A Sinkhorn-NN Hybrid Algorithm for Optimal Transport

Jonathan Geuter

September 24, 2022

Master's Thesis

Technische Universität Berlin

First Examiner: Dr. Vaios Laschos, WIAS Berlin
Second Examiner: Prof. Dr. Martin Skutella, TU Berlin

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfs-mittel angefertigt habe.

Jonathan Geuter

Datum, Ort

Abstract

The Sinkhorn algorithm [8] is the state-of-the-art to compute optimal transport distances between discrete probability distributions, using an entropic regularizer added to the optimal transport problem. The entropic problem being a convex optimization problem, the algorithm is guaranteed to converge, no matter its initialization. This lead to little attention being paid to initializing it, and simple vectors like the vector with all entries equal to 1 are common choices. We present a Sinkhorn-NN hybrid algorithm, in which a pretrained neural network predicts an approximation of the optimal potential of the optimal transport dual problem given two distributions, which can then be used to compute a starting vector for the Sinkhorn algorithm. The network is universal in the sense that it is able to generalize to any pair of distributions of fixed dimension. We show that this initialization can significantly accelerate convergence of the Sinkhorn algorithm.

A PyTorch implementation can be found at `https://github.com/j-geuter/DualOTComputations`.

Deutsche Zusammenfassung

Short Summary in German

Optimal Transport oder, auf Deutsch, *Optimaler Transport* ist ein Teilgebiet der Mathematik, das sich mit dem Transport zwischen Wahrscheinlichkeitsverteilungen beschäftigt. Gegeben zwei polnische Wahrscheinlichkeitsräume $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ mit ihren jeweiligen Borel-σ-Algebren und eine Kostenfunktion $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$, lautet das *Kantorovich-Problem*

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

wobei $\Pi(\mu, \nu)$ die Menge aller *Transportpläne* enthält, das heißt aller Wahrscheinlichkeitsmaße γ auf dem Produktraum $\mathcal{X} \times \mathcal{Y}$, sodass $\gamma \circ \pi_{\mathcal{X}}^{-1} = \mu,\ \gamma \circ \pi_{\mathcal{Y}}^{-1} = \nu$. Das duale Problem lautet

$$\sup \left\{ \int_{\mathcal{X}} \psi(x)\, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \varphi(y)\, \mathrm{d}\nu(y) :\ \psi \in L^1(\mu),\ \varphi \in L^1(\nu),\ \psi + \varphi \leq c \right\},$$

und beide Probleme besitzen stets Optimierer, deren Optima übereinstimmen. Diese Optimierer zu berechnen ist jedoch recht rechenaufwändig, in hohen Dimensionen auch für den Fall, dass beide Maße diskret sind. Ein effizienter Algorithmus zur Approximation des Optimums und auch eines optimalen Transportplans ist der Sinkhornalgorithmus [8]. Dieser konvergiert gegen die Lösung des entropisch regularisierten Problems

$$\min_{\gamma} \left\{ \langle \gamma, c \rangle - \varepsilon H(\gamma) :\ \gamma \circ \pi_{\mathcal{X}}^{-1} = \mu,\ \gamma \circ \pi_{\mathcal{Y}}^{-1} = \nu \right\},$$

wobei $\varepsilon > 0$ ein Regularisierungskoeffizient ist und $H(\gamma) := -\sum_{ij} \gamma_{ij}(\log \gamma_{ij} - 1)$ die Entropie des Transportplans γ. Der Sinkhornalgorithmus ist ein iteratives Näherungsverfahren, das mit einem Startvektor initialisiert wird. Da die Konvergenz unabhängig vom Startvektor garantiert ist, wurde einer spezifizierten Initialisierung bisher wenig Beachtung geschenkt. Wir zeigen, dass eine gut gewählte Initialisierung die Konvergenzgeschwindigkeit drastisch verbessern kann. Dazu stellen wir unseren *Sinkhorn-NN hybrid algorithm* vor – ein Hybrid aus einem neuronalen Netz und dem Sinkhornalgorithmus. Wir trainieren ein Netz so, dass es ein optimales Potential f des diskreten dualen Transportproblems

$$\max \{ \langle f, \alpha \rangle + \langle g, \beta \rangle :\ f \in \mathbb{R}^m,\ g \in \mathbb{R}^n,\ f + g \leq c \}$$

gegeben α und β (α und β entsprechen hier Vektorrepräsentationen der Maße μ und ν, d.h. $\alpha_i = \mu(x_i)$, $\beta_j = \nu(y_j)$) approximieren kann, und zeigen, wie sich daraus ein Startvektor für den Sinkhornalgorithmus bestimmen lässt, der die Konvergenzgeschwindigkeit verglichen mit einem üblichen Startvektor deutlich verbessert. Die Arbeit enthält eine ausführliche Einführung in die Thematik Optimal Transport, inklusive eines Überblicks über das diskrete Transportproblem, den sogenannten *Wassersteindistanzen* und einer detaillierten Erklärung des Sinkhornalgorithmus. Im Anschluss werden die Details unseres Sinkhorn-NN-Algorithmus erläutert und Ergebnisse verschiedener Experimente präsentiert, die abschließend interpretiert werden.

Contents

1 Introduction **6**

2 Notation **8**

3 Optimal Transport **11**
3.1 The Monge Problem 11
3.2 The Kantorovich Problem 13
3.3 c-Transforms and the Dual Problem 17
3.4 Fundamental Theorem of Optimal Transport 23
3.5 Duality Theorem 27
3.6 Wasserstein Distances 28
3.7 Discrete Optimal Transport 31

4 Sinkhorn Algorithm **34**
4.1 Entropic Optimal Transport 34
4.2 Sinkhorn Algorithm 37
4.3 Initializing Sinkhorn's Algorithm 39

5 A Trained Initialization for the Sinkhorn Algorithm **40**
5.1 Sinkhorn-NN Hybrid Algorithm 40
5.2 Training Data 41
5.3 Network Architecture 43
5.4 Why Not...? 44

6 Results **46**

7 Discussion **47**

A Appendix **48**
A.1 Measure Theory 48

References **52**

Index **54**

1 Introduction

Optimal Transport is what it sounds like - the theory of optimally transporting something somewhere. French mathematician Gaspard Monge laid the foundation of optimal transport in the 18^{th} century. In his 1781 publication *Mémoire sur la théorie des déblais et des remblais* [17], he considers the following problem: Assume you extract soil from various places, and this soil needs to be transported to various other places, e.g. construction sites. You know how much soil you extract in each location, as well as how much is needed at each construction site. You also know how much it costs you to transport a certain amount of soil from a to b. What you are looking for is a *transport plan*, i.e. an assignment that tells you how much soil to transport from each extraction point to each construction site. In mathematical terms, this reads as follows: Given two Polish probability spaces $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$, equipped with their Borel-σ-algebras, and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$, the task is to find a measurable *transport map* $T : \mathcal{X} \to \mathcal{Y}$ which "transports" mass from the measure μ to mass from the measure ν, meaning we require $\nu = \mu \circ T^{-1}$, while minimizing the total cost

$$\int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\mu(x).$$

While this formulation is very intuitive and simple, it has one major drawback: There is no guarantee that such an optimizer exists. There may not even exist any transport map at all - consider the case where μ is a Dirac measure and ν is not. The formulation of the problem which is most common today is a relaxation of Monge's original formulation, and was derived by Soviet mathematician Leonid Vitaliyevich Kantorovich. The crucial change Kantorovich proposed was the following: Instead of requiring the existence of a transport map - which means that given a location where you extract soil, you can deterministically determine a single construction site this soil is transported to - we are now only interested in a *transport plan*, which allows for splitting up the soil to be transported to different construction sites. Mathematically speaking, this means we try to find a measure γ on $\mathcal{X} \times \mathcal{Y}$ which admits μ and ν as its marginals on $\mathcal{X}$ and $\mathcal{Y}$, i.e.

$$\gamma \circ \pi_{\mathcal{X}}^{-1} = \mu, \quad \gamma \circ \pi_{\mathcal{Y}}^{-1} = \nu$$

(where π denotes the projection), and minimizes

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

This problem has a dual problem:

$$\sup \left\{ \int_{\mathcal{X}} \psi(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \varphi(y) \, \mathrm{d}\nu(y) : \ \psi \in L^1(\mu), \ \varphi \in L^1(\nu), \ \psi + \varphi \leq c \right\},$$

and both problems admit optimal solutions and their optimal values coincide; however, computing these solutions tends to be computationally expensive, particularly in higher dimensions. In the discrete setting, i.e. where $\mathcal{X}$ and $\mathcal{Y}$ are both finite, an efficient way to compute an approximation of the solution is the Sinkhorn algorithm [8], an iterative algorithm converging to the solution of the *entropic optimal transport problem*, which consists of adding an entropic regularizer to the Kantorovich

problem:

$$\min_{\gamma} \left\{ \langle \gamma, c \rangle - \varepsilon H(\gamma) : \ \gamma \circ \pi_{\mathcal{X}}^{-1} = \mu, \ \gamma \circ \pi_{\mathcal{Y}}^{-1} = \nu \right\}$$

where $\varepsilon > 0$ is a regularizing coefficient and $H(\gamma) := -\sum_{ij} \gamma_{ij}(\log \gamma_{ij} - 1)$ the entropy of γ. This problem can be shown to be ε-strongly convex, hence it admits a unique solution, and the Sinkhorn algorithm is guaranteed to converge to this solution. However, carefully choosing a starting vector for the algorithm can significantly improve its convergence speed. This is why we suggest a *Sinkhorn-NN hybrid algorithm* (NN signifying *neural network*), where a neural network is pretrained to predict an optimal potential f of the discrete dual problem

$$\max_{\substack{f \in \mathbb{R}^m, \ g \in \mathbb{R}^n \\ f+g \leq c}} \langle f, \alpha \rangle + \langle g, \beta \rangle$$

given α and β, where α and β are vector representations of μ and ν (i.e. $\alpha_i = \mu(x_i)$, $\beta_j = \nu(y_j)$). We will show how this dual potential can be used to compute a starting vector for the Sinkhorn algorithm, and that this approach significantly improves the convergence speed of the Sinkhorn algorithm compared to a fixed initialization commonly used.

The thesis is structured as follows: in section 2, all notation used throughout the thesis is defined. The following section 3 is devoted to a thorough introduction to optimal transport. The Monge and Kantorovich problems are defined, and two major theorems – the fundamental theorem of optimal transport 3.4.1 and the duality theorem 3.5.1 – are proven. Additionaly, the well-known *Wasserstein distances* are defined, and we will see what the optimal transport problem looks like in the discrete case. Section 4 features the entropic optimal transport problem and the Sinkhorn algorithm. In section 5, we will discuss the details of our algorithm and its implementation, such as the training data and network structure we used. Experiments and results will be presented in section 6. A final discussion, interpreting the results and outlining the scope and limits of the idea presented, can be found in section 7. The appendix A contains some basics omitted during the thesis and further explanations.

2 Notation

In this section, we list some notations that will be used throughout the thesis. Definitions and results corresponding to these notations can be found in the appendix, section A. Also, we will mention some conventions that will be used throughout the thesis.

- $\mathbb{N} := \{0, 1, 2, 3, ...\}$
- $\mathbb{N}_{>0} := \{1, 2, 3, 4, ...\}$
- for $m, n \in \mathbb{N}_{>0}$, $m \leq n$, $[\![m, n]\!] := \{m, m+1, ..., n\}$ and $[\![n]\!] := [\![1, n]\!]$
- For $r \in \mathbb{R}$: $[0, r] := \{x \in \mathbb{R} : 0 \leq x \leq r\}$, $[0, r) := \{x \in \mathbb{R} : 0 \leq x < r\}$, etc.
- S_n with $n \in \mathbb{N}_{>0}$ denotes the set of all permutations of $[\![n]\!]$

For $n \in \mathbb{N}_{>0}$, I_n denotes the identity in $\mathbb{R}^{n\times n}$ and $1_n \in \mathbb{R}^n$ the n-dimensional vector with all entries equal to 1. Similarily, by $0_n \in \mathbb{R}^n$ we denote the 0-vector. If it is clear what space is meant, we will sometimes write 0 instead. Let $\Delta^n = \{v \in \mathbb{R}^n_{\geq 0} : \sum_i v_i = 1\}$ be the n-dimensional probability simplex and $\Delta^n_{>0} = \{v \in \Delta^n : v_i > 0 \text{ for all } i \in [\![n]\!]\}$. For a matrix $A \in \mathbb{R}^{m\times n}$, a_{ij} refers to the entry in the i^{th} row and j^{th} column. For $m, n, k, l \in \mathbb{N}_{>0}$, $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{k\times l}$, $A \otimes B \in \mathbb{R}^{mk\times nl}$ denotes the *Kronecker product* of A and B, i.e. the matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix}.$$

For two vectors $a, b \in \mathbb{R}^n$, $\langle a, b\rangle$ is the usual scalar product and $\mathrm{diag}(a) \in \mathbb{R}^{n\times n}$ the matrix with diagonal entries $\mathrm{diag}(a)_{ii} = a_i$ and all other entries equal to 0. For matrices $A, B \in \mathbb{R}^{m\times n}$, we use $\langle \cdot, \cdot \rangle$ to denote the *Frobenius dot-product*:

$$\langle A, B\rangle := \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ij}.$$

Now, let $(\mathcal{X}, d)$ be a metric space and $x \in \mathcal{X}$. For $r > 0$, we denote by $B_r(x)$ the open ball of radius r around x and by $\overline{B_r(x)}$ its closure. More generally, for any set $A \subset \mathcal{X}$, we will denote its closure by $\overline{A}$. A *neighbourhood* of a point $x \in \mathcal{X}$ is a set $V \subset \mathcal{X}$ containing an open set U such that $x \in U \subset V$. $\mathcal{X}$ is called *totally bounded* if for any $\varepsilon > 0$, we can cover $\mathcal{X}$ by finitely many open balls of radius ε. A *Polish space* is a complete, separable metric space.[1] We will oftentimes deal with the product space of two Polish spaces (which is again a Polish space), each equipped with its own σ-algebra. The σ-algebra on the product space will then be the product-σ-algebra.[2]

[1] Note that some authors define a Polish space to be a separable, completely metrizable topological space, which is a space homeomorphic to a separable, complete metric space.

[2] More details on the product of two Polish spaces and its σ-algebra can be found in the appendix, see e.g. remark A.9, where we also explain why it does not matter whether we choose the product σ-algebra or the σ-algebra generated by the product topology on the product space of two Polish spaces.

For any set X, by Id_X we refer to the identity function on X. If it is clear what identity function is meant, we will sometimes only write Id.
If Y is another set, then $\pi_{\mathcal{X}}$ is the projection $X \times Y \to X$, $(x, y) \mapsto x$.
Let $(\mathcal{X}, d)$ be a metric space. A function $f : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ is called *lower semicontinuous* if

$$f(x_0) \leq \liminf_{x \to x_0} f(x) \quad \text{for all } x \in \mathcal{X}.$$

The support $\mathrm{supp}(f)$ of a function $f : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ is the set $\overline{\{x \in \mathcal{X} : f(x) \neq 0\}}$.
By $C(\mathcal{X})$ we denote the space of all continuous functions $f : \mathcal{X} \to \mathbb{R}$, and the space of all continuous and bounded functions is denoted by $C_b(\mathcal{X})$.

For a topological space $\mathcal{X}$, its Borel σ-algebra is denoted by $\mathcal{B}(\mathcal{X})$. A measure μ on a measurable space $(\mathcal{X}, \mathcal{A})$ is called *Borel measure* if $\mathcal{B}(\mathcal{X}) \subset \mathcal{A}$, and *finite* if $\mu(\mathcal{X}) < \infty$.
For $A \in \mathcal{B}(\mathcal{X})$, the indicator function $\mathbb{1}_A : \mathcal{X} \to \mathbb{R}$ is defined via

$$\mathbb{1}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}.$$

The *Dirac measure* at $x \in \mathcal{X}$ is defined as $\delta_x : \mathcal{B}(\mathcal{X}) \to \mathbb{R}$, $\delta_x(A) = \mathbb{1}_A(x)$.
Throughout the thesis, **we will always consider all measure spaces to be Polish probability spaces, equipped with their Borel σ-algebra**, unless stated otherwise. Also, **all functions on measure spaces will always be assumed to be measurable** (in cases where there are no measures involved, for example for functions defined on mere sets, this is of course not assumed).
The set of all Borel probability measures on $\mathcal{X}$ will be denoted by $P(\mathcal{X})$. A set $N \subset \mathcal{X}$ is said to be μ-negligible if it is contained in a Borel set of measure 0. A measure μ on $\mathcal{X}$ is said to be *concentrated* on $C \subset \mathcal{X}$ if $\mathcal{X} \setminus C$ is μ-negligible. The support $\mathrm{supp}(\mu)$ of $\mu \in P(\mathcal{X})$ is the smallest closed set on which μ is concentrated.
If T is a map $\mathcal{X} \to \mathcal{Y}$ and μ a measure on $\mathcal{X}$, then the *pushforward measure* of μ by T is the measure $\mu \circ T^{-1}$ on $\mathcal{Y}$.[3]
The weak topology on $P(\mathcal{X})$ is induced by convergence against functions in $C_b(\mathcal{X})$, i.e. bounded and continuous test functions. More explicitly, a sequence $(\gamma_n)_{n \in \mathbb{N}} \subset P(\mathcal{X})$ is said to converge to $\gamma \in P(\mathcal{X})$ (weakly), if for all $f \in C_b(\mathcal{X})$, we have

$$\lim_{n \to \infty} \int_{\mathcal{X}} f(x) \, \mathrm{d}\gamma_n(x) = \int_{\mathcal{X}} f(x) \, \mathrm{d}\gamma(x).$$

A fact worth noting here and one that will be used throughout the thesis is that integration against bounded and continuous test functions uniquely defines a measure (cmp. lemma A.16).
A set $M \subset P(\mathcal{X})$ is called *tight* if for any $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subset \mathcal{X}$ such that for all $\mu \in M$, we have $\mu(\mathcal{X} \setminus K_\varepsilon) < \varepsilon$.
If $\mathcal{S}_1$ is a measurable space and $X : \Omega \to \mathcal{S}_1$ is a random variable defined on a probability space $(\Omega, \mathbb{P})$,

[3] there are many different notations in the literature for the pushforward measure, including $T_{\#}\mu$, $T\#\mu$, $T(\mu)$, $T\mu$, or μT^{-1}.

its pushforward measure $\mathbb{P} \circ X^{-1}$ on the image space $\mathcal{S}_1$ is also called the *law* of X and will be denoted by $\mathcal{L}(X)$. Similarily, if $\mathcal{S}_2$ is another measurable space and $Y : \Omega \to \mathcal{S}_2$ is another random variable defined on $(\Omega, \mathbb{P})$, the pushforward measure $\mathbb{P} \circ (X, Y)^{-1}$ of $(X, Y) : \Omega \to \mathcal{S}_1 \times \mathcal{S}_2$ will be denoted by $\mathcal{L}(X, Y)$.

Sometimes we will make statements like "for all $x \in \mathcal{X}$, we have $c(x) \leq a(x)$", where $a \in L^1(\mu)$ for some measure μ on $\mathcal{X}$. Of course, a is not defined point-wise, hence statements like this are to be understood as "μ-almost surely, we have...".

3 Optimal Transport

In this chapter, we introduce the optimal transport problem in its two well-known formulations, the *Monge* and *Kantorovich Problem*, in sections 3.1 and 3.2. We will then derive a dual formulation of the Kantorovich Problem in section 3.3 and get to know the concepts of *c-cyclical monotonicity*, *c-concavity*, and *c-transforms*. Leveraging these new concepts, we prove the *Fundamental Theorem of Optimal Transport* in section 3.4. A duality theorem can easily be derived, as is shown in section 3.5. Section 3.6 deals with the case where the source and target space are the same, which gives rise to the so-called *Wasserstein distances*. Finally, in section 3.7, we will focus on the special case where both the source and target distributions are discrete. Amongst many other applications, this is the case when considering distributions that are derived from image data, where the pixels can be taken to be a discrete metric space.
In particular sections 3.1–3.3 and 3.6 are based on [25].
Be reminded we are *always* considering Polish probability spaces equipped with their Borel-σ-algebra if not stated otherwise. Also note that some properties of Polish spaces will be used implicitly, such as the fact that the product σ-algebra of two Polish spaces is the same as the Borel σ-algebra on the product space, or the fact that the product of two Polish spaces is again a Polish space. See the appendix, in particular remark A.9, for more details.

3.1 The Monge Problem

The most basic structure we will use time and time again is the so-called *coupling*, something we already got to know in the introduction.

Definition 3.1.1 (Coupling)**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces. Let $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ be two random variables on a probability space $(\Omega, \mathbb{P})$ such that their laws are equal to μ and ν, i.e. $\mathcal{L}(X) = \mu$, $\mathcal{L}(Y) = \nu$. Then (X, Y) is called a* coupling *of μ and ν. Oftentimes, the joint law $\mathcal{L}(X, Y)$ is also referred to as a coupling.*

A coupling can be seen as transforming the measure μ into the measure ν, or, put differently, transporting mass from μ to ν. Hence, couplings are also called *transport plans*. This gets more clear by realizing that coupling μ and ν is nothing else but constructing a measure γ on $\mathcal{X} \times \mathcal{Y}$ which admits μ and ν as its *marginals*, meaning:

$$\gamma \circ \pi_{\mathcal{X}}^{-1} = \mu, \quad \gamma \circ \pi_{\mathcal{Y}}^{-1} = \nu.$$

(Indeed, note if we are given a coupling (X, Y), such a measure γ is given by the joint law of (X, Y) as for any $A \in \mathcal{B}(\mathcal{X})$, we have

$$\mathcal{L}(X, Y) \circ \pi_{\mathcal{X}}^{-1}(A) = \mathcal{L}(X, Y)(A \times \mathcal{Y}) = \mathcal{L}(X)(A) = \mu(A)$$

which is equivalent to the marginal condition on $\mathcal{X}$, and the marginal condition on $\mathcal{Y}$ follows in the same way). The set of all such γ is denoted by $\Pi(\mu, \nu)$. We will refer to measures in $\Pi(\mu, \nu)$ as couplings as well.

Remark 3.1.2. For a probability measure γ on $\mathcal{X} \times \mathcal{Y}$, the following conditions are equivalent to γ being a coupling of μ and ν:

1. For all measurable sets $A \in \mathcal{B}(\mathcal{X})$ and $B \in \mathcal{B}(\mathcal{Y})$ it holds $\gamma(A \times \mathcal{Y}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu(B)$.

2. For all $(\phi, \psi) \in L^1(\mu) \times L^1(\nu)$ it holds

$$\int_{\mathcal{X}\times\mathcal{Y}} \phi(x) + \psi(y) \, \mathrm{d}\gamma(x, y) = \int_{\mathcal{X}} \phi(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \, \mathrm{d}\nu(y).$$

3. For all $(\phi, \psi) \in C_b(\mu) \times C_b(\nu)$ it holds

$$\int_{\mathcal{X}\times\mathcal{Y}} \phi(x) + \psi(y) \, \mathrm{d}\gamma(x, y) = \int_{\mathcal{X}} \phi(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \, \mathrm{d}\nu(y).$$

   This is an immediate consequence of lemma A.16 and a fact that we will use again later, as integration against functions in C_b is what defines weak convergence.

Remark 3.1.3. Note that there always exists a coupling between any measures μ and ν: Simply set $\gamma = \mu \otimes \nu$, which is also called the *trivial coupling*. This means the corresponding random variables X and Y are independent. In this case, knowledge of X does not provide any information about Y. The other extreme case is the following.

Definition 3.1.4 (Deterministic Coupling)**.** *A coupling (X, Y) of μ and ν is called* deterministic coupling *if there exists a map $T : \mathcal{X} \to \mathcal{Y}$ s.t. $Y = T(X)$. T is called a* transport map*.*

Remark 3.1.5. In terms of measures, the equivalent definition is the following: A coupling $\gamma \in \Pi(\mu, \nu)$ is deterministic if there exists a map $T : \mathcal{X} \to \mathcal{Y}$ such that $\gamma = \mu \circ (\mathrm{Id}, T)^{-1}$.
Note that for deterministic couplings, ν is given as the push-forward measure of μ by T, as for any $B \in \mathcal{B}(\mathcal{Y})$ we have

$$\mu \circ T^{-1}(B) = \mu((\mathrm{Id}, T)^{-1}(\mathcal{X} \times B)) = \gamma(\mathcal{X} \times B) = \nu(B).$$

With this definition at hand, we are now able to precisely define the Monge problem. As we have seen in the introduction, we are interested in transport maps that are optimal with respect to a given cost function c.
In the Monge problem, we integrate the cost function with respect to a deterministic coupling γ and optimize over all such deterministic couplings. As deterministic couplings have the property that they concentrate the mass on the graph of a function T, this problem can be more conveniently and intuitively formulated as follows:

Problem 3.1.6 (Monge Problem)**.** Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ a cost function. The *Monge Problem* is defined as:

$$\inf_T \left\{ \int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\mu(x) : \quad T : \mathcal{X} \to \mathcal{Y}, \ \nu = \mu \circ T^{-1} \right\}.$$

Example 3.1.7. Consider the following simple example illustrating the concept of transport maps. Let λ be the Lebesgue measure on $\mathbb{R}$, $\mathcal{X} = \mathcal{Y} = [0, n + 1]$, and $c(x, y) = |x - y|^p$ for some $p > 0$

(building a cost function using a distance is very common; we will see this again later when introducing *Wasserstein distances*). Let $\mu = \frac{1}{n}\lambda|_{[0,n]}$ and $\nu = \frac{1}{n}\lambda|_{[1,n+1]}$ be the uniform distributions on $[0, n]$ and $[1, n+1]$ resp. for some $n > 1$. Consider the following transport maps:

$$T_1(x) := x + 1, \qquad T_2(x) := \begin{cases} x + n, & x \in [0, 1], \\ x, & x \in (1, n]. \end{cases}$$

The corresponding transport costs are:

$$\int_{\mathcal{X}} c(x, T_1(x)) \, \mathrm{d}\mu(x) = \int_0^n |x - (x+1)|^p \, \mathrm{d}\mu(x) = \frac{1}{n} \int_0^n 1 \, \mathrm{d}x = 1$$

and

$$\int_{\mathcal{X}} c(x, T_2(x)) \, \mathrm{d}\mu(x) = \int_0^1 |x - (x+n)|^p \, \mathrm{d}\mu(x) = \frac{1}{n} \int_0^1 n^p \, \mathrm{d}x = n^{p-1}.$$

Hence, we can see that T_1 yields a better transport cost if and only if $p > 1$, and T_2 if and only if $p < 1$. For $p = 1$, the transport costs are the same. Intuitively this makes sense, as T_1 moves all mass along $\mathbb{R}$ equally, whereas T_2 leaves as much mass as possible in place while only moving some from the "start" to the "end", which should not be favourable if transport costs grow faster than linearly (i.e. $p > 1$), while being favourable if the opposite is the case.

As we have seen in the introduction, the Monge problem faces a serious drawback: transport maps between measures need not exist. In the following section, we will get to know a famous relaxation of this problem, the *Kantorovich Problem*.

3.2 The Kantorovich Problem

Problem 3.2.1 (Kantorovich Problem)**.** Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ a cost function. The *Kantorovich Problem* is defined as:

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

Remark 3.2.2. As we will see later on, there exists a dual problem to the Kantorovich problem. That's why problem 3.2.1 is also referred to as the *primal problem of optimal transport*.

Definition 3.2.3 (Optimal Transport Plan)**.** *A transport plan $\gamma \in \Pi(\mu, \nu)$ which achieves the infimum in problem 3.2.1 is called an* optimal transport plan.

As any transport map T induces a transport plan via $\gamma = \mu \circ (\mathrm{Id}, T)^{-1}$ (see remark 3.1.5), this is indeed a relaxation of Monge's problem. The important difference is the following: Whereas in the Monge problem, mass from $x \in \mathcal{X}$ gets transported entirely to $T(x) \in \mathcal{Y}$, the Kantorovich problem allows for splitting the mass.
This relaxation comes with many nice properties. For example, under very mild assumptions on the cost function, we can guarantee the existence of an optimal transport plan. In the following, we will prove this result. The proof makes use of *Prokhorov's Theorem* (see A.13): A subset $\mathcal{P} \subset P(\mathcal{X})$ has compact closure with respect to the weak topology if and only if it is tight. To apply it, we will need

tightness of $\Pi(\mu,\nu)$, which is what the following lemma gives us (in combination with the fact that $\{\mu\}$ and $\{\nu\}$ are tight subsets of $P(\mathcal{X})$ and $P(\mathcal{Y})$ resp., which we will prove soon).

Lemma 3.2.4 (Tightness of Transport Plans)**.** *Let $\mathcal{P} \subset P(\mathcal{X})$ and $\mathcal{Q} \subset P(\mathcal{Y})$ be two tight subsets of $P(\mathcal{X})$ and $P(\mathcal{Y})$ respectively. Then the set of all transport plans in $\mathcal{P}$ and $\mathcal{Q}$, namely $\Pi(\mathcal{P},\mathcal{Q}) := \bigcup_{\mu\in\mathcal{P},\ \nu\in\mathcal{Q}} \Pi(\mu,\nu)$, is tight in $P(\mathcal{X}\times\mathcal{Y})$.*

Proof. Let $\varepsilon > 0$. By assumption, there exist compact sets $K_\varepsilon \subset \mathcal{X}$ and $L_\varepsilon \subset \mathcal{Y}$ such that

$$\mu(\mathcal{X}\setminus K_\varepsilon) \le \frac{\varepsilon}{2} \text{ for all } \mu\in\mathcal{P}; \quad \nu(\mathcal{Y}\setminus L_\varepsilon) \le \frac{\varepsilon}{2} \text{ for all } \nu\in\mathcal{Q}.$$

Now let $\mu\in\mathcal{P}$ and $\nu\in\mathcal{Q}$ be two measures and $\gamma\in\Pi(\mu,\nu)$ a transport plan. Then:

$$\gamma(\mathcal{X}\times\mathcal{Y}\setminus(K_\varepsilon\times L_\varepsilon)) \le \mu(\mathcal{X}\setminus K_\varepsilon) + \nu(\mathcal{Y}\setminus L_\varepsilon) \le \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon$$

and the claim follows from the fact that $K_\varepsilon \times L_\varepsilon$ is again compact in $\mathcal{X}\times\mathcal{Y}$. □

The main idea in the proof will be to use *Weierstraß' Theorem* (A.12) on the functional

$$F : \Pi(\mu,\nu) \to \mathbb{R}\cup\{+\infty\},\ \gamma \mapsto \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\gamma.$$

(Note F cannot take the value $-\infty$ as we will only consider cost functions bounded from below.) Once we know that F is lower semicontinuous, this will yield a minimizer.

We will apply the previous lemma to the sets $\{\mu\} \subset P(\mathcal{X})$ and $\{\nu\} \subset P(\mathcal{Y})$. While it may seem intuitive that these sets are tight in Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, it requires a little bit of work to prove. First, we establish a well-known fact from topology.

Lemma 3.2.5. *Let $(\mathcal{X},d)$ be a complete, metric space and $K\subset\mathcal{X}$ be a closed subset. Then K is compact if and only if it is totally bounded.*

Proof. This proof is loosely based on Lemma 3.1 in [19].
First, let K be compact. Then for any $\varepsilon>0$, $\bigcup_{x\in K} B_\varepsilon(x)$ is an open cover of K and hence has a finite subcover. This shows that K is totally bounded.
Now, let K be totally bounded. We show that any sequence in K contains a convergent subsequence, from which compactness follows (cmp. proposition A.11). Thus, let $(x_n)_{n\in\mathbb{N}_{>0}}$ be a sequence in K. Now we construct a subsequence $(x_{n_k})_{k\in\mathbb{N}_{>0}}$ inductively: For $m=1$, we can cover K with finitely many balls of radius 1, hence one of them, B_1, contains infinitely many x_n. Let $N_1 := \{n\in\mathbb{N}_{>0} : x_n \in B_1\}$ and $n_1\in N_1$. Now assume for some $l\in\mathbb{N}_{>0}$ we have a ball B_l of radius $\frac{1}{l}$ such that $N_l := \{n > n_{l-1} :\ x_n \in \bigcap_{i=1}^{l} B_i\}$ is infinite (where $n_{l-1}\in B_{l-1}$). Choose $n_l\in N_l$. As K is totally bounded, we can cover $\bigcap_{i=1}^{l} B_i$ with finitely many balls of radius $\frac{1}{l+1}$, one of which contains infinitely many x_n. Call this ball B_{l+1}. Then also $N_{l+1} := \{n > n_l :\ x_n \in \bigcap_{i=1}^{l+1} B_i\}$ is infinite. This construction results in a sequence $(x_{n_k})_{k\in\mathbb{N}_{>0}}$ for which $x_{n_l}\in B_k$ for all $l\ge k$. Hence, it is a Cauchy sequence in K and as $\mathcal{X}$ is complete, it converges. Since K is closed by assumption, its limit point lies in K. □

Lemma 3.2.6. *For a Polish space $\mathcal{X}$, any $\mu\in P(\mathcal{X})$ is tight (viewed as the set $\{\mu\}$).*

Proof. This proof follows that of Theorem 3.2 in [19].
Let $\varepsilon > 0$. We need to show that there exists a compact set $K \subset \mathcal{X}$ such that $\mu(\mathcal{X} \setminus K) \leq \varepsilon$. Let $\{a_1, a_2, ...\}$ be a dense subset of $\mathcal{X}$. For any $m \in \mathbb{N}_{>0}$, there exists an integer n_m such that

$$\mu\left(\bigcup_{i=1}^{n_m} B_{\frac{1}{m}}(a_i)\right) > \mu(\mathcal{X}) - \frac{\varepsilon}{2^m}.$$

Let

$$K := \bigcap_{m=1}^{\infty} \bigcup_{i=1}^{n_m} \overline{B_{\frac{1}{m}}(a_i)}.$$

Then K is closed. We now show that K is totally bounded. Let $\delta > 0$ and choose m such that $\frac{1}{m} < \delta$. Then

$$K \subset \bigcup_{i=1}^{n_m} \overline{B_{\frac{1}{m}}(a_i)} \subset \bigcup_{i=1}^{n_m} B_\delta(a_i).$$

Hence, K is compact by lemma 3.2.5. Furthermore,

$$\begin{aligned}\mu(\mathcal{X} \setminus K) &= \mu\left(\bigcup_{m=1}^{\infty}\left(\mathcal{X} \setminus \bigcup_{i=1}^{n_m} \overline{B_{\frac{1}{m}}(a_i)}\right)\right) \leq \sum_{m=1}^{\infty} \mu\left(\mathcal{X} \setminus \bigcup_{i=1}^{n_m} \overline{B_{\frac{1}{m}}(a_i)}\right) \\ &= \sum_{m=1}^{\infty}\left(\mu(\mathcal{X}) - \mu\left(\bigcup_{i=1}^{n_m} \overline{B_{\frac{1}{m}}(a_i)}\right)\right) < \sum_{m=1}^{\infty} \frac{\varepsilon}{2^m} = \varepsilon.\end{aligned}$$

□

A well-known fact about lower semicontinuous functions is that if they are bounded from below, they can be approximated from below by a sequence of continuous functions.

Lemma 3.2.7. *Let $(\mathcal{X}, d)$ be a metric space and $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be lower semicontinuous and bounded from below. Then there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of continuous functions $f_n : \mathcal{X} \to \mathbb{R}$ that converge to f pointwise from below.*

Proof. Set

$$f_n(x) := \inf_{y\in\mathcal{X}} \{f(y) + nd(x, y)\}.$$

Then all f_n are continuous, as every map $x \mapsto f(y) + nd(x, y)$ for fixed $y \in \mathcal{X}$ is continuous. Furthermore, it is clear that $f_0 \leq f_1 \leq ... \leq f$, as $f(x) = f(x) + nd(x, x) \geq \inf_y\{f(y) + nd(x, y)\}$ for all $x \in \mathcal{X}$ and all $n \in \mathbb{N}$. Hence, for fixed $x \in \mathcal{X}$, $\lim_{n\to\infty} f_n(x)$ exists and $\lim_{n\to\infty} f_n(x) \leq f(x)$. To finish the proof, it suffices to show that $\lim_{n\to\infty} f_n(x) \geq f(x)$.
Without loss of generality, we may assume $l := \lim_{n\to\infty} f_n(x) < \infty$ (otherwise, the inequality we want to prove trivially holds). For each $n \in \mathbb{N}$, we can choose $y_n \in \mathcal{X}$ such that

$$f_n(x) \leq f(y_n) + nd(x, y_n) < f_n(x) + \frac{1}{n}. \tag{1}$$

Thus, using the fact that f is lower bounded, we get

$$d(x, y_n) < \frac{f_n(x) + \frac{1}{n} - f(y_n)}{n} \leq \frac{l + \frac{1}{n} - f(y_n)}{n} \leq \frac{C}{n}$$

for some constant C not depending on n. This yields $d(x, y_n) \to 0$ as $n \to \infty$, i.e. y_n converges to x. As we have $f_n(x) + \frac{1}{n} > f(y_n)$ by equation (1), we get

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} f_n(x) + \frac{1}{n} \geq \liminf_{n\to\infty} f(y_n) \geq f(x),$$

where we used the lower semicontinuity of f in the last estimate. □

Remark 3.2.8. An even stronger statement holds true: f is lower semicontinuous if and only if it can be written as the pointwise limit from below of a sequence of k-Lipschitz functions. To prove this version of the statement, one can use the same functions as in the proof above, as they are already k-Lipschitz by definition.

With this proposition at hand, we are now able to show that the functional F from above is lower semicontinuous.

Proposition 3.2.9 (Lower Semicontinuity of the Cost Functional)**.** *Let $c : \mathcal{X} \times \mathcal{Y}$ be a lower semicontinuous, bounded from below cost function. Then the functional*

$$F : \Pi(\mu, \nu) \to \mathbb{R} \cup \{+\infty\},\ \gamma \mapsto \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\gamma$$

is lower semicontinuous (where $\Pi(\mu, \nu)$ is equipped with the weak topology in $P(\mathcal{X} \times \mathcal{Y})$).

Proof. As c is bounded from below and lower semicontinuous, by lemma 3.2.7 there exists a sequence $(c_n)_{n\in\mathbb{N}}$ with $c_n : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $c(x, y) = \lim_{n\to\infty} c_n(x, y)$ from below for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Let $(\gamma_l)_{l\in\mathbb{N}} \subset \Pi(\mu, \nu)$ be a sequence converging weakly to some $\gamma \in \Pi(\mu, \nu)$. Then

$$\begin{aligned} F(\gamma) = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\gamma(x, y) &= \lim_{n\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c_n(x, y)\, \mathrm{d}\gamma(x, y) \\ &= \lim_{n\to\infty} \lim_{l\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c_n(x, y)\, \mathrm{d}\gamma_l(x, y) \\ &\leq \liminf_{l\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\gamma_l(x, y) \\ &= \liminf_{l\to\infty} F(\gamma_l), \end{aligned}$$

where in the first step, we used the Monotone Convergence Theorem (theorem A.15), the second step follows by weak convergence of γ_l, and the last one by the fact that the c_n converge to c from below. □

We are now able to prove the existence of an optimal transport plan for the Kantorovich Problem.

Theorem 3.2.10 (Existence of an Optimal Transport Plan)**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ a lower semicontinuous cost function that is bounded from below. Then there exists an optimal transport plan $\gamma \in \Pi(\mu, \nu)$ minimizing the total transport cost $\int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\gamma(x, y)$.*

Proof. Let

$$F : \Pi(\mu, \nu) \to \mathbb{R} \cup \{+\infty\}, \quad \gamma \mapsto \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\gamma(x, y).$$

We need to show that F attains its minimum on $\Pi(\mu,\nu)$. By proposition 3.2.9, we know that F is lower semicontinuous. All that is left to show is that $\Pi(\mu,\nu)$ is compact; then the claim follows by Weierstraß Theorem (A.12). From lemma 3.2.6, we know that $\{\mu\}$ and $\{\nu\}$ are tight in $P(\mathcal{X})$ and $P(\mathcal{Y})$ respectively. Hence, by lemma 3.2.4, $\Pi(\mu,\nu)$ is tight in $P(\mathcal{X}\times\mathcal{Y})$ as well. By Prokhorov's Theorem (A.13), $\Pi(\mu,\nu)$ is precompact, meaning its closure (with respect to the weak topology) is compact in $P(\mathcal{X}\times\mathcal{Y})$. Hence, in order to show that $\Pi(\mu,\nu)$ is compact, it suffices to show that it is closed.
Let $(\gamma_n)_{n\in\mathbb{N}}\subset\Pi(\mu,\nu)$ be a sequence converging weakly to some $\gamma\in P(\mathcal{X}\times\mathcal{Y})$. Let $(\varphi,\psi)\in C_b(\mathcal{X}\times\mathcal{Y})$. Then

$$\int_{\mathcal{X}\times\mathcal{Y}}\varphi(x)+\psi(y)\,\mathrm{d}\gamma(x,y)=\lim_{n\to\infty}\int_{\mathcal{X}\times\mathcal{Y}}\varphi(x)+\psi(y)\,\mathrm{d}\gamma_n(x,y)=\int_{\mathcal{X}}\varphi(x)\,\mathrm{d}\mu(x)+\int_{\mathcal{Y}}\psi(y)\,\mathrm{d}\nu(y),$$

from which $\gamma\in\Pi(\mu,\nu)$ follows by remark 3.1.2. □

Remark 3.2.11. The lower boundedness of c ensures that $\int_{\mathcal{X}\times\mathcal{Y}}c\,\mathrm{d}\gamma$ is well-defined in $\mathbb{R}\cup\{+\infty\}$. Oftentimes in applications, c will be a metric, so this assumption is automatically fulfilled. However, it is possible to generalize the theorem to more general cost functions; in [25], for example, it is only assumed that $c\geq a+b$ for some $a\in L^1(\mu)$ and $b\in L^1(\nu)$.

Remark 3.2.12. The existence of an optimal coupling does not imply that this optimal transport cost is finite; simply take $c=+\infty$, for example. Hence, sometimes stronger assumptions on c are made, such as $\int_{\mathcal{X}\times\mathcal{Y}}c\,\mathrm{d}\mu\,\mathrm{d}\nu<+\infty$ which yields a finite cost for at least the independent coupling.

Remark 3.2.13. One might wonder if we could prove the existence of an optimal *transport map* in a similar fashion. However, the problem is that the set of transport maps is in general *not* a compact subset of $P(\mathcal{X}\times\mathcal{Y})$; in fact, it is oftentimes dense in $\Pi(\mu,\nu)$. Hence, we can only hope for equality of the infimum in the Monge problem and the minimum in the Kantorovich problem. There exist numerous theorems of this form for certain settings. One of the most general ones as of today is the following.

Corollary 3.2.14. *Let $(\mathcal{X},\mu)$ and $(\mathcal{Y},\nu)$ be two Polish probability spaces, where μ is non-atomic (meaning for any $A\in\mathcal{B}(\mathcal{X})$ with $\mu(A)>0$, there exists some $B\in\mathcal{B}(\mathcal{X})$ with $B\subsetneq A$ with $\mu(B)>0$). Let $c:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}\cup\{+\infty\}$ be a continuous cost function. Then the infimum in the Monge problem is equal to the minimum in the Kantorovich problem.*

Proof. This theorem, alongside its proof, can be found as Theorem B in [21]. □

3.3 c-Transforms and the Dual Problem

Duality is oftentimes a powerful tool to rephrase a problem into an equivalent dual problem. As in many other areas of mathematics, optimal transport comes with such a dual formulation as well. Many concepts and results that are known in convex analysis transfer to optimal transport if we adapt the notions of some concepts, such as concavity or cyclical monotonicity, to the cost function c.
The dual problem very naturally extends our soil-analogy from earlier. So far, we were concerned with transporting soil from extraction points in $\mathcal{X}$ to construction sites in $\mathcal{Y}$ at minimal cost $\int_{\mathcal{X}\times\mathcal{Y}}c\,\mathrm{d}\gamma$.

Now assume a company offers to do the transport for you. They will buy the soil from you for $\psi(x)$ at the extraction point $x \in \mathcal{X}$ and sell it back to you for $\varphi(y)$ at the construction site $y \in \mathcal{Y}$. This means to get soil transported from x to y, you now pay $\varphi(y) - \psi(x)$ instead of $c(x, y)$. Obviously, you are only willing to accept this offer if the company stays below the transport cost c which you would pay if you did the transport yourself. Hence, they need to set the prices ψ and φ such that

$$\varphi(y) - \psi(x) \leq c(x, y) \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

Under this condition, the company will try to maximize their profits. This naturally yields the *dual Kantorovich problem*:

Problem 3.3.1 (Dual Kantorovich Problem)**.** Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ a cost function. The *dual Kantorovich problem* is defined as:

$$\sup \left\{ \int_{\mathcal{X}} \psi(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \varphi(y) \, \mathrm{d}\nu(y) : \ \psi \in L^1(\mu), \ \varphi \in L^1(\nu), \ \psi + \varphi \leq c \right\}.$$

The functions ψ and ϕ are also called *(dual) potentials*.

Note that for the sake of simplicitly, we changed the sign of ψ in our formulation of the dual. This means that $\psi(x)$ would correspond to *what you pay to the company* at $x \in \mathcal{X}$. A pair of price functions (ψ, φ) satisfying the condition $\psi + \varphi \leq c$ will be called *competitive*. Ultimately, as is usually the case with dual formulations, we would like to show equality of the optima appearing in the primal and dual problems. One inequality is both intuitive and easy to show: As any pair of competitive prices stays below the cost function, the value of the dual problem should be at most the value of the primal problem. Indeed, if $\gamma \in \Pi(\mu, \nu)$ is a transport plan and (ψ, φ) is a pair of competitive prices, then

$$\int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) \geq \int_{\mathcal{X} \times \mathcal{Y}} \psi(x) + \varphi(y) \, \mathrm{d}\gamma(x, y) = \int_{\mathcal{X}} \psi(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \varphi(y) \, \mathrm{d}\nu(y),$$

which yields

$$\begin{aligned} \inf_{\gamma \in \Pi(\mu, \nu)} & \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) \right\} \\ & \geq \sup \left\{ \int_{\mathcal{X}} \psi(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \varphi(y) \, \mathrm{d}\nu(y) : \ \psi \in L^1(\mu), \ \varphi \in L^1(\nu), \ \psi + \varphi \leq c \right\}. \quad (2) \end{aligned}$$

This means if we can find a transport plan γ and a competitive pair (ψ, φ) which yield equality, both are optimal for the primal and dual respectively.
For a given point $x \in \mathcal{X}$, the company will of course try to maximize $\psi(x)$ (as this is what you pay them). Under the premise of competitiveness, the maximum value $\psi(x)$ can take is $\inf_y c(x, y) - \varphi(y)$. Similarly, the company will try to maximize $\varphi(y)$ for a given $y \in \mathcal{Y}$, which they can set to a maximum of $\inf_x c(x, y) - \psi(x)$. In light of this consideration, we refer to a pair of prices (ψ, φ) as *tight* if

$$\psi(x) = \inf_y c(x, y) - \varphi(y) \ \text{ for all } x \in \mathcal{X}, \quad \varphi(y) = \inf_x c(x, y) - \psi(x) \ \text{ for all } y \in \mathcal{Y}. \quad (3)$$

As functions of this form will play a vital role, they get a name: They are called *c-transforms*.

Definition 3.3.2 (c-Transforms)**.** *Let* $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$, $\psi : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ *and* $\varphi : \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$.
The c-transform $\psi^c : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ *of* ψ *is defined via*

$$\psi^c(y) = \inf_{x \in \mathcal{X}} c(x, y) - \psi(x).$$

Similarily, the c-transform $\varphi^c : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ *of* φ *is defined via*

$$\varphi^c(x) = \inf_{y \in \mathcal{Y}} c(x, y) - \varphi(y).$$

If we have an arbitrary pair of competitive prices (ψ, φ), we could improve ψ by setting $\psi(x) = \varphi^c(x)$ everywhere. Then, in turn, we could improve φ by setting $\varphi(y) = \psi^c(y)$ everywhere. As can be easily seen, we cannot improve ψ and φ any further in this way; (3) now holds. Hence, it makes sense to restrict to tight pairs of functions in the dual problem. Since we can reconstruct φ from ψ using (3), we can consider ψ as the only variable in the dual. However, simply choosing $\psi \in L^1(\mu)$ arbitrarily and then defining φ as in (3) will not make the first equation in (3) hold. It will hold true if and only if ψ is *c-concave* according to the following definition.

Definition 3.3.3 (c-Concavity)**.** *Let* $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ *be a function.*
A function $\psi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ *is called* c-concave *if* $\psi \not\equiv -\infty$ *and there exists a function* $\varphi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ *such that* $\psi = \varphi^c$.
Similarily, a function $\varphi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ *is called* c-concave *if* $\varphi \not\equiv -\infty$ *and there exists a function* $\psi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ *such that* $\varphi = \psi^c$.

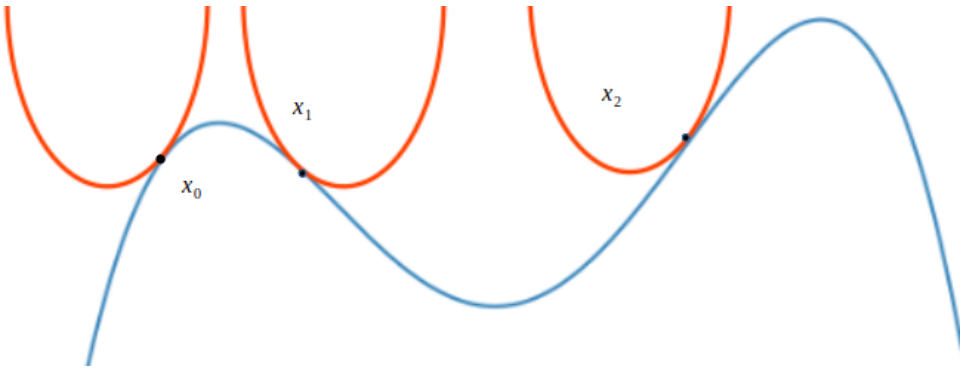

Figure 1: A c-concave function $\psi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ is one whose graph can entirely be caressed from above with the negative cost function. The points (x_i, y_i), $i \in [\![3]\!]$, come from the superdifferentials $\partial^c\psi(x_i)$ respectively, see definition 3.3.5. The blue graph shows ψ and the red graphs show the functions $c(\cdot, y_i) - \psi^c(y_i)$ respectively.

Example 3.3.4. In the special case where $c = d$ is a metric (i.e. $\mathcal{X} = \mathcal{Y}$), being c-concave is equivalent to being 1-Lipschitz (i.e. being Lipschitz continuous with Lipschitz constant equal to 1). To see this, first let $\psi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ be c-concave and φ as in definition 3.3.3, such that $\psi = \varphi^c$. Then for

$x, y \in \mathcal{X}$ we have

$$\begin{aligned}
|\psi(x) - \psi(y)| &= \left| \left(\inf_{z\in\mathcal{X}} c(x,z) - \varphi(z) \right) - \left(\inf_{z\in\mathcal{X}} c(y,z) - \varphi(z) \right) \right| \\
&= \left| \left(\sup_{z\in\mathcal{X}} \varphi(z) - c(y,z) \right) - \left(\sup_{z\in\mathcal{X}} \varphi(z) - c(x,z) \right) \right| \\
&\leq \left| \sup_{z\in\mathcal{X}} \varphi(z) - d(x,z) - \varphi(z) + d(y,z) \right| \\
&= \left| \sup_{z\in\mathcal{X}} d(y,z) - d(x,z) \right| \leq d(x,y),
\end{aligned}$$

which shows that ψ is 1-Lipschitz. On the other hand, if ψ is 1-Lipschitz, we have

$$\psi(x) \leq d(x,y) + \psi(y)$$

for all $x, y \in \mathcal{X}$ and, choosing $y = x$, we can see that this means

$$\psi(x) = \inf_{y\in\mathcal{X}} d(x,y) - (-\psi(y)) = (-\psi)^c(x),$$

i.e. $\psi = (-\psi)^c$ which shows that ψ is c-concave. Similarly, from

$$-\psi(x) \leq d(x,y) - \psi(y)$$

we can conclude that

$$-\psi(x) = \inf_{y\in\mathcal{X}} d(x,y) - \psi(y) = \psi^c(x),$$

i.e. $\psi^c = -\psi$. This shows that the c-transform of ψ is not just any function, but in this case actually equal to $-\psi$.

As can be seen from our previous considerations, the set where $\psi^c(y) = c(x,y) - \psi(x)$ is special in the sense that on this set, the infimum from (3) is attained at $\psi(x)$ for $\varphi = \psi^c$ (note that the inequality $\psi^c(y) \leq c(x,y) - \psi(x)$ always holds by definition). This set also gets a name.

Definition 3.3.5 (c-Superdifferential)**.** *Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ and let $\psi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ be c-concave. Then the c-superdifferential $\partial^c\psi \subset \mathcal{X} \times \mathcal{Y}$ of ψ is defined as*

$$\partial^c\psi = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : \ \psi^c(y) = c(x,y) - \psi(x)\}.$$

The c-superdifferential $\partial^c\psi(x)$ of ψ at $x \in \mathcal{X}$ is given by

$$\partial^c\psi(x) = \{y \in \mathcal{Y} : \ (x,y) \in \partial^c\psi\}.$$

Remark 3.3.6. In the literature, there exist many more definitions, such as c^+- and c^--transforms, c-convexity, or c-subdifferentials. However, they are all redundant in some sense. For example, a function ψ is c-convex if and only if $-\psi$ is c-concave. Hence, the definitions from above will suffice.

The next result justifies the concept of c-concavity, as it shows that c-concave functions are exactly those functions where a double c-transformation yields the same function again.

Proposition 3.3.7 (Alternative Characterization of c-Concavity)**.** *For $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, let $\psi^{cc} := (\psi^c)^c$. Then ψ is c-concave if and only if $\psi^{cc} = \psi$.*

Proof. First, we note that for any function $\phi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$, we have the identity $\phi^c = (\phi^{cc})^c =: \phi^{ccc}$, as

$$\begin{aligned}\phi^{ccc}(x) &= \inf_{y\in\mathcal{Y}} \left[c(x,y) - \inf_{\tilde{x}\in\mathcal{X}} \left(c(\tilde{x},y) - \inf_{\tilde{y}\in\mathcal{Y}} \left(c(\tilde{x},\tilde{y}) - \phi(\tilde{y}) \right) \right) \right] \\ &= \inf_{y\in\mathcal{Y}} \sup_{\tilde{x}\in\mathcal{X}} \inf_{\tilde{y}\in\mathcal{Y}} \left[c(x,y) + c(\tilde{x},\tilde{y}) - \phi(\tilde{y}) - c(\tilde{x},y) \right]\end{aligned}$$

and $\tilde{x} = x$ yields

$$\phi^{ccc}(x) \geq \inf_{y\in\mathcal{Y}} \inf_{\tilde{y}\in\mathcal{Y}} \left[c(x,y) + c(x,\tilde{y}) - \phi(\tilde{y}) - c(x,y) \right] = \psi^c(x),$$

whereas $\tilde{y} = y$ yields

$$\phi^{ccc} \leq \inf_{y\in\mathcal{Y}} \sup_{\tilde{x}\in\mathcal{X}} \left[c(x,y) + c(\tilde{x},y) - \phi(y) - c(\tilde{x},y) \right] = \psi^c(x).$$

Now if ψ is c-concave, there exists a function φ as in definition 3.3.3 such that $\psi = \varphi^c$, hence $\psi^{cc} = \varphi^{ccc} = \varphi^c = \psi$. On the other hand, if $\psi = \psi^{cc}$, then ψ is the c-transform of ψ^c and c-concave by definition. □

An alternative characterization of the c-superdifferential at a point x which we will need later on is the following.

Proposition 3.3.8 (Alternative Characterization of the c-Superdifferential)**.** *Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ and let $\psi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ be c-concave. A point $y \in \mathcal{Y}$ lies in the c-superdifferential of ψ at $x \in \mathcal{X}$ if and only if*

$$\psi(x) - c(x,y) \geq \psi(z) - c(z,y) \quad \text{for all } z \in \mathcal{X}.$$

Proof. "⇒": Let $y \in \partial^c\psi(x)$. We have

$$\begin{aligned}\psi(x) - c(x,y) = -\psi^c(y) &= -\inf_{z\in\mathcal{X}} c(z,y) - \psi(z) \\ &= \sup_{z\in\mathcal{X}} \psi(z) - c(z,y) \geq \psi(z) - c(z,y) \quad \text{for all } z \in \mathcal{X}.\end{aligned}$$

"⇐": Let $\psi(x) - c(x,y) \geq \psi(z) - c(z,y)$ hold for all $z \in \mathcal{X}$. Then

$$\begin{aligned}\psi(x) - c(x,y) &\geq \sup_{z\in\mathcal{X}} \psi(z) - c(z,y) \\ &= -\inf_{z\in\mathcal{X}} c(z,y) - \psi(z) = -\psi^c(y),\end{aligned}$$

but as we have seen before, the reverse inequality $\psi^c(y) \leq c(x,y) - \psi(x)$ always holds by definition, which gives us $\psi^c(y) = c(x,y) - \psi(x)$, hence $y \in \partial^c\psi(x)$. □

Another concept that we will need is that of *c-cyclical monotonicity*. Before we define it, let's motivate its definition by our analogy again. Say you have a transport plan, but you think you could improve it. In order to do so, you decide to reroute one unit of soil that was originally sent from x_1 to y_1 to

go to y_2 instead. This means you reduce the transport cost by $c(x_1, y_1) - c(x_1, y_2)$. Now as you have excess soil at y_2, you send one unit that was sent from x_2 to y_2 to go to y_3 instead. If you keep going like this, at some point you will have to send a unit of soil that was going from x_n to y_n to go to y_1 instead, as y_1 was still lacking one unit from earlier. This means your new transport plan is better than the old one if and only if

$$c(x_1, y_2) + c(x_2, y_3) + ... + c(x_n, y_1) < c(x_1, y_1) + c(x_2, y_2) + ... + c(x_n, y_n).$$

If you can find such a cycle improving the transport cost, this shows your original plan was not optimal. Conversely, if you cannot find such a cycle, it seems likely that your original plan was indeed optimal (and we will see later that under mild assumptions, this is in fact true), and it operates on a *c-cyclically monotone set.*

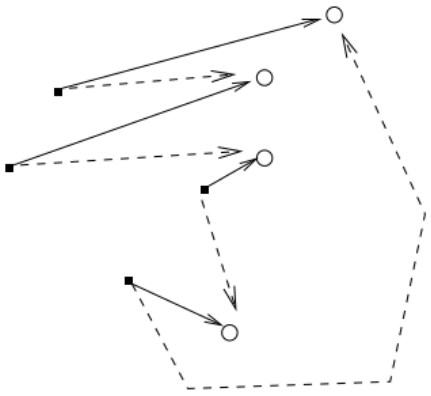

Figure 2: Trying to reduce the transport cost by finding a cycle of lower cost. Solid arrows stand for the original transport plan, dashed arrows for the rerouted mass.[4]

Definition 3.3.9 (c-Cyclical Monotonicity)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two sets and $x : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ a function. A subset $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ is called* c-cyclically monotone *if for all $n \in \mathbb{N}_{>0}$, all $(x_1, y_1), ..., (x_n, y_n) \in \Gamma$, and all permutations $\sigma \in S_n$, there holds*

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

A transport plan is said to be c-cyclically monotone *if it is concentrated on a c-cyclically monotone set.*

One nice result is that c-superdifferentials are always c-cyclically monotone.

Proposition 3.3.10. *Let $\psi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ be c-concave. Then $\partial^c \psi$ is a c-cyclically monotone set.*

Proof. Let $n \in \mathbb{N}_{>0}$ and $(x_i, y_i) \in \partial^c \psi$, $i \in [\![n]\!]$. Let $\sigma \in S_n$. Then

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

□

[4]Source: [25], figure 5.1.

In the following section, we will see as part of the Fundamental Theorem of Optimal Transport 3.4.1 that under mild assumptions on c, *every* c-cyclically monotone set can in turn be obtained from the c-superdifferential of a c-concave function.

3.4 Fundamental Theorem of Optimal Transport

This section is devoted to the Fundamental Theorem of Optimal Transport and its proof, which is partly also based on [1].

Theorem 3.4.1 (Fundamental Theorem of Optimal Transport)**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous and bounded from below cost function, such that for some $a \in L^1(\mu)$ and $b \in L^1(\nu)$,*

$$c(x, y) \leq a(x) + b(y) \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

Let $\gamma \in \Pi(\mu, \nu)$ be an arbitrary transport plan. Then the following three statements are equivalent:

(i) γ is optimal for the Kantorovich problem,

(ii) $\operatorname{supp}(\gamma)$ is a c-cyclically monotone set in $\mathcal{X} \times \mathcal{Y}$,

(iii) there exists a c-concave function ψ such that $\max\{\psi, 0\} \in L^1(\mu)$ and $\operatorname{supp}(\gamma) \subset \partial^c \psi$.

Proof. First, notice that $c \in L^1(\tilde{\gamma})$ for any $\tilde{\gamma} \in \Pi(\mu, \nu)$ as c is bounded from below and

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \mathrm{d}\tilde{\gamma}(x, y) \leq \int_{\mathcal{X}\times\mathcal{Y}} a(x) + b(y)\, \mathrm{d}\tilde{\gamma}(x, y) = \int_{\mathcal{X}} a(x)\, \mathrm{d}\mu)(x) + \int_{\mathcal{Y}} b(y)\, \mathrm{d}\nu(y) < \infty.$$

$(i) \Rightarrow (ii)$:
Intuitively, it is quite clear what to do: Assume that $\operatorname{supp}(\gamma)$ is not c-cyclically monotone, find a set on which we can reduce the transport cost, and "shift" γ along this set in order to construct a new transport plan which has lower total cost than γ, which yields a contradiction.
More explicitly: We assume for the sake of contradiction that $\operatorname{supp}(\gamma)$ is not c-cyclically monotone. That means we can find $n \in \mathbb{N}_{>0}$, $(x_i, y_i) \in \operatorname{supp}(\gamma)$, $i \in [\![n]\!]$, and some $\sigma \in S_n$ such that

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

As c is continuous, we can find neighbourhoods $U_i \times V_i \in \mathcal{B}(\mathcal{X} \times \mathcal{Y})$ of (x_i, y_i) for all i such that

$$\sum_{i=1}^{n} c(u_i, v_{\sigma(i)}) - c(u_i, v_i) < 0 \quad \text{for all } (u_i, v_i) \in U_i \times V_i,\ i \in [\![n]\!]. \tag{4}$$

Now we will construct a signed measure η (see definition A.1) on $\mathcal{B}(\mathcal{X} \times \mathcal{Y})$ such that the "variation" $\tilde{\gamma} := \gamma + \eta$ has a lower total cost than γ. To this end, η needs to fulfill the following three conditions:

(1) $\eta^- \leq \gamma$ (where η^- is the lower variation of η, see definition A.4, such that $\tilde{\gamma} \geq 0$ is a measure)

(2) $\eta \circ \pi_{\mathcal{X}}^{-1} = 0$, $\eta \circ \pi_{\mathcal{Y}}^{-1} = 0$ (i.e. the marginals are zero, s.t. $\tilde{\gamma} \in \Pi(\mu, \nu)$)

(3) $\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\eta(x,y) < 0$ (s.t. γ is not optimal)

Note that the second condition will also imply that $0 = \eta(\pi_{\mathcal{X}}^{-1}(\mathcal{X})) = \eta(\mathcal{X}\times\mathcal{Y})$, i.e. $\tilde{\gamma}(\mathcal{X}\times\mathcal{Y}) = 1$.
Let $\Omega := \Pi_{i=1}^n U_i \times V_i$ and $P \in P(\Omega)$ be defined as $P = \gamma_1 \otimes \gamma_2 \otimes ... \otimes \gamma_n$, where $\gamma_i := \frac{1}{m_i}\gamma|_{U_i\times V_i}$ with $m_i := \gamma(U_i \times V_i)$, $i \in [\![n]\!]$. Then each γ_i is a probability measure on $U_i \times V_i$ and P is a probability measure on Ω. Let $\pi_{U_i} : \Omega \to U_i$ and $\pi_{V_i} : \Omega \to V_i$ be the projections onto U_i and V_i. Set

$$\eta := \frac{\min_i m_i}{n} \sum_{i=1}^n P \circ (\pi_{U_i}, \pi_{V_{\sigma(i)}})^{-1} - P \circ (\pi_{U_i}, \pi_{V_i})^{-1}.$$

Let $(A \times B) \in \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{Y})$ be arbitrary. Since for all i we have $(A \times B) \cap (U_i \times V_i) \subset A \times B$, we have

$$\eta^-(A \times B) \leq \frac{\min_i m_i}{n} \sum_{i=1}^n \frac{1}{m_i}\gamma(A \times B) \leq \frac{1}{n} \cdot n\gamma(A \times B) = \gamma(A \times B),$$

and since $\mathcal{B}(\mathcal{X} \times \mathcal{Y})$ is generated by sets of this form, it follows that $\eta^- \leq \gamma$, which proves (1).
Regarding (2), for $B \in \mathcal{B}(\mathcal{Y})$, we have

$$\begin{aligned}
\eta \circ \pi_{\mathcal{Y}}^{-1}(B) &= \eta(\mathcal{X} \times B) \\
&= \frac{\min_i m_i}{n} \sum_{i=1}^n P\left((U_1 \times V_1) \times ... \times (U_{\sigma(i)} \times (V_{\sigma(i)} \cap B)) \times ... \times (U_n \times V_n)\right) \\
&\qquad - P\left((U_1 \times V_1) \times ... \times (U_i \times (V_i \cap B)) \times ... \times (U_n \times V_n)\right) \\
&= \frac{\min_i m_i}{n} \sum_{i=1}^n \gamma_{\sigma(i)}(U_{\sigma(i)} \times (V_{\sigma(i)} \cap B)) - \gamma_i(U_i \times (V_i \cap B)) = 0, \qquad (5)
\end{aligned}$$

hence $\eta \circ \pi_{\mathcal{Y}}^{-1} = 0$. Similarily, $\eta \circ \pi_{\mathcal{X}}^{-1} = 0$ follows, except that in the analogous derivation as in equation (5), the permutation will disappear. This proves (2).
Regarding (3), note that from (4) it follows that $\int c\,\mathrm{d}\eta \leq 0$, but since c is continuous we even have $\int c\,\mathrm{d}\eta < 0$.
To sum it up, we now have a transport plan $\tilde{\gamma} \in \Pi(\mu, \nu)$ such that

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\tilde{\gamma}(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\gamma(x,y) + \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\eta(x,y) < \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\gamma(x,y)$$

which contradicts the optimality of γ.

(ii) $\Rightarrow$ *(iii)* :
We will prove an even more general statement: Any c-cyclically monotone set $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ is contained in the c-superdifferential of a c-concave function ψ s.t. $\max\{\psi, 0\} \in L^1(\mu)$.
Let Γ be a c-cyclically monotone set and $(\bar{x}, \bar{y}) \in \Gamma$. Since we want $\Gamma \subset \partial^c\psi$ to hold, for any $n \in \mathbb{N}_{>0}$

and any $(x_i, y_i) \in \Gamma$, $i \in [\![n]\!]$, there has to hold

$$\begin{aligned}\psi(x) &\leq c(x, y_1) - \psi^c(y_1) = c(x, y_1) - c(x_1, y_1) + \psi(x_1) \\ &\leq (c(x, y_1) - c(x_1, y_1)) + c(x_1, y_2) - \psi^c(y_2) \\ &= ... \\ &\leq (c(x, y_1) - c(x_1, y_1)) + (c(x_1, y_2) - c(x_2, y_2)) + ... + (c(x_n, \bar{y}) - c(\bar{x}, \bar{y})) + \psi(\bar{x}).\end{aligned}$$

Hence, it makes sense to define ψ as the infimum over all such expressions. However, we leave out $\psi(\bar{x})$ in the end. Note that a function ψ is c-concave if and only if $\psi + k$ for a constant $k \in \mathbb{R}$ is c-concave, and that the c-superdifferentials of ψ and $\psi + k$ are identical, thus we are free to ignore $\psi(\bar{x})$. We define

$$\psi(x) := \inf_{\substack{n \in \mathbb{N}_{>0} \\ (x_i, y_i) \in \Gamma, i \in [\![n]\!]}} (c(x, y_1) - c(x_1, y_1)) + (c(x_1, y_2) - c(x_2, y_2)) + ... + (c(x_n, \bar{y}) - c(\bar{x}, \bar{y})) \,. \tag{6}$$

Now for $n = 1$ and $(x_1, y_1) = (\bar{x}, \bar{y})$ we get $\psi(\bar{x}) \leq c(\bar{x}, \bar{y}) - c(\bar{x}, \bar{y}) = 0$, whereas we get $\psi(\bar{x}) \geq 0$ from the fact that $\psi(\bar{x})$ is defined as the infimum over expressions which by c-cyclical monotonicity of Γ are all non-negative. Thus, $\psi(\bar{x}) = 0$ (which yields $\psi \not\equiv -\infty$, as is needed by definition of c-concave functions). In order to see that ψ is indeed c-concave, set

$$\varphi(y) := \sup_{\substack{n \in \mathbb{N}_{>0} \\ (x_1, y), (x_i, y_i) \in \Gamma, i \in [\![n]\!]}} c(x_1, y) - c(x_1, y_2) + c(x_2, y_2) - ... - c(x_n, \bar{y}) + c(\bar{x}, \bar{y}),$$

for $y \in \pi_{\mathcal{Y}}(\Gamma)$, and $\varphi(y) := -\infty$ for $y \notin \pi_{\mathcal{Y}}(\Gamma)$. Then, replacing y_1 by y in the definition of ψ, we can see that

$$\begin{aligned}\psi(x) &= \inf_{y \in \mathcal{Y}} \inf_{\substack{n \in \mathbb{N}_{>0} \\ (x_1, y), (x_i, y_i) \in \Gamma, i \in [\![n]\!]}} c(x, y) - c(x_1, y) + c(x_1, y_2) - c(x_2, y_2) + ... + c(x_n, \bar{y}) - c(\bar{x}, \bar{y}) \\ &= \inf_{y \in \mathcal{Y}} c(x, y) - \varphi(y).\end{aligned}$$

Choosing $n = 1$ and $(x_1, y_1) = (\bar{x}, \bar{y})$ again, we get

$$\psi(x) \leq c(x, \bar{y}) - c(\bar{x}, \bar{y}) \leq a(x) + b(\bar{y}) - c(\bar{x}, \bar{y}),$$

and as $a \in L^1(\mu)$, this yields $\max\{\psi, 0\} \in L^1(\mu)$. Now all that is left to show is that $\Gamma \subset \partial^c \psi$. To this end, let $(\tilde{x}, \tilde{y}) \in \Gamma$. We need to show that $(\tilde{x}, \tilde{y}) \in \partial^c \psi$. Let $(x_1, y_1) = (\tilde{x}, \tilde{y})$. Then from (6) we get

$$\begin{aligned}\psi(x) &\leq c(x, \tilde{y}) - c(\tilde{x}, \tilde{y}) + \inf_{\substack{n \in \mathbb{N} \\ (x_i, y_i) \in \Gamma, i \in [\![2, n]\!]}} c(\tilde{x}, y_2) - c(x_2, y_2) + ... + c(x_n, \bar{y}) - c(\bar{x}, \bar{y}) \\ &\leq c(x, \tilde{y}) - c(\tilde{x}, \tilde{y}) + \inf_{\substack{n \in \mathbb{N}_{>0} \\ (x_i, y_i) \in \Gamma, i \in [\![n]\!]}} c(\tilde{x}, y_1) - c(x_1, y_1) + ... + c(x_n, \bar{y}) - c(\bar{x}, \bar{y}) \\ &= c(x, \tilde{y}) - c(\tilde{x}, \tilde{y}) + \psi(\tilde{x}),\end{aligned}$$

which holds for all $x \in \mathcal{X}$. By proposition 3.3.8, this is equivalent to $(\tilde{x}, \tilde{y}) \in \partial^c \psi$.

$(iii) \Rightarrow (i)$:

Let ψ be as in (iii) and $\operatorname{supp}(\gamma) \subset \partial^c\psi$. We have

$$\begin{aligned}\psi(x) + \psi^c(y) &= c(x,y) \quad \text{for all } (x,y) \in \operatorname{supp}(\gamma),\\ \psi(x) + \psi^c(y) &\leq c(x,y) \quad \text{for all } (x,y) \in \mathcal{X} \times \mathcal{Y}.\end{aligned}$$

Let $\tilde{\gamma} \in \Pi(\mu,\nu)$ be an arbitrary transport plan. We will show that $\int c\,\mathrm{d}\gamma \leq \int c\,\mathrm{d}\tilde{\gamma}$, which yields the claim:

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\gamma(x,y) &= \int_{\mathcal{X}\times\mathcal{Y}} \psi(x) + \psi^c(y)\,\mathrm{d}\gamma(x,y) = \int_{\mathcal{X}} \psi(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi^c(y)\,\mathrm{d}\nu(y)\\ &= \int_{\mathcal{X}\times\mathcal{Y}} \psi(x) + \psi^c(y)\,\mathrm{d}\tilde{\gamma}(x,y) \leq \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\tilde{\gamma}(x,y).\end{aligned}$$

□

Remark 3.4.2. One interesting statement worth noting was proven: under the assumptions made on c in theorem 3.4.1, *any* c-cyclically monotone set is contained in the c-superdifferential of a c-concave function.

Remark 3.4.3. Another important and immediate consequence of theorem 3.4.1 is the following: If γ is an optimal transport plan and $\tilde{\gamma}$ is another arbitrary transport plan with $\operatorname{supp}(\tilde{\gamma}) \subset \operatorname{supp}(\gamma)$, then $\tilde{\gamma}$ is also optimal. This means that optimality does not depend on how the mass is distributed within the support, but only on the support itself.

Remark 3.4.4. In theorem 3.4.1 we showed that for every optimal γ, there exists a c-concave function ψ such that $\operatorname{supp}(\gamma) \subset \partial^c\psi$. Actually, an ever stronger statement holds true: For any other optimal transport plan $\tilde{\gamma}$, we have $\operatorname{supp}(\tilde{\gamma}) \subset \partial^c\psi$ *with the same function* ψ. To see this, first note that $\max\{\psi^c, 0\} \in L^1(\nu)$ as

$$\psi^c(y) = \inf_{x\in\mathcal{X}} c(x,y) - \psi(y) \leq c(x,y) - \psi(x) \leq a(x) + b(y) - \psi(x)$$

for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. As $b \in L^1(\nu)$ (and $a, \psi \not\equiv -\infty$), this yields $\max\{\psi^c, 0\} \in L^1(\nu)$. Hence,

$$\begin{aligned}\int_{\mathcal{X}} \psi(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi^c(y)\,\mathrm{d}\nu(y) &= \int_{\mathcal{X}\times\mathcal{Y}} \psi(x) + \psi^c(y)\,\mathrm{d}\gamma(x,y)\\ &\leq \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\tilde{\gamma}(x,y)\\ &= \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\gamma(x,y)\\ &= \int_{\mathcal{X}\times\mathcal{Y}} \psi(x) + \psi^c(y)\,\mathrm{d}\gamma(x,y),\end{aligned}$$

hence the inequality is an equality which means $(x,y) \in \partial^c\psi$ $\tilde{\gamma}$-surely. By continuity of c, it follows $\operatorname{supp}(\tilde{\gamma}) \subset \partial^c\psi$.

Remark 3.4.5. As we have seen in theorem 3.2.10, continuity of c is not needed for an optimal transport plan to exist, lower semicontinuity suffices. So one might wonder whether theorem 3.4.1 carries over to this setting. This is, in general, not the case. One can show that, under the same assumptions on c as in theorem 3.4.1 with continuity replaced by lower semicontinuity, the following

implication holds: If $\gamma \in \Pi(\mu, \nu)$ is optimal, then it is concentrated on a c-cyclically monotone set. However, this set need not be closed in general, hence the support of γ does not equal this set.

3.5 Duality Theorem

We are now ready to prove that the infimum in the primal Kantorovich problem and the supremum in the dual problem coincide under the same assumptions as in theorem 3.4.1.

Theorem 3.5.1 (Duality Theorem)**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a continuous and bounded from below cost function such that*

$$c(x, y) \leq a(x) + b(y) \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

for some $a \in L^1(\mu)$, $b \in L^1(\nu)$. Then there is duality:

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) = \sup_{\substack{(\psi,\varphi) \in L^1(\mu) \times L^1(\nu) \\ \psi + \varphi \leq c}} \int_{\mathcal{X}} \psi(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \varphi(y) \, \mathrm{d}\nu(y). \tag{7}$$

Furthermore, both the infimum and the supremum are attained, and the maximizing couple (ψ, φ) in the supremum is of the form (ψ, ψ^c) for some c-concave function ψ.

Proof. In theorem 3.2.10, we proved (under even milder assumptions on c) the existence of an optimal transport plan for the primal problem, i.e. the left hand side in (7). As we have seen before, the infimum on the left hand side needs to be greater or equal than the supremum on the right hand side, see equation (2). For the reverse inequality, let $\gamma \in \Pi(\mu, \nu)$ be optimal. By theorem 3.4.1 and remark 3.4.4, we know there exists a c-concave function ψ such that $\mathrm{supp}(\gamma) \subset \partial^c \psi$, $\max\{\psi, 0\} \in L^1(\mu)$, and $\max\{\psi^c, 0\} \in L^1(\nu)$. This gives us

$$\begin{aligned} \infty > \int_{\mathcal{X}} a(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} b(y) \, \mathrm{d}\nu(y) &= \int_{\mathcal{X}\times\mathcal{Y}} a(x) + b(y) \, \mathrm{d}\gamma(x, y) \geq \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) \\ = \int_{\mathcal{X}\times\mathcal{Y}} \psi(x) + \psi^c(y) \, \mathrm{d}\gamma(x, y) &= \int_{\mathcal{X}} \psi(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi^c(y) \, \mathrm{d}\nu(y) \end{aligned}$$

and as c is bounded from below, this gives us $\psi \in L^1(\mu)$ and $\psi^c \in L^1(\nu)$ which proves that (ψ, ψ^c) is an admissible couple for the dual problem. This proves the reverse inequality and in particular, as

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y) = \int_{\mathcal{X}} \psi(x) \, \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi^c(y) \, \mathrm{d}\nu(y),$$

shows that (ψ, ψ^c) is optimal for the dual problem. □

Remark 3.5.2. Again, an ever stronger statement holds true: For *any* c-concave maximizing couple (ψ, ψ^c) and any optimal $\gamma \in \Pi(\mu, \nu)$ we have $\mathrm{supp}(\gamma) \subset \partial^c \psi$: By remark 3.4.4, we know that there

exists some c-concave $\tilde{\psi}$ such that $\tilde{\psi} \in L^1(\mu)$, $\tilde{\psi}^c \in L^1(\nu)$ and $\text{supp}(\gamma) \subset \partial^c \tilde{\psi}$. This yields

$$\begin{aligned}\int_{\mathcal{X}} \psi(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi^c(y)\,\mathrm{d}\nu(y) &\geq \int_{\mathcal{X}} \tilde{\psi}(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} \tilde{\psi}^c(y)\,\mathrm{d}\nu(y) = \int_{\mathcal{X}\times\mathcal{Y}} \tilde{\psi}(x) + \tilde{\psi}^c(y)\,\mathrm{d}\gamma(x,y) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,\mathrm{d}\gamma(x,y) \geq \int_{\mathcal{X}\times\mathcal{Y}} \psi(x) + \psi^c(y)\,\mathrm{d}\gamma(x,y) \\ &= \int_{\mathcal{X}} \psi(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi^c(y)\,\mathrm{d}\nu(y),\end{aligned}$$

hence there must be equality which shows $\text{supp}(\gamma) \subset \partial^c \psi$. In particular, this shows that any admissible couple $(\tilde{\psi}, \tilde{\psi}^c)$ for the dual for which we have $\text{supp}(\gamma) \subset \partial^c \tilde{\psi}$ for an optimal γ is in fact optimal. This means also the c-concave function ψ appearing in theorem 3.4.1 is optimal for the dual problem.

Remark 3.5.3. Again, we can ask ourselves whether or not theorem 3.5.1 carries over to the setting where c is only lower semicontinuous. Indeed, under the same assumptions on c with continuity replaced by lower semicontinuity, the duality between the optima in the primal and dual problem still holds. However, it is not guaranteed to be attained by a couple (ψ, ψ^c) of c-concave functions anymore.

3.6 Wasserstein Distances

This section deals with the special case where $\mathcal{X} = \mathcal{Y}$. This allows one to use the metric on $\mathcal{X}$ as a cost function, giving rise to *Wasserstein distances*. This means we will shift our attention now; so far we had been interested in the *minimizer* of the optimal transport problem. Now we are interested in the *minimum*. As it turns out, this minimum defines a metric between the origin and target distribution.

Definition 3.6.1 (Wasserstein Space, Wasserstein distance)**.** *Let $(\mathcal{X}, d)$ be a Polish space and $p \in [1, \infty)$. Then the* Wasserstein space of order p *is defined as:*

$$\mathcal{P}_p(\mathcal{X}) = \left\{ \mu \in P(\mathcal{X}) : \int_{\mathcal{X}} d(x_0, x)^p \,\mathrm{d}\mu(x) < \infty \right\}$$

for an arbitrary $x_0 \in \mathcal{X}$.[5]
For $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ the Wasserstein distance of order p between μ and ν *is defined as*

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \,\mathrm{d}\gamma(x,y) \right)^{\frac{1}{p}}.$$

Remark 3.6.2. The terminology in the literature is not very coherent. Wasserstein distances had been discovered and rediscovered by multiple authors over the span of the twentieth century, including Wasserstein whose name is actually spelled "Vasershtein"; so "Wasserstein" as a name is very doubtful for that reason alone. Other names, such as "Wasserstein metric" or "Kantorovich distance", exist as well. In particular, the distance W_1 for the case $p = 1$ is known under different names as well, such as "Kantorovich-Rubinstein distance" or, more recently and mostly in image processing and computer science, "Earth Mover's distance".

[5] Note that the space does not depend on the choice of x_0.

Remark 3.6.3 (Special Case W_1)**.** In the case $p = 1$ the Wasserstein distance takes another special form. We know from theorem 3.5.1 that

$$W_1(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)\,\mathrm{d}\gamma(x,y) = \sup_{\substack{\psi\in L^1(\mu)\\ \psi \text{ c-concave}}} \int_{\mathcal{X}} \psi(x)\,\mathrm{d}\mu(x) + \int_{\mathcal{X}} \psi^c(x)\,\mathrm{d}\nu(x).$$

In example 3.3.4, we saw that in the case $c(x,y) = d(x,y)$, being c-concave is actually equivalent to being 1-Lipschitz. Furthermore, we saw that in this case, $\psi^c = -\psi$. This gives us the following formula for $W_1(\mu,\nu)$:

$$W_1(\mu,\nu) = \sup_{\psi \text{ 1-Lipschitz}} \int_{\mathcal{X}} \psi(x)\,\mathrm{d}\mu(x) - \int_{\mathcal{X}} \psi(x)\,\mathrm{d}\nu(x).$$

In particular, we see that the value of $W_1(\mu,\nu)$ does not really depend on μ *and* ν; it only depends on their difference $\mu - \nu$.

Example 3.6.4. In the special case $\mu = \delta_x$ for some $x \in \mathcal{X}$ and $\nu = \delta_y$ for some $y \in \mathcal{Y}$, we have $W_p(\delta_x,\delta_y) = d(x,y)$. In particular, $W_p(\delta_x,\delta_y)$ does not depend on p which is generally not the case.

Next, we will show that, as the name suggests, W_p defines a metric on $\mathcal{P}_p(\mathcal{X})$. To prove this result, we need the following lemma.

Lemma 3.6.5 (Gluing)**.** *Let $(\mathcal{X}_i,\mu_i)$, $i \in [\![3]\!]$, be Polish probability spaces and $\gamma_{12} \in \Pi(\mu_1,\mu_2)$ and $\gamma_{23} \in \Pi(\mu_2,\mu_3)$ be transport plans. Then there exists a probability measure $\gamma \in P(\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3)$ with marginals γ_{12} on $\mathcal{X}_1\times\mathcal{X}_2$ and γ_{23} on $\mathcal{X}_2\times\mathcal{X}_3$. In particular, this means γ has marginals μ_i on $\mathcal{X}_i$ for $i \in [\![3]\!]$.*

Proof. Let $\mathcal{X}$ and $\mathcal{Y}$ be any Polish probability spaces and $\gamma \in P(\mathcal{X}\times\mathcal{Y})$ with $\gamma\circ\pi_1^{-1} = \mu$ for some $\mu \in P(\mathcal{X})$. Let $\pi_{\mathcal{X}} : \mathcal{X}\times\mathcal{Y} \to \mathcal{X}$ be the usual projection. By the disintegration theorem A.14 we can find measures $(\gamma_x) \subset P(\mathcal{X}\times\mathcal{Y})$ for $x \in \mathcal{X}$ which are μ-almost uniquely defined s.t.

$$0 = \gamma_x\left((\mathcal{X}\times\mathcal{Y}) \setminus \pi_{\mathcal{X}}^{-1}(x)\right) = \gamma_x\left((\mathcal{X}\times\mathcal{Y}) \setminus (\{x\}\times\mathcal{Y})\right) \quad \mu\text{-almost everywhere,}$$

i.e. the measures γ_x are concentrated on the fibres $\{x\}\times\mathcal{Y}$ a.e., meaning we can identify γ_x with a measure on $\mathcal{Y}$ (which we will do in the following), and for any $\phi : \mathcal{X}\times\mathcal{Y} \to [0,\infty]$ we have

$$\int_{\mathcal{X}\times\mathcal{Y}} \phi(x,y)\,\mathrm{d}\gamma(x,y) = \int_{\mathcal{X}}\int_{\mathcal{Y}} \phi(x,y)\,\mathrm{d}\gamma_x(y)\,\mathrm{d}\mu(x).$$

We will write this as

$$\gamma = \int_{\mathcal{X}} \delta_x \otimes \gamma_x\,\mathrm{d}\mu(x).$$

Now disintegrating γ_{12} and γ_{23} both with respect to μ_2 gives us measures $(\gamma_{12,x_2})_{x_2\in\mathcal{X}_2} \subset P(\mathcal{X}_1)$ and $(\gamma_{23,x_2})_{x_2\in\mathcal{X}_2} \subset P(\mathcal{X}_3)$ such that

$$\gamma_{12} = \int_{\mathcal{X}_2} \delta_{x_2}\otimes\gamma_{12,x_2}\,\mathrm{d}\mu_2(x_2) \quad \text{and} \quad \gamma_{23} = \int_{\mathcal{X}_2} \delta_{x_2}\otimes\gamma_{23,x_2}\,\mathrm{d}\mu_2(x_2).$$

Set

$$\gamma := \int_{\mathcal{X}_2} \gamma_{12,x_2}\otimes\delta_{x_2}\otimes\gamma_{23,x_2}\,\mathrm{d}\mu_2(x_2).$$

Then $\gamma \in P(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3)$. For $\phi : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \to [0, \infty]$ we have

$$\int_{\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3} \phi(x_1, x_2, x_3) \, \mathrm{d}\gamma(x_1, x_2, x_3) = \int_{\mathcal{X}_2} \int_{\mathcal{X}_1 \times \mathcal{X}_3} \phi(x_1, x_2, x_3) \, \mathrm{d}\gamma_{12,x_2} \otimes \gamma_{23,x_2}(x_1, x_3) \, \mathrm{d}\mu_2(x_2).$$

If $\phi(x_1, x_2, x_3) = \phi(x_1, x_2)$ is a function only depending on x_1 and x_2, this gives us

$$\begin{aligned}
\int_{\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3} \phi(x_1, x_2) \, \mathrm{d}\gamma &= \int_{\mathcal{X}_2} \int_{\mathcal{X}_1 \times \mathcal{X}_3} \phi(x_1, x_2) \, \mathrm{d}\gamma_{12,x_2} \otimes \gamma_{23,x_2}(x_1, x_3) \, \mathrm{d}\mu_2(x_2) \\
&= \int_{\mathcal{X}_2} \int_{\mathcal{X}_1} \phi(x_1, x_2, x_3) \, \mathrm{d}\gamma_{12,x_2}(x_1) \, \mathrm{d}\mu_2(x_2) \\
&= \int_{\mathcal{X}_1 \times \mathcal{X}_2} \phi(x_1, x_2) \, \mathrm{d}\gamma_{12}(x_1, x_2),
\end{aligned}$$

i.e. γ admits γ_{12} as its marginal on $\mathcal{X}_1 \times \mathcal{X}_2$ (cmp. lemma A.16, as this holds in particular for test functions in $C_b(\mathcal{X}_1 \times \mathcal{X}_2)$). Similarily, one shows that the marginal of γ on $\mathcal{X}_2 \times \mathcal{X}_3$ is γ_{23}, which finishes the proof. □

We are now ready to prove that the Wasserstein distance indeed defines a metric on the Wasserstein space.

Theorem 3.6.6 (Wasserstein Distance is a Metric)**.** *Let $(\mathcal{X}, d)$ be a Polish space and $p \in [0, \infty)$. Then W_p defines a metric on $\mathcal{P}_p(\mathcal{X})$.*

Proof. We need to prove four properties (letting μ_1, μ_2, μ_3 be arbitrary elements of $\mathcal{P}_p(\mathcal{X})$):

(i) W_p is finite and nonnegative,

(ii) $W_p(\mu_1, \mu_2) = 0$ if and only if $\mu_1 = \mu_2$,

(iii) $W_p(\mu_1, \mu_2) = W_p(\mu_2, \mu_1)$,

(iv) $W_p(\mu_1, \mu_3) \leq W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3)$.

Regarding (i), we have for any $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, any $\gamma \in \Pi(\mu, \nu)$ and any $z \in \mathcal{X}$ (cmp. lemma A.10):

$$\begin{aligned}
\int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, \mathrm{d}\gamma(x, y) &\leq 2^p \int_{\mathcal{X} \times \mathcal{X}} d(x, z)^p + d(z, y)^p \, \mathrm{d}\gamma(x, y) \\
&= 2^p \int_{\mathcal{X}} d(x, z)^p \, \mathrm{d}\mu(x) + 2^p \int_{\mathcal{X}} d(z, y)^p \, \mathrm{d}\nu(y) < \infty.
\end{aligned}$$

Also, by definition, W_p is nonnegative. This proves (i).
Regarding (ii), note that the implication "$\mu_1 = \mu_2 \Rightarrow W_p(\mu_1, \mu_2) = 0$ is trivial. For the reverse implication, let $W_p(\mu_1, \mu_2) = 0$. We need to show $\mu_1 = \mu_2$. Let $\gamma \in \Pi(\mu_1, \mu_2)$ be an optimal transport plan (which exists by theorem 3.2.10). Then

$$0 = W_p(\mu_1, \mu_2) = \left(\int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{\frac{1}{p}},$$

which implies γ has to be concentrated on the diagonal $\{(x, y) \in \mathcal{X} \times \mathcal{X} : \ x = y\}$. This means for

any test function $\phi \in C_b(\mathcal{X})$ we have

$$\int_{\mathcal{X}} \phi(x)\,\mathrm{d}\mu_1(x) = \int_{\mathcal{X}\times\mathcal{X}} \phi(x)\,\mathrm{d}\gamma(x,y) = \int_{\mathcal{X}\times\mathcal{X}} \phi(y)\,\mathrm{d}\gamma(x,y) = \int_{\mathcal{X}} \phi(y)\,\mathrm{d}\mu_2(y),$$

which gives us $\mu_1 = \mu_2$ (cmp. lemma A.16). This proves (ii).
Next up, (iii) trivially holds by definition.
Regarding (iv), let $\gamma_{12} \in \Pi(\mu_1,\mu_2)$ and $\gamma_{23} \in \Pi(\mu_2,\mu_3)$ be optimal transport plans (which again exist by theorem 3.2.10). Let $\mathcal{X}_i = \mathcal{X}$ for $i \in [\![3]\!]$ and define γ as in the gluing lemma 3.6.5, i.e.

$$\gamma := \int_{\mathcal{X}_2} \gamma_{12,x_2} \otimes \delta_{x_2} \otimes \gamma_{23,x_2}\,\mathrm{d}\mu_2(x_2).$$

Let γ_{13} be the marginal of γ on $\mathcal{X}_1 \times \mathcal{X}_3$. Then $\gamma_{13} \in \Pi(\mu_1,\mu_3)$, but it is not necessarily optimal. It holds

$$\begin{aligned}
W_p(\mu_1,\mu_3) &\le \left(\int_{\mathcal{X}_1\times\mathcal{X}_3} d(x_1,x_3)^p\,\mathrm{d}\gamma_{13}(x_1,x_3)\right)^{\frac{1}{p}} \\
&= \left(\int_{\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3} d(x_1,x_3)^p\,\mathrm{d}\gamma(x_1,x_2,x_3)\right)^{\frac{1}{p}} \\
&\le \left(\int_{\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3} (d(x_1,x_2)+d(x_2,x_3))^p\,\mathrm{d}\gamma(x_1,x_2,x_3)\right)^{\frac{1}{p}} \\
&\le \left(\int_{\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3} d(x_1,x_2)^p\,\mathrm{d}\gamma(x_1,x_2,x_3)\right)^{\frac{1}{p}} + \left(\int_{\mathcal{X}_1\times\mathcal{X}_2\times\mathcal{X}_3} d(x_2,x_3)^p\,\mathrm{d}\gamma(x_1,x_2,x_3)\right)^{\frac{1}{p}} \\
&= \left(\int_{\mathcal{X}_1\times\mathcal{X}_2} d(x_1,x_2)^p\,\mathrm{d}\gamma_{12}(x_1,x_2)\right)^{\frac{1}{p}} + \left(\int_{\mathcal{X}_2\times\mathcal{X}_3} d(x_2,x_3)^p\,\mathrm{d}\gamma_{23}(x_2,x_3)\right)^{\frac{1}{p}} \\
&= W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3),
\end{aligned}$$

where in the fourth step we used Minkowski's inequality (cmp. proposition A.17). □

An interesting property of the metric space $(\mathcal{P}_p(\mathcal{X}), W_p)$, which we will state here without proof (as the proof requires a lot of preparation), is that it is Polish if $\mathcal{X}$ is. A proof can be found in [25], theorem 6.18.

Theorem 3.6.7. *If $\mathcal{X}$ is a Polish probability space and $p \in [1,\infty)$, then $(\mathcal{P}_p(\mathcal{X}), W_p)$ is a Polish space as well.*

3.7 Discrete Optimal Transport

In this section, we will have a look at how the optimal transport problem changes when both measures are discrete probability measures. This is of particular interest, as we will be dealing with optimal transport between two-dimensional black and white images later on. These images can easily be converted to discrete probability distributions. We will see that the OT problem reduces to a classical linear program, hence all methods and algorithms known in linear programming can be applied in the discrete OT setting as well. This section is partly based on [20].

We will start off with formulating the OT problem in the discrete setting.
Let $\mathcal{X} = \{x_1, ..., x_m\}$ and $\mathcal{Y} = \{y_1, ..., y_n\}$ for some $n, m \in \mathbb{N}_{>0}$. Further let $\alpha \in \Delta^m$ and $\beta \in \Delta^n$ in the probability simplex, and define two measures on $\mathcal{X}$ and $\mathcal{Y}$ via $\mu := \sum_i \alpha_i \delta_{x_i}$, $\nu := \sum_j \beta_j \delta_{y_j}$. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a cost function and set $c_{ij} := c(x_i, y_j)$ for $i \in [\![m]\!]$, $j \in [\![n]\!]$. Then the Kantorovich problem 3.2.1 becomes

Problem 3.7.1 (Discrete Optimal Transport Problem)**.**

$$\begin{aligned} \min_{\gamma \in \mathbb{R}^{m \times n}_{\geq 0}} & \sum_{i,j} c_{ij}\gamma_{ij} \\ \text{s.t.} & \sum_j \gamma_{ij} = \alpha_i \text{ for all } i \in [\![m]\!], \\ & \sum_i \gamma_{ij} = \beta_j \text{ for all } j \in [\![n]\!]. \end{aligned} \tag{8}$$

Note that the discrete equivalent of $\Pi(\mu, \nu)$ becomes

$$\Pi(\mu, \nu) = \left\{ \gamma \in \mathbb{R}^{m \times n}_{\geq 0} : \sum_j \gamma_{ij} = \alpha_i \text{ for all } i \in [\![m]\!], \sum_i \gamma_{ij} = \beta_j \text{ for all } j \in [\![n]\!] \right\}.$$

This set is a polytope, as can more easily be seen from the following reformulation of the problem. In order to rewrite (8) in matrix-vector form, we set $\Gamma := [\gamma_{ij}]_{ij} \in \mathbb{R}^{m \times n}$, $\gamma := \text{vec}(\Gamma) \in \mathbb{R}^{mn}$, $C := [c_{ij}]_{ij} \in \mathbb{R}^{m \times n}$ and $c := \text{vec}(C) \in \mathbb{R}^{mn}$. Furthermore, set (cmp. section 2 for notations)

$$A := \begin{bmatrix} 1_n^\top \otimes I_m \\ I_n \otimes 1_m^\top \end{bmatrix} \in \mathbb{R}^{(m+n) \times mn}, \quad b = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in \mathbb{R}^{m+n}.$$

Then we can reformulate (8) as follows:

$$\begin{aligned} \min_{\gamma \in \mathbb{R}^{mn}} & \langle c, \gamma \rangle \\ \text{s.t. } & A\gamma = b, \\ & \gamma \geq 0, \end{aligned} \tag{9}$$

which is nothing else but a regular linear program (see e.g. equation (1.3) in [5], p. 4). We can also easily verify that the dual optimal transport problem 3.3.1 turns into the dual linear program. If we let $\psi \in \mathbb{R}^m$ and $\phi \in \mathbb{R}^n$ and set $y = \begin{bmatrix} \psi^\top & \phi^\top \end{bmatrix}^\top$, 3.3.1 becomes:

$$\begin{aligned} \max_{y \in \mathbb{R}^{m+n}} & \langle y, b \rangle \\ \text{s.t. } & y^\top A \leq c^\top, \end{aligned}$$

which is exactly the dual linear program to 9 (cmp. [5], p. 143 for a definition of the dual linear program).
Now a standard result in linear programming states that the optimum of a linear program is attained at a vertex of the polytope to be optimized over.[6] This can e.g. be found as Theorem 2.7. in [5], p.

[6]Under some additional assumptions. These assumptions can for example be: There exists an optimal solution, and

65. Using this property, one can prove the following proposition.

Proposition 3.7.2. *Let $P \in \Pi(\mu, \nu)$ be a vertex of the polytope of feasible measures to (8). Then P does not have more than $m + n - 1$ nonzero entries. In particular, there exists an optimal solution to (8) with at most $m + n - 1$ nonzero entries.*

Proof. A proof can for example be found in [20], Proposition 3.4. □

Leveraging this property, there exist multiple algorithms which can solve (8) precisely. For example, the *Network Simplex Algorithm* is a version of the well-known *Simplex Algorithm* which works particularily well in this case, cmp. also [20], chapter 3.5. It makes use of the dual formulation in the discrete case and iteratively improves a feasible solution for the primal until it reaches optimality. Orlin ([18]) was able to prove a polynomial complexity bound on the algorithm which was shortly after improved by Tarjan ([23]) in 1997. However, once $\mathcal{X}$ and $\mathcal{Y}$ become high-dimensional (a few hundred upwards), this algorithm tends to be prohibitively slow. Other algorithms exist, such as the *Hungarian Algorithm* (see [15]) or the *Auction Algorithm* ([4]), but none of them tends to be fast in high-dimensional spaces. This is why Cuturi proposed a different approach in his seminal work [8]: The so-called *Sinkhorn Algorithm*. We will see how it works in the following chapter.

the polytope has at least one vertex. Both of these are fulfilled in our case.

4 Sinkhorn Algorithm

In this section, we will see what the Sinkhorn algorithm is and how it works. We will also discuss its advantages and disadvantages. The idea underlying the algorithm is to introduce an *entropic regularizer* to the optimal transport problem. Let $\mathcal{X}$, $\mathcal{Y}$, m, n, α, β, μ, ν, and c as in the previous subsection 3.7.

4.1 Entropic Optimal Transport

Definition 4.1.1 (Entropy)**.** *For a matrix $P = [p_{ij}]_{ij} \in \mathbb{R}^{m\times n}_{\geq 0}$, we define its* entropy $H(P)$ *as*

$$H(P) := -\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}(\log p_{ij} - 1)$$

if all entries are positive, and $H(P) := -\infty$ if at least one entry is equal to 0*.*

Remark 4.1.2. Note that for transport plans, this notion reduces to $1 - \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\log p_{ij}$. The definition of entropy is not consistent in the literature. We follow the definition in [20]. Other definitions exist, such as $H(P) = -\sum_{i,j} p_{ij}\log p_{ij}$, which can e.g. be found in [8]. Also, the basis of the logarithm does not really matter, as it will only alter the entropy by a constant, and we will see later on that such scaling is irrelevant for our applications. However, the usual convention is to use $\log_2$.

Remark 4.1.3. Entropy is a concept that also exists in physics. There, it measures the randomness or disorder of a system. At maximum entropy, the system reaches a stable state of equilibrium. The mathematical entropy can be thought of in a similar manner: Let's assume P is a transport plan. Then the transport plan of maximal entropy is the trivial coupling (remark 3.1.3), which can be thought of as an equilibrium.

With entropy at hand, we can now define the entropic OT problem.

Problem 4.1.4 (Entropic Optimal Transport Problem)**.** With c, μ, ν as before (cmp. 8) and $\varepsilon > 0$, the entropic optimal transport problem is defined as:

$$P^{\varepsilon}(\mu,\nu) := \min_{\gamma\in\Pi(\mu,\nu)} \langle c,\gamma\rangle - \varepsilon H(\gamma). \tag{10}$$

The term $-\varepsilon H(\gamma)$ is referred to as the entropic regularizer.

Remark 4.1.5. As the entropy is a 1-strongly convex function[7] on all transport plans (since $\partial^2 H(P) = -\mathrm{diag}(\frac{1}{p_{ij}})$ and $p_{ij} \leq 1$), the objective in (10) is ε-strongly convex and hence admits a unique optimal solution.

With ε, we can vary the impact of the regularizer on the solution. As $\varepsilon \to \infty$, the unique solution to (10) converges to the transport plan of maximum entropy, which, as we just saw, is the trivial coupling, and as $\varepsilon \to 0$, the solution to (10) indeed converges to the maximum entropy optimal coupling of the unregularized problem.[8]

[7] A function f is called l-strongly convex if $l\,\|x-y\|_2^2 \leq (\nabla f(x) - \nabla f(y))^{\top}(x-y)$ for all x, y.

[8] We will refer to the regular optimal transport problem as the *unregularized* one, while problem 4.1.4 will interchangeably be called the *entropic* or the *regularized* problem.

Proposition 4.1.6. *Let γ_ε be the unique solution to problem 4.1.4 for $\varepsilon > 0$. Then*

$$\gamma_\varepsilon \xrightarrow{\varepsilon \to 0} \arg\min \left\{ -H(\gamma) : \ \gamma \in \Pi(\mu, \nu), \ \langle \gamma, c \rangle = \min_{\gamma' \in \Pi(\mu,\nu)} \langle \gamma', c \rangle \right\}$$

and

$$\gamma_\varepsilon \xrightarrow{\varepsilon \to \infty} \alpha \beta^\top.$$

Proof. A proof can e.g. be found in [20], proposition 4.1. □

Remark 4.1.7. With a different definition to the entropic optimal transport problem, Cuturi [8] showed that, similar to the Wasserstein distances, the optimal value of the entropic problem defines a metric on the distributions (cmp. [8], theorem 1; a slight modification in multiplying the value by $\mathbb{1}_{\mu \neq \nu}$ is needed in order to ensure that it takes the value 0 if and only if $\mu = \nu$).

Definition 4.1.8 (Gibbs Kernel)**.** *For a cost function $c \in \mathbb{R}^{m \times n}$ and a regularizing constant $\varepsilon > 0$, we define the* Gibbs kernel $K \in \mathbb{R}^{m \times n}$ *of c via*

$$K_{ij} = \exp(-\frac{c_{ij}}{\varepsilon}), \quad i \in [\![m]\!], \ j \in [\![n]\!].$$

The unique solution of (10) always takes a particular form.

Proposition 4.1.9 (Solution of Entropic Optimal Transport)**.** *The solution of (10) is unique and takes the form*

$$\gamma_{ij} = u_i K_{ij} v_j, \quad i \in [\![m]\!], \ j \in [\![n]\!],$$

where K is the Gibbs kernel and $u \in \mathbb{R}^m_{>0}$ and $v \in \mathbb{R}^n_{>0}$ are two positive vectors uniquely defined up to a scaling constant (i.e. scaling u by some $\lambda > 0$ and v by $\frac{1}{\lambda}$).

Proof. The proof of this proposition can e.g. be found in [20], proposition 4.3. □

As does the unconstrained problem, the entropic version also comes with its own dual problem.

Problem 4.1.10 (Entropic Dual Problem)**.** The dual problem is defined as (again α and β being the vectors representing μ and ν as in 8):

$$D^\varepsilon(\mu, \nu) := \max_{f \in \mathbb{R}^m, \ g \in \mathbb{R}^n} \langle f, \alpha \rangle + \langle g, \beta \rangle - \varepsilon \langle e^{f/\varepsilon}, K e^{g/\varepsilon} \rangle.$$

Proposition 4.1.11. *There exists an optimal solution to the dual 4.1.10 and there is duality, i.e.*

$$P^\varepsilon(\mu, \nu) = D^\varepsilon(\mu, \nu).$$

Furthermore, vectors u and v as in proposition 4.1.9 and optimal f, g for problem 4.1.10 are linked via

$$(u, v) = (e^{f/\varepsilon}, e^{g/\varepsilon}).$$

Proof. A short proof can e.g. be found in [20], proposition 4.4. □

Remark 4.1.12. Again, the solution to the dual is not unique, as one can replace f by $f - k$ and g by $g + k$ for a constant $k \in \mathbb{R}$. Note that the link between the scaling vectors and the dual solution

enables us to recover the optimal transport plan γ for the entropic primal from a solution (f, g) to the dual:

$$\gamma_{ij} = e^{(f_i+g_j-c_{ij})/\varepsilon},$$

which follows immediately from the representation of γ in proposition 4.1.9 and proposition 4.1.11. This is a property that the unregularized problem did not have.

A solution (f, g) of the unregularized problem 3.3.1 approximates the solution of the regularized dual in the following sense.

Proposition 4.1.13. *Let (f, g) be a solution to the unregularized dual problem 3.3.1 and $(f^\star, g^\star)$ a solution to the regularized dual problem 4.1.10 for some $\varepsilon > 0$. Then $(f^\star, g^\star)$ is feasible for the unregularized problem, i.e. $f^\star + g^\star \leq c$, and*

$$0 \leq D^\varepsilon(\mu, \nu) - \left[\langle f, \alpha\rangle + \langle g, \beta\rangle - \varepsilon\langle e^{f/\varepsilon}, Ke^{g/\varepsilon}\rangle\right] \leq mn\varepsilon,$$

i.e. (f, g) in the entropic dual differs from the optimal value by at most a factor of $mn\varepsilon$. In particular, if $\varepsilon \to 0$, the optimum of the entropic dual converges to its value at (f, g), and the value the unregularized dual takes at $(f^\star, g^\star)$ converges to its optimum, i.e.

$$\langle f^\star, \alpha\rangle + \langle g^\star, \beta\rangle \xrightarrow{\varepsilon\to 0} \langle f, \alpha\rangle + \langle g, \beta\rangle. \tag{11}$$

Proof. Let γ be the solution of the entropic primal problem. As we have

$$1 \geq \gamma_{ij} = e^{(f_i^\star+g_j^\star-c_{ij})/\varepsilon} \text{ for all } i \in [\![m]\!],\ j \in [\![n]\!],$$

it follows that $f_i^\star + g_j^\star - c_{ij} \leq 0$ for all i and j, i.e. $f^\star + g^\star \leq c$. This makes $(f^\star, g^\star)$ feasible for the unregularized dual. From optimality of (f, g) we get

$$\langle f, \alpha\rangle + \langle g, \beta\rangle \geq \langle f^\star, \alpha\rangle + \langle g^\star, \beta\rangle.$$

This gives us

$$\begin{aligned}
&D^\varepsilon(\mu, \nu) - \left[\langle f, \alpha\rangle + \langle g, \beta\rangle - \varepsilon\langle e^{f/\varepsilon}, Ke^{g/\varepsilon}\rangle\right] \\
=&\langle f^\star, \alpha\rangle + \langle g^\star, \beta\rangle - \varepsilon\langle e^{f^\star/\varepsilon}, Ke^{g^\star/\varepsilon}\rangle - \left[\langle f, \alpha\rangle + \langle g, \beta\rangle - \varepsilon\langle e^{f/\varepsilon}, Ke^{g/\varepsilon}\rangle\right] \\
\leq&\varepsilon\left[\langle e^{f/\varepsilon}, Ke^{g/\varepsilon}\rangle - \langle e^{f^\star/\varepsilon}, Ke^{g^\star/\varepsilon}\rangle\right] \\
\leq&\varepsilon\sum_{i,j} e^{(f_i+g_j-c_{ij})/\varepsilon} \leq mn\varepsilon,
\end{aligned}$$

where in the last step we used the fact that $f + g \leq c$. Also note that the starting expression is always greater or equal to 0 by optimality of $(f^\star, g^\star)$. This also implies (11), as

$$\begin{aligned}
0 &\geq \langle f^\star, \alpha\rangle + \langle g^\star, \beta\rangle - \langle f, \alpha\rangle + \langle g, \beta\rangle \\
&\geq -\varepsilon\left[\langle e^{f/\varepsilon}, Ke^{g/\varepsilon}\rangle - \langle e^{f^\star/\varepsilon}, Ke^{g^\star/\varepsilon}\rangle\right] \geq -\varepsilon mn.
\end{aligned}$$

□

4.2 Sinkhorn Algorithm

Since the optimal solution γ needs to fulfill the marginal constraints, the following equalities need to hold (where $\odot$ denotes the entry-wise vector multiplication):

$$u \odot Kv = \alpha,$$
$$v \odot K^\top u = \beta.$$

These equations lie at the heart of the *Sinkhorn-Knopp fixpoint iteration*, which iteratively finds the solution from proposition 4.1.9. Upon initialization of $v^0 \in \mathbb{R}^n_{>0}$ (which can be initialized arbitrarily or according to some initialization scheme), the updates of the algorithm are as follows:

$$u^{l+1} = \frac{\alpha}{Kv^l}, \quad v^{l+1} = \frac{\beta}{K^\top u^{l+1}}, \quad l = 0, 1, 2, ..., \tag{12}$$

where the fractions are to be understood as element-wise division. These iterations indeed converge to the optimal solution.

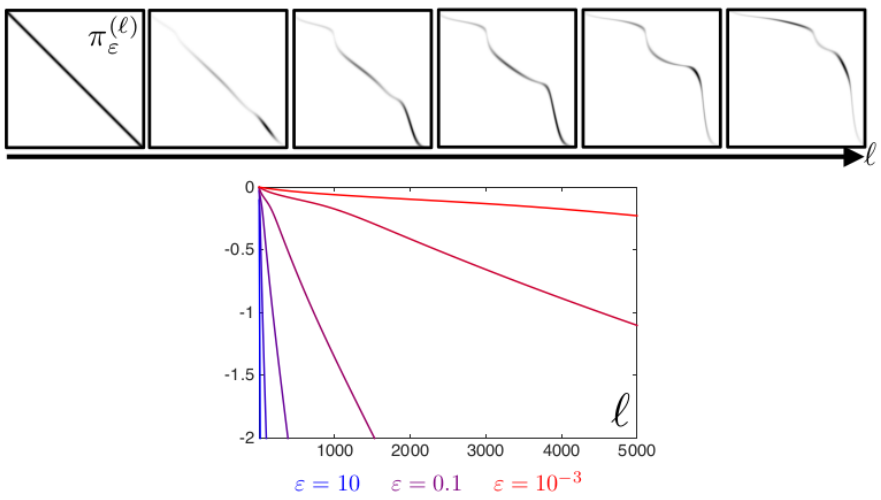

Figure 3: Top: Evolution of $\pi_\varepsilon^{(l)} := \mathrm{diag}(u^l) K \mathrm{diag}(v^l)$ as l increases, for $\varepsilon = 0.1$, $c(x, y) = |x - y|^2$ and one-dimensional distributions on $[0, 1]$. Bottom: Impact of the regularizing constant ε on the convergence rate of the algorithm, measured in terms of the marginal contraint violation on β, $\log\left(\left\|1_n^\top \pi_\varepsilon^{(l)} - \beta^\top\right\|_1\right)$.[9]

Proposition 4.2.1. *The iterates u^l and v^l as in (12) converge to the up to a constant uniquely defined u and v in proposition 4.1.9 as $l \to \infty$.*

Proof. This statement goes back to the original work of Sinkhorn and Knopp, [22], where convergence is proven. □

Remark 4.2.2. Note that with varying initilizations v^0, the limit points of u^l and v^l will also change as they are only unique up to a constant. However, the limit point $\lim_{l\to\infty} \mathrm{diag}(u^l) K \mathrm{diag}(v^l)$ will always be the unique solution from proposition 4.1.9.

[9] Source: [20], figure 4.5.

This algorithm lies at the heart of the seminal work of Cuturi [8]. The algorithm from his paper can be seen in figure 4, while the equivalent algorithm using our iteration as above can be seen in algorithm 1.

Algorithm 1 Sinkhorn Algorithm

in $c \in \mathbb{R}^{m\times n}$, $\varepsilon > 0$, $\alpha \in \Delta^m_{>0}$, $\beta \in \Delta^n_{>0}$
initialize v^0 (e.g. $v^0 \leftarrow 1_n$), $l \leftarrow 0$, $K \leftarrow e^{-c/\varepsilon}$
repeat
    $u^{l+1} \leftarrow \alpha./Kv^l$
    $v^{l+1} \leftarrow \beta./K^\top u^{l+1}$
until stopping criterion is met
$\gamma \leftarrow \mathrm{diag}(u^l)K\mathrm{diag}(v^l)$
out γ, $\langle \gamma, c\rangle$

Algorithm 1 Computation of $d^\lambda_M(r,c)$ using Sinkhorn-Knopp's fixed point iteration

Input `M`, λ, `r`, `c`.
`I=(r>0); r=r(I); M=M(I,:); K=exp(-λ*M)`
Set `x=ones(length(r),size(c,2))/length(r);`
while `x` changes **do**
  `x=diag(1./r)*K*(c.*(1./(K'*(1./x))))`
end while
`u=1./x; v=c.*(1./(K'*u))`
d^λ_M`(r,c)=sum(u.*((K.*M)*v))`

Figure 4: Sinkhorn algorithm as in [8].

He uses M for c, λ for $\frac{1}{\varepsilon}$ and r, c for α, β. Note that his algorithm is equivalent to what we have established; what he calls x in the iteration is $\frac{1}{u}$ for us, and he performs both updates of the iteration in a single line. In practice, when using this algorithm to compute transport costs, one usually implements fitting stopping criteria such as the violations on the marginal constraints, as can also be seen in the description of figure 3.

Cuturi convincingly showed that this approach can drastically speed up computations of optimal transport costs, in particular in higher dimensions. One main advantage is that it allows for efficient parallelization of multiple optimal transport problems at once: Given a fixed cost and regularizer ε, and a collection of pairs of distributions $(\alpha_i, \beta_i)_{i\in[\![k]\!]}$, writing $\mathrm{A} := \begin{bmatrix}\alpha_1 & \dots & \alpha_k\end{bmatrix}$, $\mathrm{B} := \begin{bmatrix}\beta_1 & \dots & \beta_k\end{bmatrix}$ we can solve these problems simultaneously by initializing some $V^0 \in \mathbb{R}^{n\times k}_{>0}$ and updating $U, V \in \mathbb{R}^{n\times k}_{>0}$ as follows:

$$U^{l+1} = \frac{\mathrm{A}}{KV^l}, \quad V^{l+1} = \frac{\mathrm{B}}{K^\top U^{l+1}}, \quad l = 0, 1, 2, ...$$

Furthermore, Sinkhorn's algorithm allows for computations of optimal transport costs that are differentiable in the inputs. However, the Sinkhorn algorithm comes with a few drawbacks and pitfalls. One obvious such drawback is that the algorithm merely computes the solution to the entropic optimal transport problem and not the unregularized one. One can argue that in certain cases, this is actually desirable as solutions that come from the regularized problem oftentimes come closer to what can be observed in real life, such as traffic flow patters. Also, choosing the regularizer sufficiently small

will ensure solutions that are close to the unregularized one. However, when ε gets too small, this might result in entries of K being stored as zeros due to numerical rounding errors, which in turn can cause a division by 0 in the iterative updates from (12). For similar reasons, Sinkhorn does not support computations on distributions that contain zeros or very small values. To some extend, these problems can be dealt with by shifting computations to the log domain. Details can be found in [20], section 4.4.

4.3 Initializing Sinkhorn's Algorithm

We have seen that the entropic optimal transport problem admits a unique solution, and that the iterates from the Sinkhorn algorithm converge to the vectors u and v corresponding to that solution. Hence, one might think that it does not matter how v^0 is initialized, as the iterates are guaranteed to converge anyways. While it is true that convergence is guaranteed no matter the initialization, initialization matters in terms of *convergence speed*. If for example v^0 is already very close to v from the optimal solution (more precisely: $a\ v$, as v is only uniquely defined up to a constant), this initialization will lead to much faster convergence than a random one.[10] Comparatively little attention has been paid to improving initializion of Sinkhorn's algorithm. Thornton and Cuturi [24] propose using dual vectors recovered from the unregularized 1D optimal transport problem, or from known transport maps in a Gaussian setup, and were able to significantly speed up convergence. Amos et al. [2] use a learned approach, where a neural network learns to approximate one of the two dual potentials of the (unregularized) OT problem which can then be fed into Sinkhorn's algorithm as an initialization. While this idea is in parts similar to what we will propose in the following chapter, there are two key differences: Firstly, Amos et al. use a loss that's based on the Wasserstein distance approximation that the dual potential approximation yields (we will see how exactly in section 5.4); this has the clear advantage that you can simply attempt to minimize the loss (i.e. the negative of the Wasserstein distance approximation) *without* having to know the ground truth, i.e. the actual Wasserstein distance. However, using a loss on the Wasserstein distance instead of one on the potential directly means that vital information on how the potential looks like can be lost resulting in less accurate approximations of the potential. This might be the reason for the second key difference: In Amos et al. [2], such networks are only trained for very specific datasets such as MNIST as their intrinsic structure allows for much easier approximations of the potential. We will show that one can actually train a *universal* network which is not dataset-dependent using a loss on the dual potential.

[10] While intuitively clear, this will also become empirically evident in section 6.

5 A Trained Initialization for the Sinkhorn Algorithm

We now present our Sinkhorn-NN hybrid algorithm. The main idea is: Train a neural network to predict the dual potential of the unconstrained optimal transport problem and then use that to initialize the Sinkhorn algorithm. First, we will have a a more detailed look at the idea itself, and then at its implementation. Let $\mathcal{X}$, $\mathcal{Y}$, m, n, α, β, μ, ν, and c as in subsection 3.7 on discrete optimal transport and $K = e^{-c/\varepsilon}$ for $\varepsilon > 0$.

5.1 Sinkhorn-NN Hybrid Algorithm

Ultimately, we want to be able to quickly approximate optima for discrete optimal transport problems. In light of proposition 4.1.6, the entropic optimal transport problem 4.1.4 is a reasonable approximation of the regular problem 8 for $\varepsilon > 0$ small enough. An efficient way to approximate the solution to the entropic problem is the Sinkhorn algorithm 1. It converges to a tuple (u, v) of vectors from which an optimal transport plan γ to the entropic problem can be recovered via $\gamma = \mathrm{diag}(u)K\mathrm{diag}(v)$, see proposition 4.2.1. Usually, the Sinkhorn algorithm is initialized with the 1-vector, as convergence is guaranteed. However, more precise initializations can lead to much quicker convergence, as we will see. In light of proposition 4.1.13, it is reasonable to believe that a solution (f, g) of the dual optimal transport problem 3.3.1 can be used to compute a good starting vector v^0 via $v^0 = e^{g/\varepsilon}$, as we know that the limit point v of the algorithm can be written as $v = e^{g^\star/\varepsilon}$ for a solution $(f^\star, g^\star)$ of the entropic dual problem 4.1.10, cmp. proposition 4.1.11. Hence, we will let a neural network learn to approximate the function $(\alpha, \beta) \mapsto (f, g)$ which maps two distributions to the solution of the unregularized dual problem. In fact, it suffices to consider the function $(\alpha, \beta) \mapsto f$ as we know that at optimality, we can recover g from f via $g = f^c$, cmp. theorem 3.5.1.[11] We will do so using training data containing ground truth dual potentials. After training is finished, given two distributions $\alpha \in \Delta^m_{>0}$ and $\beta \in \Delta^n_{>0}$, we can compute $f \approx \mathrm{net}(\alpha, \beta)$, $g = f^c$, $v^0 = e^{g/\varepsilon}$ for a given $\varepsilon > 0$, and use this vector v^0 as a starting vector the for Sinkhorn algorithm. A pseudocode of the Sinkhorn-NN hybrid algorithm can be seen in algorithm 2.

Algorithm 2 Sinkhorn-NN Hybrid Algorithm

in $c \in \mathbb{R}^{m \times n}$
generate training data $d = (\alpha_{\text{train}}, \beta_{\text{train}}, f_{\text{train}}) \in \Delta^m \times \Delta^n \times \mathbb{R}^m$
initialize net_θ with $(m+n)$-dim. input and m-dim. output with parameters θ
train net_θ on d with $\mathrm{loss}(\alpha_{\text{train}}, \beta_{\text{train}}) \leftarrow \mathrm{MSE}(\mathrm{net}_\theta(\alpha_{\text{train}}, \beta_{\text{train}}), f_{\text{train}})$
in $\varepsilon > 0$, $\alpha \in \Delta^m_{>0}$, $\beta \in \Delta^n_{>0}$
$g \leftarrow c$-transform$(\mathrm{net}_\theta(\alpha, \beta))$
$v^0 \leftarrow e^{g/\varepsilon}$, $l \leftarrow 0$, $K \leftarrow e^{-c/\varepsilon}$
repeat
    $u^{l+1} \leftarrow \alpha./Kv^l$
    $v^{l+1} \leftarrow \beta./K^\top u^{l+1}$
until stopping criterion is met
$\gamma \leftarrow \mathrm{diag}(u^l)K\mathrm{diag}(v^l)$
out γ, $\langle \gamma, c \rangle$

[11] We could also directly approximate g instead, as this is the vector used to initialize v. However, in practice, approximating f and recovering g from it yields better results due to the fact that this ensures that g is c-concave (by definition of c-concavity, see definition 3.3.3), as opposed to only being an approximation of a c-concave function.

Note that once the training in steps $1-4$ has completed, steps $5-13$ can be repeated, i.e. once a network is trained, it can be used for all future Sinkhorn computations for that dimension and cost matrix.
In what follows, we will have a closer look at some of the implementation details, such as how training data is generated or what network architecture is used. The PyTorch implementation can be found at `https://github.com/j-geuter/DualOTComputations`.

5.2 Training Data

As we want the algorithm to be able to generalize to any type of data, we have to make sure that our training data is rich enough to capture the structure of the entire function $\Delta^m \times \Delta^n \to \mathbb{R}^m$, $(\alpha, \beta) \mapsto f$, and not only train the network on a small subset of $\Delta^m \times \Delta^n$. Thus, we cannot train on a specific dataset like MNIST. Setting each of the m resp. n data points in a training sample to a random number between 0 and 1 and then normalizing the sample to sum to 1 works in theory in generating training data that captures the entire domain of the function, but does not work particularily well in practice for two reasons: First, one should prevent data points from being 0 or close to 0. Not only does the Sinkhorn algorithm require the distributions to be strictly positive everywhere, but also do zeros make the dual potentials more arbitrary: If $\alpha_i = 0$ for some i, then changing f_i does not alter the value of $\langle f, \alpha\rangle + \langle f^c, \beta\rangle$. Thus, enforcing all data points to be larger than some small threshold larger than 0 ensures some kind of uniqueness of the dual potential, and is not restrictive of the problem as the Sinkhorn algorithm only works on strictly positive data anyways.[12] Second, particularily in larger dimensions, this procedure of producing samples will make distributions tend to be very similar in the sense that their mass is roughly evenly distributed across all datapoints, which leads to approximately equal transport distances across all samples. This problem can be alleviated as follows: Letting r be a random number in $[0, 1]$, instead of setting a datapoint to r we will set it to r^k for some $k > 1$. This means the data point values are not evenly spread between 0 and 1 anymore but tend towards the extremes 0 and 1, making the distribution's mass concentrate on fewer data points. This ensures that the transport distances between samples are sufficiently large, and the samples differ more distinctly from one another. Another important factor in producing data is the following: As we have seen, the dual value $\langle f, \alpha\rangle + \langle f^c, \beta\rangle$ is invariant under adding a constant to f, so we will only use potentials that sum to 0 to ensure some kind of uniqueness with respect to this constant. Combining these ideas, the data generation procedure can be seen in algorithm 3.
Importantly, adding k_2 should happen *after* exponentiating by k_1 to ensure no data points are too close to 0 in the end. The dual potential f can be computed using one of the precise algorithms mentioned in subsection 3.7, for example the network simplex algorithm.
Now, we will have a look at how precisely we implemented this idea. We will be dealing with 28×28-dimensional distributions only, i.e. 784-dimensional data. However, the algorithm can easily be transferred to data of different dimension. Empirically, in this case, $k_1 = 3$ and $k_2 = 0.001$ are good choices, and these are the constants we used for all training data generation.[13]

[12] As a random number between 0 and 1 is almost surely greater than 0 anyways, and a finite number of random numbers will be larger than a positive threshold, this property holds automatically. However, we found adding a small constant to all datapoints to slightly improve learning, as this lets us control the threshold. If training on specific datasets that contain zeros by default – such as MNIST – adding a constant vastly improves learning.

[13] Note that these values are very specific to the sample size. With 14×14-dimensional data, for instance, $k_1 = 2$ proved to be better.

Algorithm 3 Training Data Generation

in $k_1 > 1$, $k_2 > 0$
data $\leftarrow$ list()
for $i = 1, 2, ..., N$ **do**
    $\alpha \in [0,1]^m$, $\beta \in [0,1]^n$ random
    $\alpha \leftarrow \alpha^{k_1}$, $\beta \leftarrow \beta^{k_1}$
    $\alpha \leftarrow \alpha + k_2$, $\beta \leftarrow \beta + k_2$
    $\alpha \leftarrow \frac{\alpha}{\sum_i \alpha_i}$, $\beta \leftarrow \frac{\beta}{\sum_i \beta_i}$
    $f \leftarrow \text{DualPotential}(\alpha, \beta)$
    $f \leftarrow f - \frac{\sum_i f_i}{m}$
    append(data,(α, β, f))
end for
out data

The cost matrix used throughout all experiments is the squared Euclidean distance, i.e. the cost to get from α_{ij} (considering α to be 28×28-dimensional instead of 784-dimensional) to $\beta_{i'j'}$ is $(i-i')^2+(j-j')^2$. This means the optimal transport costs correspond to the squared Wasserstein-2 distances, cmp. definition 3.6.1. This is a very common choice,[14] but any other cost function could have also been used.[15] Figure 5 visualizes the training data and compares it to the case $k_1 = k_2 = 0$.

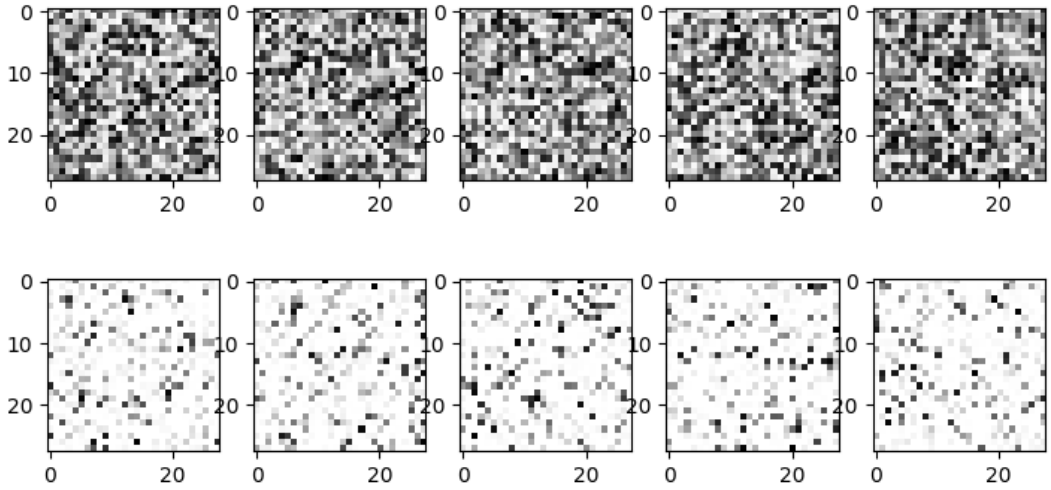

Figure 5: Top: data with $k_1 = k_2 = 0$. Bottom: our training data with $k_1 = 3$ and $k_2 = 0.001$.

[14] Another common choice would be the regular Euclidean distance, yielding the Wasserstein-1 distance. As we have seen in remark 3.6.3, the Wasserstein-1 distance admits a very unique structure in the optimal transport problem, where the value of the dual only depends on $\mu - \nu$ and not on both measures separately. This fact is often exploited, e.g. in the famous Wasserstein GAN paper [3]. However, we neither need nor want to use this property, as our algorithm also works in more general settings. Yet, we want to mention that in case this cost matrix is used, one could utilize the additional structure of the problem by feeding $\alpha - \beta$ into the network instead of (α, β), resulting in an input size of 784 instead of 1568.

[15] However, we will see in the following subsection that we make some use of the fact that the Wasserstein distance is a metric. With a different cost function, we might lose this property; but it is not essential to the algorithm and only used for fine-tuning the network.

5.3 Network Architecture

The network is a very simple feed-forward neural network. It is medium-sized, which proved to be superior over larger, deeper networks as well as over smaller networks. It consists of three layers, the first with $2 \cdot 784 = 1568$ in- and $6 \cdot 784 = 4704$ outputs, the second $6 \cdot 784$ in- and outputs, and the last $6 \cdot 784$ in- and 784 outputs. This totals roughly 33 million trainable parameters, which results in training taking approximately 10 minutes per 100.000 samples on a 4 core, 1.6 GHz CPU. The first two layers contain a batch normalization layer ([12]) and a ReLU activation function (rectified linear unit), whereas the last layer comes without either. The loss is the mean squared error loss (MSE) on the dual potential with the Adam optimizer function, see [14], and varying learning rates depending on the experiment. Since we know that our optimal transport distances are equal to the squared Wasserstein-2 distance which is a metric (cmp. theorem 3.6.6), we know it is symmetric. Symmetry is easy to enforce: Instead of returning $\text{net}(\alpha, \beta)$, we can return $\frac{\text{net}(\alpha,\beta)+\text{net}(\beta,\alpha)^c}{2}$, i.e. computing the network's output for (α, β) and then switching the order of the input distributions. Note that we have to compute the c-transform of $\text{net}(\beta, \alpha)$ as if we switch the distributions' order, this means instead of maximizing $\langle \cdot, \alpha \rangle + \langle \cdot, \beta \rangle$ we now maximize $\langle \cdot, \beta \rangle + \langle \cdot, \alpha \rangle$, and if (f, g) maximizes the former, we have $g = f^c$ by theorem 3.5.1 and the maximum equals $\langle f, \alpha \rangle + \langle f^c, \beta \rangle$, with f appearing in the first scalar product. However, this means (g, f) is optimal for the latter, again with $g = f^c$, and $\langle f^c, \beta \rangle + \langle f, \alpha \rangle$ is the optimum, with f^c appearing first. Hence, when switching the order of the distributions, we need to consider the c-transform of the network's output. In practice, enforcing symmetry during training did not improve performance, but switching it on after training was over reduced the error on the dual potential by up to 10% (as outlined in algorithm 2, we use the mean squared error (MSE)).
Furthermore, we know that the network's output should be a c-concave function. Again, this is easy to enforce: Instead of returning $\text{net}(\alpha, \beta)$, we can return its '$c$-concavification' $\text{net}(\alpha, \beta)^{cc}$ instead which can be thought of as the c-concave approximation of a non-c-concave function, also cmp. proposition 3.3.7. Again, this did not improve performance if turned on during training, but did improve results if turned on for testing. However, note that we chose to learn f instead of g in algorithm 2 for precisely this reason: As we need g to initialize v for the Sinkhorn part, we need to compute it via $g = f^c$ which makes it c-concave already, so we get a c-concave result 'for free'.
Note that we can also use this network to approximate the squared Wasserstein-2 distance directly, without having to feed the outputs to the Sinkhorn algorithm, by computing

$$W_2^2(\alpha, \beta) \approx \langle \text{net}(\alpha, \beta), \alpha \rangle + \langle \text{net}(\alpha, \beta)^c, \beta \rangle$$

. The MSE-error on the Wasserstein distance approximation by the network could even be reduced by up to 50% by each of the two ideas – symmetry and c-concavification – respectively.[16] As enforcing the metric properties of the Wasserstein-2 distance proved successful with respect to its symmetry, one might wonder if the network's performance could also somehow benefit from enforcing the other two metric properties, the triangle inequality and the metric being zero if and only if its inputs are identical. However, both of these are not easy to enforce, unlike the symmetry. Regarding the latter of the two, one could attempt to at least add training samples of pairs of identical distributions to the training data; however, this did not prove to improve performance.

[16] At this point, one might wonder why even include the Sinkhorn part then, and not use the network on its own to approximate the optimal transport cost. We will investigate this question in section 5.4.

5.4 Why Not...?

By now, there are a few important questions one might be wondering which we have to address. These include: Why don't we use the Sinkhorn algorithm in creating data, as this would give us the potential from the regularized problem in the training data, which is the actual limit point of the Sinkhorn algorithm? Why don't we use the network's predictions directly to approximate the transport cost, instead of feeding its outputs to the Sinkhorn algorithm? And why don't we use a loss on the transport cost approximation computed from the network's output, which would let us train the network without needing ground truth potentials? In the following, we will answer these questions.

Why not use the Sinkhorn algorithm in creating training data?
We want the network to compute good starting vectors for the Sinkhorn algorithm. The Sinkhorn algorithm converges to the potential of the entropic dual problem 4.1.10. So why don't we use potentials from the entropic problem in our training data instead? One the one hand, we know by proposition 4.1.13 that the solution of the unregularized dual problem 3.3.1 approximates the entropic dual. On the other hand, if we did use the entropic potential in our training data, we would have to fix a regularizer $\varepsilon > 0$ for generating the data. This might increase convergence speed for *that particular* regularizer. But we want a universal network which can be used with the Sinkhorn algorithm with varying regularizers.

Why not use the network directly for approximating transport distances?
Our network approximates the dual potential f, and given an optimal f we know by the duality theorem 3.5.1 that

$$\min_{\gamma \in \Pi(\alpha,\beta)} = \langle f, \alpha \rangle + \langle f^c, \beta \rangle.^{17}$$

This means we could approximate the transport distance between α and β by calculating $\langle \text{net}(\alpha,\beta), \alpha \rangle + \langle \text{net}(\alpha,\beta)^c, \beta \rangle$. So why do we need to feed $\text{net}(\alpha,\beta)$ into the Sinkhorn algorithm in order to approximate this exact same quantity? First, one nice aspect of the entropic problem is that we can recover the transport plan from the dual solution (or from the scaling vector returned by Sinkhorn), cmp. proposition 4.1.9 and remark 4.1.12. This would not be possible if we used only the network. Second, and more importantly, the function $(\alpha,\beta) \mapsto f$ is very complex and hard to learn for a neural network. This means the approximations computed by the network are not very accurate. They work very well in accelerating the Sinkhorn algorithm, but on their own, they do not approximate the dual potential to a satisfying degree. ADD FIGURE

Why not use a loss on the transport distance?
For a sample (α, β, f) from the training data, the loss we use is

$$\text{loss}(\alpha,\beta) = \text{MSE}(\text{net}(\alpha,\beta), f).$$

[17] Slightly abusing notation; by $\Pi(\alpha,\beta)$ we of course refer to the set of transport plans between the measures *corresponding* to α and β, as these are merely vectors.

However, since we know that the expression $\langle \cdot, \alpha \rangle + \langle \cdot, \beta \rangle$ is maximized on optimal potentials, and at optimality the second potential is the c-transform of the first (cmp. theorem 3.5.1), we could also use

$$\text{loss}(\alpha, \beta) = -\langle \text{net}(\alpha, \beta), \alpha \rangle + \langle \text{net}(\alpha, \beta)^c, \beta \rangle$$

as a loss. This is in fact the loss used by Amos et al. [2]. Using this loss has a distinct advantage over our choice: It does not require ground truth optimal potentials f in the training data. This means we do not have to solve the optimal transport problem between the two distributions of each training sample for generating training data. However, as can be seen in ADD FIGURE this leads to significantly worse approximations on the potential than our approach. This might be the reason that no universal network is presented in [2], but only one trained on a specific dataset like MNIST. In order to learn a universal network able to generalize to any data, a loss on the potential is to be preferred. This means that training data generation takes more time, but for a given distribution size, training data needs to be created just *once*; afterwards, it can be used for various networks.

6 Results

7 Discussion

A Appendix

A.1 Measure Theory

In this section we recall some definitions and properties from measure theory that are used throughout the thesis, in particular in chapter 3 on optimal transport. For definitions not provided here, one might resort to introductions to measure theory such as the one provided in [11].

Definition A.1 (Measure, Finite Measure, Discrete Measure, Signed Measure)**.** *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space.*[18] *A* measure *on $(\mathcal{X}, \mathcal{A})$ is a map $\mu : \mathcal{A} \to [0, \infty]$ such that $\mu(\emptyset) = 0$ and*

$$\mu\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} \mu(A_i)$$

for every countable collection $(A_i)_{i\in\mathbb{N}} \subset \mathcal{A}$ of pairwise disjoint sets in $\mathcal{A}$. The triple $(\mathcal{X}, \mathcal{A}, \mu)$ is called a measure space.
The measure is called finite *if $\mu(\mathcal{X}) < \infty$. It is called* discrete *if it is concentrated on a countable set, i.e. there exist $(x_i)_{i\in\mathbb{N}}$ such that $\mu(\mathcal{X} \setminus (\cup_{i\in\mathbb{N}}\{x_i\})) = 0$.*
A signed measure *on $(\mathcal{X}, \mathcal{A})$ is a map $\nu : \mathcal{A} \to [-\infty, \infty]$ such that $\nu(\emptyset) = 0$, ν takes at most one of the two values $-\infty$ and ∞ on all of $\mathcal{A}$, and*

$$\nu\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} \nu(A_i)$$

for every countable collection $(A_i)_{i\in\mathbb{N}} \subset \mathcal{A}$ of pairwise disjoint sets in $\mathcal{A}$.

Definition A.2 (Mutually Singular Measures)**.** *Let μ and ν be two measures on a measurable space $(\mathcal{X}, \mathcal{A})$. The measures μ and ν are called* mutually singular*, written $\mu \perp \nu$, if there exist two disjoint sets $\mathcal{X}_\mu$ and $\mathcal{X}_\nu$ such that $\mathcal{X} = \mathcal{X}_\mu \cup \mathcal{X}_\nu$ and for every $A \in \mathcal{A}$ we have*

$$\mu(A) = \mu(A \cap \mathcal{X}_\mu), \quad \nu(A) = \nu(A \cap \mathcal{X}_\nu).$$

Theorem A.3 (Jordan Decomposition)**.** *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and μ a signed measure on $(\mathcal{X}, \mathcal{A})$. Then there exists a unique pair (μ^+, μ^-) of mutually singular measures on $(\mathcal{X}, \mathcal{A})$, one of which is finite, such that $\mu = \mu^+ - \mu^-$.*

Proof. A proof can e.g. be found in [10], theorem 2. □

Definition A.4 (Upper Variation, Lower Variation, Total Variation)**.** *Let μ be a signed measure on a measurable space $(\mathcal{X}, \mathcal{A})$ and (μ^+, μ^-) its Jordan decomposition. Then μ^+ is called the* upper variation *of μ, μ^- its* lower variation *and $\|\mu\| := \mu^+ + \mu^-$ its* total variation*.*

Definition A.5 (Product σ-Algebra)**.** *Given two measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, the* product σ-algebra *of $\mathcal{A}$ and $\mathcal{B}$ is the σ-algebra generated by all sets of the form $A \times B$ for $A \in \mathcal{A}$ and $B \in \mathcal{B}$. It is denoted by $\mathcal{A} \otimes \mathcal{B}$.*

[18] A *measurable space* is a nonempty set $\mathcal{X}$ together with a σ-algebra $\mathcal{A}$ on that space.

Definition A.6 (σ-Finite)**.** *A measure space $(\mathcal{X}, \mathcal{A}, \mu)$ is called* σ-finite *if there exist sets $(X_i)_{i\in\mathbb{N}} \subset \mathcal{A}$ such that*

$$\mu(X_i) < \infty \text{ for all } i \in \mathbb{N} \text{ and } \mathcal{X} = \bigcup_{i\in\mathbb{N}} X_i.$$

Proposition A.7 (Product Measure)**.** *Let $(\mathcal{X}, \mathcal{A}, \mu)$ and $\mathcal{Y}, \mathcal{B}, \nu)$ be two σ-finite measure spaces. Then there exists a unique measure on $\mathcal{A} \otimes \mathcal{B}$, called the* product measure *of μ and ν and denoted by $\mu \otimes \nu$, such that*

$$\mu \otimes \nu(A \times B) = \mu(A)\nu(B) \quad \text{for all } A \in \mathcal{A},\ B \in \mathcal{B}.$$

Proof. The proof of this well-known fact can e.g. be found in [16], chapter 8.2, theorem A. □

Definition A.8 (Product Topology, Borel σ-Algebra on Product Space)**.** *Let $(\mathcal{X}, \mathcal{T}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{T}_\mathcal{Y})$ be topological spaces. Then the* product topology *on $\mathcal{X} \times \mathcal{Y}$ is the topology generated by sets of the form $\pi_\mathcal{X}^{-1}(X)$ for $X \in \mathcal{T}_\mathcal{X}$ and $\pi_\mathcal{Y}^{-1}(Y)$ for $Y \in \mathcal{T}_\mathcal{Y}$, i.e. the coarsest topology such that $\pi_\mathcal{X}$ and $\pi_\mathcal{Y}$ are continuous. We denote it by $\mathcal{T}_{\mathcal{X}\times\mathcal{Y}}$.*
The Borel σ-algebra on the product space is the σ-algebra $\mathcal{B}(\mathcal{X} \times \mathcal{Y})$ generated by $\mathcal{T}_{\mathcal{X}\times\mathcal{Y}}$.

Remark A.9 (Properties of Polish Spaces)**.** In the thesis, we will often face product spaces of two spaces, say $\mathcal{X}$ and $\mathcal{Y}$. As one can see from the definitions above, if both of these are equipped with their Borel σ-algebras $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$, there are really two logical choices for what σ-algebra we could consider on the product space now: We could either take the product σ-algebra $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y})$, or the Borel σ-algebra on the product space, $\mathcal{B}(\mathcal{X} \times \mathcal{Y})$. In general, these two σ-algebras do not coincide. However, if both $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces, the two σ-algebras on the product space turn out to indeed be the same.[19] This means we do not have to worry about which σ-algebra to choose.
Also note that the product of two Polish spaces, equipped with the product topology, will again be a Polish space.[20]

Lemma A.10. *Let $(\mathcal{X}, d)$ be a metric space and $x, y \in \mathcal{X}$. Then for any $z \in \mathcal{X}$ there holds*

$$d(x, y)^p \le 2^p(d(x, z)^p + d(z, y)^p).$$

Proof. We have

$$d(x, y)^p \le (d(x, z) + d(z, y))^p = \sum_{k=0}^{p} \binom{p}{k} d(x, z)^{p-k} d(z, y)^k.$$

Now in the case $d(x, z) \le d(z, y)$ this gives us

$$d(x, y)^p \le \sum_{k=0}^{p} \binom{p}{k} d(z, y)^p = 2^p d(z, y)^p,$$

[19] A proof of this result can e.g. be found in [13], lemma 1.2. Note that completeness of the spaces is not even needed.
[20] Here we make a slight abuse of notation, as our definition of a Polish space as in section 2 was that it is a complete, separable, *metric* space; however, in this case, the product space is *metrizable*. This means we can find a metric that induces the product topology. Choose, e.g., $d_{\mathcal{X}\times\mathcal{Y}} := \max(d_\mathcal{X}, d_\mathcal{Y})$, i.e. $d_{\mathcal{X}\times\mathcal{Y}}((x_1, y_1), (x_2, y_2)) = \max(d_\mathcal{X}(x_1, x_2), d_\mathcal{Y}(y_1, y_2))$. Then it is straightforward to show that this is indeed a metric on the product space that induces the product topology, and that $\mathcal{X} \times \mathcal{Y}$ equipped with this metric is a Polish space.

whereas in the case $d(x,z) \geq d(z,y)$ we get

$$d(x,y)^p \leq \sum_{k=0}^{p} \binom{p}{k} d(x,z)^p = 2^p d(x,z)^p.$$

Combining these two inequalities yields the claim. □

Proposition A.11. *Let $(\mathcal{X}, d)$ be a metric space. Then $\mathcal{X}$ is compact if and only if it is sequentially compact, meaning any sequence in $\mathcal{X}$ contains a convergent subsequence.*

Proof. First, assume that $\mathcal{X}$ is compact and let $(x_i)_{i\in\mathbb{N}}$ be a sequence in $\mathcal{X}$. Assume for the sake of contradiction that it does not contain a convergent subsequence. This means it does not contain a cluster point, hence for any $x \in \mathcal{X}$ we can find an open neighbourhood U_x of x such that $\{i \in \mathbb{N} : x_i \in U_x\}$ is finite. This means $\{U_x : x \in \mathcal{X}\}$ is an open cover of $\mathcal{X}$, and by compactness we can find a finite subcover. However, this finite subcover can only contain a finite number of points x_i, which is a contradiction.
Now assume that $\mathcal{X}$ is sequentially compact. We will first prove that this implies that $\mathcal{X}$ is compete and totally bounded, and then show that this in turn implies that it is compact (note that this means we actually prove equivalence between "compact", "sequentially compact" and "complete + totally bounded"). Completeness follows immediately, as every Cauchy sequence contains a convergent subsequence by assumption. Now assume for the sake of contradiction $\mathcal{X}$ was not totally bounded, i.e. there exists some $\varepsilon > 0$ such that $\mathcal{X}$ cannot be covered by finitely many balls of radius ε. Let $B_\varepsilon(x_0)$ be an arbitrary ball of that radius for some $x_0 \in \mathcal{X}$. For $i \in \mathbb{N}_{>0}$, let $x_i \in \mathcal{X} \setminus (\cup_{j=0}^{i-1} B_\varepsilon(x_j))$. Then $(x_i)_{i\in\mathbb{N}}$ is a sequence in $\mathcal{X}$ with the property that $d(x_i, x_j) \geq \varepsilon$ whenever $i \neq j$. This means it cannot contain a convergent subsequence, which is a contradiction.
Now let $\mathcal{X}$ be complete and totally bounded. We need to prove it is compact. Assume for the sake of contradiction it was not, i.e. there exists an open cover $(U_i)_{i\in I}$ of $\mathcal{X}$ without a finite subcover. As $\mathcal{X}$ is totally bounded, we can cover it by finitely many sets $C_1^1, ..., C_{p_1}^1$ of diameter less than 1, and by assumption one of these sets, call it C^1, cannot be covered by finitely many U_i. Now C^1 can be covered by finitely many sets $C_1^2, ..., C_{p_2}^2$ of radius less than $\frac{1}{2}$ (without loss of generality let $C_i^2 \subset C^1$ for all $i \in [\![p_2]\!]$), and again one of them, C^2, cannot be covered by finitely many U_i. Proceeding like this, we find a sequence $C^1 \supset C^2 \supset C^3 \supset ...$ of sets C^i with diameters less than $\frac{1}{i}$, and each of them cannot be covered by finitely many U_i. Let $x_i \in C^i$ be an arbitrary point for each $i \in \mathbb{N}_{>0}$. Then $(x_i)_{i\in\mathbb{N}_{>0}}$ is a Cauchy sequence by construction, hence it converges to some $x \in \mathcal{X}$ by completeness. As $(U_i)_i$ is a covering of $\mathcal{X}$, we can find $i \in I$ such that $x \in U_i$. Let $\delta > 0$ such that $B_\delta(x) \subset U_i$. Let $N \in \mathbb{N}$ such that $d(x, x_N) < \frac{\delta}{2}$ and $\frac{1}{N} < \frac{\delta}{2}$. Then

$$C^N \subset B_{\frac{1}{N}}(x_N) \subset B_{\frac{\delta}{2}}(x_N) \subset B_\delta(x) \subset U_i,$$

which contradicts the fact that C^N cannot be covered by finitely many U_i. □

Theorem A.12 (Weierstraß)**.** *Let $(\mathcal{X}, d)$ be a metric space, $K \subset \mathcal{X}$ compact, and $I : \mathcal{X} \to [-\infty, \infty]$ a lower semicontinuous function. Then I attains its minimum on K, meaning there exists some $x_0 \in K$ such that*

$$I(x_0) = \min_{x\in K} I(x).$$

Proof. This follows directly from Theorem 3.6 in [11], as the statement there is a generalization of our proposition to topological spaces. □

Theorem A.13 (Prokhorov)**.** *Let $\mathcal{X}$ be a Polish space. Then a set $\mathcal{P} \subset P(\mathcal{X})$ is precompact for the weak topology, meaning its closure w.r.t. the weak topology is compact, if and only if it is tight.*

Proof. This can be found as Theorem 8.6.2 in [7]. □

Theorem A.14 (Disintegration)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces and $\mu \in P(\mathcal{X})$. Let $\pi : \mathcal{X} \to \mathcal{Y}$ be a Borel map and $\nu := \mu \circ \pi^{-1}$. Then there exists a ν-almost everywhere uniquely defined family $(\mu_y)_{y\in\mathcal{Y}} \subset P(\mathcal{X})$ of Borel measures on $\mathcal{X}$ such that*

$$\mu_y\left(\mathcal{X} \setminus \pi^{-1}(y)\right) = 0 \quad \text{for } \nu\text{-almost every } y \in \mathcal{Y},$$

and for any Borel map $f : \mathcal{X} \to [0, \infty]$ there holds

$$\int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x) = \int_{\mathcal{Y}} \int_{\pi^{-1}(y)} f(x)\,\mathrm{d}\mu_y(x)\,\mathrm{d}\nu(y).$$

Proof. A proof to this theorem can be found in [9], Chapter III, theorem 70 on page 78. It holds true in the even more general setting of Radon spaces. □

Theorem A.15 (Monotone Convergence)**.** *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and $(f_n)_{n\in\mathbb{N}} : \mathcal{X} \to [c, \infty]$ be an increasing sequence of measurable functions, where $c \in \mathbb{R}$. Then the pointwise supremum of these functions, $f := \sup_{n\in\mathbb{N}} f_n$, is measurable and*

$$\lim_{n\to\infty} \int_{\mathcal{X}} f_n(x)\,\mathrm{d}\mu(x) = \int_{\mathcal{X}} f(x)\,\mathrm{d}\mu(x).$$

Proof. A proof can e.g. be found in [6], Theorem 2.8.2. □

Lemma A.16 (Uniqueness of Measures integrated over $C_b(\mathcal{X})$)**.** *Let $(\mathcal{X}, d)$ be a metric space and μ and ν be two finite Borel measures on $\mathcal{X}$. Then $\mu = \nu$ if and only if $\int_{\mathcal{X}} \varphi\,\mathrm{d}\mu = \int_{\mathcal{X}} \varphi\,\mathrm{d}\nu$ for all $\varphi \in C_b(\mathcal{X})$.*

Proof. If $\mu = \nu$, then $\int_{\mathcal{X}} \varphi\,\mathrm{d}\mu = \int_{\mathcal{X}} \varphi\,\mathrm{d}\nu$ for all $\varphi \in C_b(\mathcal{X})$ is clear. Now let $\int_{\mathcal{X}} \varphi\,\mathrm{d}\mu = \int_{\mathcal{X}} \varphi\,\mathrm{d}\nu$ hold for all $\varphi \in C_b(\mathcal{X})$. Let $A \in \mathcal{B}(\mathcal{X})$ be closed and for $\varepsilon > 0$ define $f_\varepsilon : \mathcal{X} \to [0, 1]$, $f_\varepsilon(x) = \max\{1 - \frac{1}{\varepsilon} d(x, A), 0\}$ (where $d(x, A) := \inf_{y\in A} d(x, y)$). Then $f_\varepsilon \in C_b(\mathcal{X})$ and f_ε converges pointwise to $\mathbb{1}_A$ from above. Hence, by proposition A.15, $\mu(A) = \nu(A)$. □

Proposition A.17 (Minkowski Inequality)**.** *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and $p \in [1, \infty)$. Let $f, g \in L^p(\mu)$. Then $f + g \in L^p(\mu)$ and*

$$\left(\int_{\mathcal{X}} |f + g|^p\,\mathrm{d}\mu\right)^{\frac{1}{p}} \le \left(\int_{\mathcal{X}} |f|^p\,\mathrm{d}\mu\right)^{\frac{1}{p}} + \left(\int_{\mathcal{X}} |g|^p\,\mathrm{d}\mu\right)^{\frac{1}{p}}.$$

Proof. A proof of this famous inequality can e.g. be found in [6], theorem 2.11.9. □

References

[1] L. Ambrosio, N. Gigli. *A user's guide to optimal transport.* Modelling and Optimisation of Flows on Networks, pages 1–155, Springer, 2013.

[2] B. Amos, S. Cohen, G. Luise, I. Redko. *Meta Optimal Transport.* arXiv:2206.05262v1 [cs.LG], 2022.

[3] M. Arjovsky, S. Chintala, L. Bottou. *Wasserstein GAN.* arXiv:1701.07875 [stat.ML], 2017.

[4] D. P. Bertsekas. *A new algorithm for the assignment problem.* Mathematical Programming, 21(1):152–171, 1981.

[5] D. Bertsimas, J. Tsitsiklis. *Introduction to Linear Programming.* Athena Scientific *and* Dynamic Ideas, Belmont, Massachusetts, 1997.

[6] V. Bogachev. *Measure Theory Volume I.* Springer, Berlin, 2007.

[7] V. Bogachev. *Measure Theory Volume II.* Springer, Berlin, 2007.

[8] M. Cuturi. *Sinkhorn distances: lightspeed computation of optimal transport.* Advances in Neural Information Processing Systems 26, pages 2292–2300, 2013.

[9] C. Dellacherie, P.-A. Meyer. *Probabilities and Potential.* Hermann, Paris, 1978.

[10] T. Fischer. *Existence, uniqueness, and minimality of the Jordan measure decomposition.* arXiv:1206.5449 [math.ST], 2012.

[11] I. Fonseca, G. Leoni. *Modern Methods in the Calculus of Variations: L^p Spaces.* Springer, New York, 2007.

[12] S. Ioffe, C. Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.* arXiv:1502.03167 [cs.LG], 2015.

[13] O. Kallenberg. *Foundations of Modern Probability.* 2^{nd} edition, Springer, New York, 2002.

[14] D. P. Kingma, J. Ba. *Adam: A Method for Stochastic Optimization.* arXiv:1412.6980 [cs.LG], 2017.

[15] H. W. Kuhn. *The hungarian method for the assignment problem.* Naval Research Logistics Quarterly, 2:83–97, 1955.

[16] M. Loève. *Probability Theory I.* 4^{th} edition, Springer, 1977.

[17] G. Monge. *Mémoire sur la théorie des déblais et des remblais.* Royale Sci. Paris, 3, 1781.

[18] J. B. Orlin. *A polynomial time primal network simplex algorithm for minimum cost flows.* Mathematical Programming, 78(2):109–129, 1997.

[19] K. R. Parthasarathy. *Probability Measures on Metric Spaces.* Academic Press Inc., 1967.

[20] G. Peyré, M. Cuturi. *Computational Optimal Transport.* Foundations and Trends in Machine Learning, vol. 11, number 5-6, pages 355–607, 2019.

[21] A. Pratelli. *On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation*. Elsevier Masson SAS, 2006.

[22] R. Sinkhorn, P. Knopp. *Concerning nonnegative Matrices and doubly stochastic Matrices*. Pacific Journal of Mathematics, Vol. 21, No. 2, 1967.

[23] R. E. Tarjan. *Dynamic trees as search trees via euler tours, applied to the network simplex algorithm*. Mathematical Programming, 78(2):169–177, 1997.

[24] J. Thornton, M. Cuturi. *Rethinking Initialization of the Sinkhorn Algorithm*. arXiv:2206.07630v1 [stat.ML], 2022.

[25] C. Villani. *Optimal Transport Old and New*. Springer, Berlin Heidelberg, 2009.

Index

$(\mathcal{P}_p(\mathcal{X}), W_p)$, 30
1-Lipschitz, 19
$B_r(x)$, 8
$C(\mathcal{X})$, 9
$C_b(\mathcal{X})$, 9, 51
H, 34
$P(\mathcal{X})$, 9
S_n, 8
W_1, 29
W_p, 28
$\Pi(\mu, \nu)$, 11
δ_x, 9
$\langle \cdot, \cdot \rangle$, 8
$[\![m, n]\!]$, 8
$[\![n]\!]$, 8
$\mathcal{A} \otimes \mathcal{B}$, 48
$\mathcal{L}(X)$, 10
$\mathcal{L}(X, Y)$, 10
$\mathcal{P}_p(\mathcal{X})$, 28
$\mathcal{T}_{\mathcal{X} \times \mathcal{Y}}$, 49
$\mu \otimes \nu$, 49
μ^+, 48
μ^-, 48
$\partial^c \psi$, 20
$\partial^c \psi(x)$, 20
$\pi_{\mathcal{X}}$, 9
ψ^c, 18
σ-finite, 48
Id, 9
φ^c, 18
c-concavity, 19
c-convexity, 20
c-cyclical monotonicity, 22
c-subdifferential, 20
c-superdifferential, 20
c-transform, 18

algorithm
  auction, 33
  Hungarian, 33
  network simplex, 33
  simplex, 33
  Sinkhorn, 33, 37
    initialization, 39

Borel σ-algebra on product space, 49

competitiveness, 18
coupling, 11
  deterministic, 12
  trivial, 12, 34

duality
  entropic optimal transport, 35
  optimal transport, 27

Earth Mover's distance, 28
entropy, 34

Frobenius dot-product, 8
function
  strongly convex, 34

Gibbs kernel, 35
gluing lemma, 29

Jordan decomposition, 48

Kantorovich, 6
  dual problem, 18
  problem, 13, 32
Kantorovich-Rubinstein distance, 28
Kronecker product, 8

law, 10
linear program
  dual, 32
  optimal solution, 32
  primal, 32
lower semicontinuity, 9
  of the cost functional, 16

marginal, 11
measurable space, 48
measure, 48
  Borel, 9
  concentrated, 9
  Dirac, 9

discrete, 48
finite, 9, 48
mutually singular, 48
product, 49
pushforward, 9
signed, 23, 48
support, 9
measure space, 48
Minkowski inequality, 31, 51
Monge, 6
problem, 12

negligible, 9
neighbourhood, 8

optimal transport
discrete, 31
discrete dual, 32
discrete primal, 32
dual problem, 18
entropic
solution, 35
entropic dual, 35
entropic primal, 34
primal problem, 13

Polish space, 8, 49
potential, 18
precompact, 51
product σ-algebra, 48
product topology, 49

regularizer, 34

sequentially compact, 50
Sinkhorn-Knopp fixpoint iteration, 37
support
of a function, 9
of a measure, 9

theorem
disintegration, 29, 51
duality, 27
fundamental theorem of optimal transport, 23
monotone convergence, 16, 51
Prokhorov, 13, 51
Weierstraß, 14, 50
tightness
of a measure, 14
of a set, 9
of prices, 18
of transport plans, 14
totally bounded, 8
transport map, 12
transport plan, 11
optimal, 13
existence, 16

variation
lower, 23, 48
total, 48
upper, 48

Wasserstein, 28
distance, 28
is a metric, 30
space, 28
is Polish, 31
weak convergence, 9
weak topology, 9