3 and Theorem 2.5, it is sensible to use these estimates in applications.

When considering complex problem settings however, the ever-growing dimensions of the underlying data become problematic. As [SDF+18, Section 1, Large-scale OT] points out, the Sinkhorn algorithm is only useful when considering discrete measures over small samples, because of every single iteration requiring an $\mathcal{O}(n^2)$ cost complexity. Therefore, another approach is required when the underlying measures are continuous or the dimensions of the individual samples get too large.

The procedure proposed by [SDF+18] improves on this disadvantage: instead of only computing the approximated dual variables at a specified location (x, y), a neural network is trained to model the continuous case for all values from Ω. In the discrete case, a finite vector representing $\varphi_i = \varphi(x_i)$, and $\psi_j = \psi(y_j)$ respectively, is iteratively improved.

Before continuing, we first give a basic overview on *neural networks*. This overview is based on [GBC16, Chapter 6], and [BN20, Section 6.2]. We then go on to estimate optimal transport plans and finish with the approximation of optimal transport maps with the algorithms given by [SDF+18].

Definition 2.11 (Neural Network; adapted from [GBC16, Chapter 6], and [BN20, Section 6.2])**.** A **neural network** $\mathcal{H}$ of depth $N \in \mathbb{N}$ is a collection of tuples $\cup_{n=1}^{N}(M_n, b_n, h_n)$. These tuples consist of

- (real-valued) matrices M_n,
- (real-valued) vectors b_n, and
- (real-valued) non-linear activation functions h_n.

The dimensions of these matrices, vectors and functions are chosen in such a manner, that for fixed $d, l \in \mathbb{N}$ the neural network $\mathcal{H}$ can be evaluated for an input vector $x \in \mathbb{R}^d$, and an output vector $y \in \mathbb{R}^l$ be obtained by applying the recursive formula

$$x_{n+1} = h_n\big(M_n x_n + b_n\big),\ n = 1, \ldots, N,$$

where we define $x_1 := x, y := x_{n+1}$.

A neural network with multiple layers between the input and output layers is often called a **deep neural network**. [BN20, p. 135] specifically refers to networks with an arbitrarily large number of layers, where each layer is width-bound, i.e. the dimension of each x_n, upon which an individual layer acts, is smaller than a fixed size.

Remark (Taken from [BN20, Formulae 6.48–6.52]). In practice, examples for typical activation functions are

- $h(x) = \max\{0, x\}$ (Rectifier Linear Unit, ReLU),
- $h(x) = \mathbb{1}_{(0,\infty)}(x)$ (Threshold),
- $h(x) = \frac{1}{1+\exp(-x)}$ (Logistic Sigmoid),
- $h(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$ (Arctan Sigmoid), and
- $h(x) = \frac{1}{2\pi}\int_{-\infty}^{x} \exp\left(-\frac{1}{2}x^2\right)$ (Gaussian Sigmoid).

The entries of the matrices M_n and the vectors b_n are considered as *parameters* θ, and the gradient of the neural network can then be computed with regard to these parameters. From here, we can define the notions of $\nabla\mathcal{H}$ and for f a sufficiently differentiable function $\partial_{\mathcal{H}} f$ (or $\partial_\theta f$).

We need to adapt the previous definition of F_ε from Definition 2.1 slightly to include the dual variables:

$$F_\varepsilon\big(\varphi(x), \psi(y)\big) := \exp\left(\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon}\right).$$

This allows us to state the algorithm estimating optimal dual variables.

Algorithm 2: Transport Plan Estimation; adapted from [SDF+18, Algorithm 1]

Result: Estimates of optimal dual variables φ, ψ

Input: $\mu \in \mathcal{P}(\Omega_1), \nu \in \mathcal{P}(\Omega_2), \varepsilon > 0$, cost function c, batch size p, learning rate ρ

Initialize φ_0, ψ_0;

$n := 0$;

while not converged **do**

  sample a batch $(x_1, \dots, x_p)$ from μ;

  sample a batch $(y_1, \dots, y_p)$ from ν;

  $\varphi_{n+1} := \varphi_n + \rho \sum_{i,j=1}^{p} \Big(\nabla\varphi_n(x_i) - \varepsilon\partial_{\varphi_n} F_\varepsilon\big(\varphi_n(x_i), \psi_n(y_j)\big)\nabla\varphi_n(x_i)\Big)$;

  $\psi_{n+1} := \psi_n + \rho \sum_{i,j=1}^{p} \Big(\nabla\psi_n(x_i) - \varepsilon\partial_{\psi_n} F_\varepsilon\big(\varphi_n(x_i), \psi_n(y_j)\big)\nabla\psi_n(y_j)\Big)$;

  $n := n + 1$;

To be able to recover the estimated optimal transport plan from the dual variables, we apply Equation 2.3 in analogy to the Sinkhorn algorithm. As

mentioned above, the resulting dual variables take the form of vectors, when μ and ν are discrete, and are modelled by neural networks in the continuous case.

The most apparent advantage of this approach according to [SDF+18, Section 3, Convergence rates and computational cost comparison], is that, in the case of μ and ν being discrete, the stochastic gradient descent method converges at a rate of $\mathcal{O}\big(n^{-\frac{1}{2}}\big)$, where n is the iteration number, but only requires a cost of $\mathcal{O}(p^2)$ per iteration. In the case of both measures being continuous, the cost per iteration remains $\mathcal{O}(p^2)$, but convergence can only be ensured up to a stationary point.

Lastly, we want to apply the results of Theorem 2.5 and Corollary 2.6 to obtain an estimated optimal transport map. The computation of the desired barycentric projection is however unfeasible for a large-scale problem. To deal with this constraint, [SDF+18] train a neural network once more, since this allows for the target map to be defined on all of Ω_1. Parameterizing the map T_θ as a deep neural network with the parameters θ, the following algorithm can be stated.

Algorithm 3: Transport Map Estimation; adapted from [SDF+18, Algorithm 2]

Result: Estimate T_θ of the barycentric projection $\bar{\gamma}^\varepsilon$

Input: $\mu \in \mathcal{P}(\Omega_1), \nu \in \mathcal{P}(\Omega_2), \varepsilon > 0$, cost function c, convex cost function d, batch size p, learning rate ρ

1 Initialize T_θ;

2 Compute estimates of optimal dual variables φ, ψ using Algorithm 2 with $\mu, \nu, \varepsilon, c, p$ and ρ;

3 **while** not converged **do**

4 sample a batch $(x_1, \dots, x_p)$ from μ;

5 sample a batch $(y_1, \dots, y_p)$ from ν;

6 $\theta := \theta - \rho \sum_{i,j=1}^{p} F_\varepsilon\big(\varphi(x_i), \psi(y_j)\big) \nabla_\theta \, d\big(y_j, T_\theta(x_i)\big)$;

This algorithm can also be used to compute the opposite barycentric projection g with respect to a convex cost function d on $\Omega_1 \times \Omega_1$. The last term in line 6 of Algorithm 3 should then state $d\big(g(y_j), x_i\big)$ instead. This is due to the symmetry of the optimal transport problem, as [SDF+18] points out.

This concludes the thesis, having stated two algorithms that produce an estimated optimal transport plan and map for large-scale optimal transport problems.

Conclusion

This thesis addresses the expansion of optimal transport to a large-scale setting via mapping estimation, by expanding upon the Sinkhorn algorithm using a stochastic gradient method for learning the desired optimal maps through neural networks.

An introduction to optimal transport is given in the first part. It states the classical Monge formulation of optimal transport, highlights the main disadvantages of this approach, and makes statements about the existence of solutions for atomless starting measures. In order to overcome these issues, the Monge formulation is then relaxed to the Kantorovich problem. It allows for the existence of solutions for more general input measures, as well as the application of duality theory by making use of c- and $\bar{c}$-transforms.

In the second part, entropic regularization is introduced with the idea of numerically solving optimal transport problems. With the barycentric projection, a method of recovering optimal transport maps from regularized transport plans is highlighted. Furthermore, a regularized dual of the optimal transport problem allows for the derivation of the Sinkhorn algorithm as a starting point in optimal mapping estimation. Finally, the methods proposed by [SDF+18] are used to estimate an optimal transport plan, as well as a corresponding optimal transport map, by modelling them as neural networks and applying stochastic gradient descent algorithms to train these networks.

From here, estimated optimal transport plans and maps can be used in a manifold of applications. An approach to apply optimal transport in generative adversarial networks is highlighted in [SZRM18]. Image swapping under large initial discrepancies has been investigated by [ZFW+20]. The idea of optimal transport has found use in image segmentation, e.g. [RP15]. An overview over optimal transport in general image processing is given by [Pap15]. Further foundations for application in computer science can be seen in [LS17]. More algorithmic and heuristic approaches can be found in [PC19]; they also used optimal transport to derive a solution to auction processes. [San15] has nicely incorporated economic, logistical, probabilistic, statistical and numerical interpretations into their work, by adding discussions at the

end of each section.

Connections between optimal transport and more theoretical topics can also be made. Partial differential equations and gradient flows in particular lend themselves to this, as seen in [Amb05] and [San15]. The concepts of *Wasserstein distances* and *Wasserstein spaces* play a predominant role in this.

Going forward, optimal transport in combination with large-scale mapping estimation can be applied in contexts where complexity costs of the Sinkhorn algorithm have been limiting. In addition, convergence results may be desirable, as [SDF+18] have so far only performed numerical experiments.

Additional Theorems

The following theorems and corollaries will not be explicitly proven. Direct sources which include proofs will be provided.

Theorem A.1 (Prokhorov; adapted from [San15, Box 1.4]). *Let $(\mu_n)_{n\in\mathbb{N}}$ be a tight sequence of probability measures over a complete and separable metric space X, that is for every $\varepsilon > 0$ there exists a compact subset $K \subseteq X$, such that $\mu_n(X \setminus K) < \varepsilon$ for every $n \in \mathbb{N}$. Then there exists $\mu \in \mathcal{P}(X)$ and a subsequence $(\mu_{n_k})_{k\in\mathbb{N}}$ such that $\mu_{n_k} \rightharpoonup \mu$ for $k \to \infty$. Conversely, every sequence $\mu_n \rightharpoonup \mu$ for $n \to \infty$ is necessarily tight.*

Here, $\mu_n \rightharpoonup \mu$ is meant in the sense of Definition 1.6.

Theorem A.2 (Arzela-Ascoli; adapted from [San15, Box 1.7]). *Let X be a compact metric space and $f_n : X \to \mathbb{R}$ be equicontinuous and equibounded for all $n \in \mathbb{N}$. Equicontinuous meaning that for every $\epsilon > 0$, there exists a common $\delta > 0$ such that $|f_n(x) - f_n(y)| < \epsilon$ for all pairs $x, y \in X$ with $d(x, y) < \delta$ and for all $n \in \mathbb{N}$. Equibounded referring to the property that there is a common constant C with $|f_n(x)| \leq C$ for all $x \in X$ and all $n \in \mathbb{N}$.*

Then the sequence $(f_n)_{n\in\mathbb{N}}$ admits a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ which uniformly converges to a continuous function $f : X \to \mathbb{R}$.

Theorem A.3 (Lebesgue, Dominated Convergence; adapted from [Bog07, Volume 1, Theorem 2.8.1]). *Suppose that μ-integrable functions f_n converge almost everywhere to a function f. If there exists a μ-integrable function ϕ such that $|f_n| \leq \phi$ a.e. for every n, then the function f is integrable and*

$$\lim_{n\to\infty} \int_X f_n \, \mathrm{d}\mu = \int_X f \, \mathrm{d}\mu.$$

Theorem A.4 (Banach-Alaoglu; taken from [Kes09, Theorem 5.2.1]). *Let X be a Banach space. Then, the closed unit ball in the dual space X' is weakly-* compact.*

Bibliography

[Amb05] Luigi Ambrosio. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser Basel Boston Berlin, 2005. https://doi.org/10.1007/b137080.

[AWR18] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. 2018. https://arxiv.org/abs/1705.09634.

[BN20] Ulisses Braga-Neto. *Fundamentals of Pattern Recognition and Machine Learning*. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-27656-0.

[Bog07] Vladimir Bogachev. *Measure Theory*. Springer Berlin Heidelberg, 2007. https://doi.org/10.1007/978-3-540-34514-5.

[Cam08] Lellis Camillo. Lecture notes on rectifiable sets, densities, and tangent measures. 2008. https://doi.org/10.4171/044.

[CLMW21] Christian Clason, Dirk A. Lorenz, Hinrich Mahler, and Benedikt Wirth. Entropic regularization of continuous optimal transport problems. *Journal of Mathematical Analysis and Applications*, 494(1):124432, February 2021. https://doi.org/10.1016/j.jmaa.2020.124432.

[CM94] Roberto Cominetti and Jaime San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1-3):169–187, October 1994. https://doi.org/10.1007/bf01582220.

[FCG+21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz,

Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. http://jmlr.org/papers/v22/20-451.html.

[FPPA13] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. 2013. https://arxiv.org/abs/1307.5551.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. https://www.deeplearningbook.org.

[GCPB16] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. 2016. https://arxiv.org/abs/1605.08527.

[Kan42] Leonid V. Kantorovich. On translation of mass (in russian). *Doklady. Acad. Sci. USSR*, 37(7–8):227–229, 1942.

[Kap57] Samuel Kaplan. On the second dual of the space of continuous functions. *Transactions of the American Mathematical Society*, 86(1):70–70, January 1957. https://doi.org/10.1090/s0002-9947-1957-0090774-3.

[Kes09] Srinivasan Kesavan. *Functional Analysis*. Hindustan Book Agency, 2009. https://doi.org/10.1007/978-93-86279-42-2.

[LS17] Bruno Lévy and Erica Schwindt. Notions of optimal transport theory and how to implement them on a computer. 2017. https://arxiv.org/abs/1710.02634.

[Mon81] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

[Pap15] Nicolas Papadakis. *Optimal Transport for Image Processing*. Habilitation, Université de Bordeaux, 2015. https://hal.archives-ouvertes.fr/tel-01246096.

[PC19] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. https://arxiv.org/abs/1803.00567.

[RP15] Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. 2015. https://arxiv.org/abs/1503.01986.

[RY08] Bryan P. Rynne and Martin A. Youngson. *Linear Functional Analysis*. Springer London, 2008. https://doi.org/10.1007/978-1-84800-005-6.

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-20828-2.

[SDF+18] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. 2018. https://arxiv.org/abs/1711.02283.

[SZRM18] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. 2018. https://arxiv.org/abs/1803.05573.

[Vil09] Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009. https://doi.org/10.1007/978-3-540-71050-9.

[ZFW+20] Hao Zhu, Chaoyou Fu, Qianyi Wu, Wayne Wu, Chen Qian, and Ran He. AOT: Appearance optimal transport based identity swapping for forgery detection. 2020. https://arxiv.org/abs/2011.02674.